Hypothesis Testing 1

4 Hypothesis Testing

Hypothesis testing is the process of making decisions based on numbers calculated from available data. It will draw upon the ideas we have come across in statistics and in probability and in particular our study of random variables. The phrase 'making decisions' implies a wide use of investigations may be assisted by a Hypothesis test; we will see that in fact most of these are tests on the validity of a proposed value of a parameter. In particular, the Null Hypothesis-based statistical tests we will study here are confirmatory studies or data analysis; that is, we have a thesis we wish to confirm or reject, rather than an exploratory analysis.

This type of confirmatory analysis, the Null-hypothesis based test, dominates the physical and social sciences, so it is important to understand it fully and be aware of what it tells us and whet it does not tell us. Before we look at hypothesis testing and how it is done, we must look again at the idea of sampling and the concept of a statistic.

4.1 Statistics and Samples

So far we have been using the word 'statistics' to mean descriptive statistics, that is, calculating certain numbers from a set of data in order to describe or summarise that data.

We will formalise the link between parameters such as the mean or standard deviation calculated for a data set and the mean and standard deviation known for a random variable. Recall that the production of a value for a given random variable occurs when the experiment that leads to the expression of the variable is carried out. Therefore we treat a set of n values in a dataset as produced values of a given random variable, obtained when the experiment is run n times. We say therefore that the set of actual values is a sample from the random variable.

To see the link with the parameters of the variable and to understand the behaviour of a sample, we may also treat a sample value X_i of a variable X as being itself a random variable with the identical distribution as X itself. With this context, the calculated values of a parameter may be regarded firstly as a random variable and secondly as an estimate of a true underlying figure. Thus for random variable X, with mean μ, assume we take n values of the variable, that is, the values

$$X_i \text{ for } i = 1 \text{ to } n.$$

The mean calculated from this data, given by

$$\overline{X} = \frac{\sum_i X_i}{n}.$$

This is now also a random variable and is an estimate of the true mean μ. Using the word estimate means that

$$\lim_{n\to\infty} \overline{X} = \lim_{n\to\infty} \frac{\sum_i X_i}{n} = \mu.$$

It is reasonable to ask what the meaning of this limiting process is; it may be viewed as a probability or an expected value:

$$\lim_{x\to\infty} P[\,|\,\overline{X} - \mu| > 0] = 0.$$

A random variable such as the mean above, calculated from other random variables, is the real meaning of the word *statistic*.

4.1.1 Definition – A Statistic

A quantity is called a *statistic* if it is calculated from the possible values of one or more random variables, equivalently, if it is calculated from a set of random variables. The statistic is itself a Random Variable.

This means that quantities such as the mean or the standard deviation of a data sample of a variable are *statistics* in this truer sense of the word, because the data they are calculated from are the possible values of a random variable; in other words, the data they came from is random in nature.

4.1.2 Null Hypothesis Based Hypothesis Testing

This is the process of making a decision on the truth or otherwise of a statement using a statistic with a known distribution. A question is posed, usually framed as a statement that a particular value of a given parameter for a random variable is correct. A statistic is identified which will be a function of the parameter and its estimate from any potential data. The distribution of this statistic is identified, crucially under the assumption that the assumed value of the parameter is correct; this assumption is therefore called the Null Hypothesis, the starting position. Data is collected, regarded as expressions of the random variable in question. A value is produced for the statistic and it is decided whether this value is unlikely or not, usually by comparing it to a threshold probability. If it is an unlikely value, then the truth of the Null hypothesis is called in to question and the Null hypothesis is rejected.

The power and applicability of this method comes when we view this as an investigation into a scientific question; a default position of our understanding of the question is stated and the situation is modelled using a random variable. This default position is the Null Hypothesis. It is then stated as a value of a parameter of the random variable, where possible. Data is collected; this is viewed as a realisation of the random variable and we proceed as before.

The nature of the probabilities being considered must be emphasised at the start; it may also be regarded as a fundamental drawback of this method. Let H_0 be the statement that the Null hypothesis is true. Let D be the data we are collecting, whose probability under the Null hypothesis is measured by the given statistic. The test procedure gives us the probability

$$P[D \mid H_0].$$

If this is a low enough compared to a threshold value, we reject H_0. Always bear in mind that we do not calculate the probability of the Null Hypothesis being true, given the data we have collected:

$$P[H_0 \mid D].$$

Before moving on to an example, we recall two essential definitions. Let X be a random variable, for which μ is the true underlying mean.

- Let the symbol $\bar{x}$ be the estimate of the mean calculated from a sample of n values of X, that is, the value:

  $$\bar{x} = \frac{\sum_i x_i}{n}.$$

  This quantity is now referred to as the *sample mean*.

- Let the symbol s be the estimate of the standard deviation calculated from a sample of n values of X via:

  $$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1} = \frac{\sum_i x_i^2 - n\bar{x}^2}{n-1}.$$

  This figure is now called the *sample standard deviation*.

With this terminology, we will look at an example of testing a mean.

4.2 A Starting Example

The simplest Null-Hypothesis based statistical test and the statistic used in this test comes directly from the central limit theorem and the normal distribution. It is used to test a proposed value of a mean, calculated in this case from a list of data values. By ‘testing a mean’, our intention is to see if the actual value we get for a mean from the data, in other words the sample mean, is too far from what we think the true, underlying mean is, to be likely. This idea is illustrated first in the following example.

4.2.1 First Example

The following are the measurements of the resistance, in kΩ, of a sample of 55 components being produced by a manufacturing process. The measurements are all in kΩ. It is proposed that the value of the true underlying mean is 24 kΩ. Here is the raw data:

24.12	28.82	21.37	21.60	22.24
27.88	29.81	25.11	20.19	25.75
24.63	24.67	21.09	21.34	23.34
21.43	23.46	28.74	21.62	28.66
22.16	21.92	21.92	24.23	25.02
22.34	27.76	28.71	29.88	22.94
25.41	29.56	22.82	25.09	29.37
24.24	26.76	25.49	22.09	24.27
28.72	26.46	22.40	26.56	23.54
25.44	22.24	29.95	20.60	29.98
23.78	28.98	29.41	28.63	21.08

Since the only way we can estimate the mean is with the sample mean, we need to know how well it estimates the true mean of the wider overall population. Take note at this point that we are now inherently modelling the resistances of these components as a random variable, call it X, and when we talk of the mean μ, we are talking about the mean of this variable. First, calculate the sample mean for this set of data.

- Let n be the number of data values in a list, with $n = 55$.
- Let x_i, for $i = 1$ to n, be the list of values. The sample mean of these figures is then

$$\bar{x} = \frac{\sum_i x_i}{n} = 1375.62 \text{ k}\Omega/55 = 25.0113 \text{ k}\Omega.$$

This is close to the proposed figure of 24kΩ for μ, but we must decide if it is close enough to stick with the vale of 24Ω or different enough that we should change our estimate. We have already seen one way to look at this problem, the confidence interval. We first need s, the standard deviation. The calculation is shown here; the top line for the standard deviation calculation is then

$$34{,}905.44 - 55\text{x}(25.0113)^2 = 499.4366.$$

Therefore $s^2 = 499.4366/54$, so $s = 3.0412$ kΩ.

We will calculate two confidence intervals, firstly for 95% and again for 99%. The required values from the standard normal distribution are

$$z_{0.05/2} = 1.960 \text{ and } z_{0.01/2} = 2.576.$$

With s and knowing $n = 55$, the two confidence intervals are

$$\bar{x} \pm z_{0.005}\frac{s}{\sqrt{n}} \text{ and } \bar{x} \pm z_{0.025}\frac{s}{\sqrt{n}}.$$

Carrying out the calculations gives

$$99\% : [\ 17.248\ \text{k}\Omega,\ 32.773\ \text{k}\Omega\ ],$$

$$95\% : [\ 19.105\ \text{k}\Omega,\ 30.917\ \text{k}\Omega\ ]$$

In both cases the proposed mean is in the confidence interval, so we accept it as a reasonable figure based on our data. There is one caveat for this work; our value of $n = 55$ is not particularly high to be able to say that these confidence interval limits are properly following a normal distribution in accordance with the Central limit theorem. The best way to illustrate this would be with a simulation.

We will now treat this question as a Null-Hypothesis based test, using much the same information. We will measure what we mean by 'close enough' or 'different enough' by seeing how likely or unlikely this value is for $\bar{x}$, if it is correct that the sample of values we have is to be regarded as expressions of a random variable with $\mu = 24$ kΩ. To achieve this, recall that the Central limit theorem tells us that the random variable Z given by the equation

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

is the standard normal variable, when the number n is very large. Because z is calculated from the random variables $\bar{x}$ and s, it is a random variable and a *statistic*, in the sense defined above. It will therefore allow us to construct some measure of $P[D \mid H_0]$.

Our procedure will be a follows; we will start with the statement that the proposed value $\mu = 24$ kΩ of the mean for the random variable X is correct; this is the Null Hypothesis.

Now choose a threshold probability that we consider to be low, in other words, representing an unusual event. This probability is usually denoted by α in keeping with the ideas of the confidence interval. The values of 0.01 or 0.05 are frequently used in this context. We will now look at an event involving z with this probability. With a given value of the mean μ, the value of z could be very low or very high. This suggests using the idea of the 'middle range' of the variable Z, for the probability α. Recall our definition of z_α so that

$$P[Z > z_\alpha] = \alpha.$$

Then we are looking for the two values $\pm z_{\alpha/2}$ such that

$$P[Z > z_{\alpha/2}] = \alpha/2 \text{ and } P[Z < -z_{\alpha/2}] = \alpha/2.$$

This makes our test a *two-tailed test*; we will review this idea later. For $\alpha = 0.05$, the values are ±1.96:

$$P[Z > 1.960] = 0.025 \text{ and } P[Z < -1.960] = 0.025.$$

Therefore a z-value outside the range –1.96 to +1.96 qualifies as an unusual event, presumably inconsistent with the idea that the true mean is 24 kΩ. For $\alpha = 0.01$, the values are ±2.576. These values of a test statistic, which quantify what is regarded as an unusual event for a particular statistic, are called *critical values*.

Working out the value of z is straightforward. We saw already that $s = 3.0412$ kΩ, so that:

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{25.0113 - 24}{3.0412/\sqrt{55}} = 2.4661.$$

The value we found for z is outside the range of the two critical values, so we can say this was, by the standard we have set, an unusual event. Therefore the value of 25.0113 kΩ, found for the mean, is unusual enough to set aside our claim of 24 kΩ. We say that we reject the Null Hypothesis of $\mu = 24$ kΩ.

4.2.2 A One-Tailed Version

We will now look at the same data, but apply a slightly different question to it.

Let us say we have had an advance look at the data and know that sample mean of these figures is the previously calculated figure:

$$\bar{x} = 25.0113 \text{ k}\Omega.$$

There is now no point in pursuing the possibility that $\bar{x}$ is too low, since it is clearly above $\mu = 24$ kΩ. [This business of having some advance knowledge of what the data is saying before carrying out a test is called *data snooping*.] The events we are looking for, for our probability α, are

$$P[Z > z_\alpha] = \alpha \text{ or } P[Z < -z_\alpha] = \alpha.$$

This makes our test a *one-tailed test*. For example, for $\alpha = 0.05$, the values are ±1.645:

$$P[Z > 1.645] = 0.05 \text{ or } P[Z < -1.645] = 0.05.$$

It is the first critical value we need since we know that Z will through up a positive number. For $\alpha = 0.01$, the corresponding critical value is 2.326.

The calculation with the data goes exactly the same way, giving a value of 2.4661. Again, we reject the Null Hypothesis.

4.3 The framework of a Hypothesis test

The analysis carried out in the previous section is a Null-Hypothesis based test on a population parameter, in this case the mean. The structure is reviewed here.

In a standard hypothesis test on a population parameter, such as the mean or standard deviation, a value for the parameter has been put forward. Sample data has been collected and an estimate of the parameter has been calculated. The question addressed is whether the estimate suggests the original value should be set aside or retained.

1. The first step is to identify, and then state, the situation where the proposed value is correct. The *Null hypothesis* is a statement that the value is correct. The *alternative hypothesis* is a statement that it is incorrect. Together, identifying these two ideas is called *framing the hypotheses*.
2. Decide on a *level of significance*; this means choosing a threshold low probability. In most cases this is taken as 0.05, particularly cases involving the social sciences or biology. For cases from physics or engineering a lower value for this probability is taken, such as 0.01.
3. Decide if the test is one- or two-tailed. If the claim of the population parameter can be invalidated either by an

unusually high value of the sample value of that parameter, or an unusually low value, then the test is one-tailed. If it could be either to high or too low, the test is two-tailed.

4. We now identify a statistic with a known distribution, which can measure the difference between the estimate of the parameter obtained from the data and the proposed value under the assumption that the null hypothesis is true. This difference is typically divided by a measure of the variance of that difference, such as the $s/\sqrt{n}$. This will then allow us to decide whether the value calculated for the statistic is unlikely or not. In effect, this measures $P[D \mid H_0]$.
5. Now find a critical value of the distribution; this is a value with which we can compare the value of the statistic to decide if the unusual event has occurred. For a two-tailed test, it will be two (linked) values, e.g. the plus or minus 1.96 figures found for the probability of 0.05 for the z variable.
6. Now calculate the sample value of the statistic and compare it to the critical values; if it is found to be an unlikely value, the Null hypothesis is rejected and the alternative is accepted.

The example we have just worked on will now be framed in this way, firstly as a two tailed and secondly as a one-tailed test. For a test on a mean, the majority of cases will involve a two-tailed test. This type of test is sometimes called a z-test, because it uses the standard normal distribution to decide on whether a proposed value for a mean can be accepted.

4.3.1 A Two-tailed Test on a Mean

For the first version of the test on the resistances, we are asking the question if the data collected supports the value of 24 kΩ for the true mean, or not. Therefore the Null Hypothesis H_0 is that the true mean for the resistances of the components is 24.

$$H_0 : \mu = 24.$$

The Alternative hypothesis H_A is that the mean is not 24:

$$H_A : \mu \neq 24.$$

The test statistic is the variable Z, which is given by the equation:

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

This variable has the standard normal distribution.

We choose a level of significance of 0.05.

Since the claim for the true mean μ could be invalidated by a higher or lower value of the sample mean, then we use a two-tailed test, so divide the 0.05 significance level by 2, to give 0.025. From the z-tables we find the values are ±1.96; in other words,

$$P[Z > 1.96] = 0.025 \text{ and also } P[Z < -1.96] = 0.025.$$

Therefore the values ±1.96 are our critical values.

From the data:

- The value of the sample mean is 25.0113.

- The value of the sample standard deviation is 3.0412.
- Then the calculation for z gives 2.4661.

This value is outside the range of the two critical values, so we reject the Null Hypothesis. The data collected does not support the claim that the true mean of the resistances is 24 kilo-Ohms.

4.3.2 A One-tailed Test on a Mean

For the second version of the test on the resistances, we are asking the question if the data collected supports the claim that the true mean is at most 24 kΩ, or if the value of μ is higher. Therefore the Null Hypothesis H_0 is that the true mean for the resistances of the components is 24, stated as

$$H_0 : \mu = 24.$$

The Alternative hypothesis H_A is that the mean is above 25:

$$H_A : \mu > 24.$$

We choose a level of significance of 0.05.

Based on the question being asked, we are faced with a one-tailed test. From the z-tables we find that

$$P[Z > 1.65] = 0.05.$$

Therefore the value 1.65 is our critical value. With the sample value $z = 2.4661$ being above the critical value, we reject the Null Hypothesis. The data collected supports the claim that the true mean of the resistances is higher than 24 kilo-Ohms.

4.3.3 Another Test on the Mean

The following are the lifetimes, in 10^6 seconds, of a selection of components produced by a manufacturing facility. Test the claim that the true mean is $25\text{x}10^6$ s.

21.04	22.11	29.98	22.49	20.23	24.97	31.49	28.16
28.66	21.74	26.33	25.43	23.27	27.86	21.14	26.61
27.56	26.18	28.23	24.59	26.5	26.93	20.75	29.85
25.65	24.56	28.99	28.17	28.23	28.71	34.4	28.54
23.75	24.34	23.95	31.92	20.12	28.63	24.28	25.03
29.13	22.32	25.49	24.62	23.85	27.64	21.8	29.66
			25.78	32.11			

Frame the hypotheses:

- Null Hypothesis H_0: the true mean is $25\text{x}10^6$ s: $\mu = 25$,
- Alternative hypothesis H_A: the mean is not $25\text{x}10^6$ s: $\mu \neq 25$.

Take a significance level of 0.05. This is a two-tailed test, since there is no ‘direction’ to the question. We know that

$$P[Z > 1.960] = 0.025,$$

so the critical values are ±1.96. Now that the framework of the test has been set up, it is time to carry out the calculations on the data. The relevant sums are:

$$\sum X_i = 1303.77, \sum X_i^2 = 34{,}545.445.$$

This allows us to calculate the sample mean and the sample standard deviation, 26.075 and 3.348 respectively, in units of 10^6 seconds. Therefore the value of z is:

$$z = (26.075 - 25) / (3.348/\sqrt{50}) = 2.272$$

This figure is outside the range, so the Null hypothesis is rejected. The data is unlikely if the true mean is 25×10^6 s.

4.3.4 Summary – The Test on a Mean

Here we will summarise the steps involved in a test on a proposed value for the true mean for a population. The test will be carried out on a sample of n values taken from the wider population. The symbol μ will denote the proposed value of the true mean. The steps are as follows:

1. The first step is framing the hypotheses. The *Null hypothesis* is a statement that the proposed value of the true mean μ is correct. The *alternative hypothesis* is a statement that it is incorrect.
2. Decide on a level of significance. In most cases this is taken as 0.05 or 0.01.
3. Decide if the test is one- or two-tailed. For a test on a mean, if the alternative hypothesis is that the proposed value for the mean was simply not correct then the test is two-tailed. If the alternative hypothesis is that it is higher or lower, then the test is one tailed.
4. The statistic Z, given by the equation

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

follows the standard normal distribution. It will measure the difference between the sample mean and the true mean.

5. Now find a critical value of the distribution. Set α to be the level of significance, then we need z_α such that

$$P[Z > z_\alpha] = \alpha \text{ or } P[Z < -z_\alpha] = \alpha.$$

   For a two-tailed test, we use $\alpha/2$ and the critical values are $\pm z_{\alpha/2}$.

6. Calculate the sample mean $\bar{x}$ and sample standard deviation s from the data and from this, calculate the sample value of z with the equation

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

7. The value found for the test statistic is compared to the critical values; if it is found to be an unlikely value, the Null hypothesis is rejected and the alternative is accepted.

We now look at the Confidence Interval alternative to the Null-Hypothesis-based test on a mean.

4.3.5 Confidence Interval Approach

The greatest philosophical drawback to this simplest of Null-Hypothesis based tests was noted; this drawback applies to all such tests. The test does the following

- The investigative question must first be framed as a model for a random variable and then as a proposed value for a particular parameter,
- This proposed value forms the Null Hypothesis; we do not ask 'is μ equal to a given value', we have proposed a value and ask if it can be falsified,
- Ultimately we decide on the Null Hypothesis if we can show that $P[D|H_0]$ is remote, where D is our data, or rather, the value of a test statistic calculated from our data, and H_0 is the Null Hypothesis.

We restate here the concept of a confidence interval. For a level of significance value α, define a number z_α such that

$$P[Z > z_\alpha] = \alpha.$$

Then recall that the central limit theorem gave us the result that

$$P\left[\overline{X} - \frac{z_{\frac{1}{2}\alpha}}{S/\sqrt{n}} < \mu < \overline{X} + \frac{z_{\frac{1}{2}\alpha}}{S/\sqrt{n}}\right] = 1 - \alpha.$$

The confidence interval is then

$$\overline{X} - \frac{z_{\frac{1}{2}\alpha}}{S/\sqrt{n}} \text{ to } \overline{X} + \frac{z_{\frac{1}{2}\alpha}}{S/\sqrt{n}}.$$

This method of presenting information gleaned from our data is now quite widely used. We will apply a similar framework to other tests on a parameter.

4.4 A Test on a Proportion

The following example shows the mechanism of a test on a proportion. Consider a list of subjects or units who may or may not have an attribute. If we intend to count the number of subjects who have this attribute, then each individual is treated as a Bernoulli trial, with a 1 as a 'yes' and this is the binomial distribution as a sum of Bernoulli trials. Let B_i be a random variable which takes on one of two discrete values 0, 1, where

$$P[B = 1] = p, \text{ so that } P[B = 0] = 1 - p.$$

The density function was:

$$f(0) = 1 - p, f(1) = p.$$

The parameters for the distribution for B are:

$$\mu = p,\ \sigma^2 = p(1 - p).$$

The binomial distribution is then

$$N = \sum_{i=1}^{n} B_i\,,$$

Apply the central limit theorem; N is normally distributed with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$. We know what these values are, so then for the variable N:

$$\mu = np,\ \sigma^2 = np(1 - p).$$

Let us look at the mean number for variable B as well:

$$Z = \frac{\overline{B} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{B} - p}{\sqrt{p(1 - p)/n}}.$$

Now the mean of the sample values of the variable B is in fact an estimate of the proportion of the subjects with this attribute. Therefore if the true proportion for the random variable B is p and our estimate is $\hat{p}$, then the variable Z is

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim \mathrm{N}(0,1).$$

Two approaches can then be taken. First, we can take a large sample size and assume that s tends to σ, so that when we substitute in the usual expression for S^2 calculated from the sample values B_i, we can show that

$$Z = \sqrt{\frac{n-1}{\hat{p}(1-\hat{p})}}\left(\hat{p} - p\right) \sim \mathrm{N}(0,1).$$

The distinction of using $n - 1$ becomes less relevant for large n. It can be shown that this convergence of s to σ will happen for smaller n if p, so presumably the estimate $\hat{p}$, are not close to 0 or 1.

Alternatively, if we know that the proportion is going to be in the neighbourhood of 0.5, for example, estimating the proportion in a close referendum, a two-way election or the toss of an allegedly fair coin, then the true variance $\sigma^2 = p(1-p)$ will in fact be close to ¼.

4.4.1 Confidence Interval for a Proportion

The same logic we applied to the standard normal variable constructed from the mean of the sample from of a normal random variable leads to a confidence interval for the estimate of a proportion.

$$P\left[\hat{p}-z_{\frac{1}{2}\alpha}\sqrt{\frac{p(1-p)}{n}}<p<\hat{p}+z_{\frac{1}{2}\alpha}\sqrt{\frac{p(1-p)}{n}}\right]=1-\alpha\,.$$

Using sample information, in other words the estimate $\hat{p}$, this is

$$P\left[\hat{p}-z_{\frac{1}{2}\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}<p<\hat{p}+z_{\frac{1}{2}\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right]=1-\alpha\,.$$

Consider finally the option of taking ¼ as a value of $p(1-p)$;

$$P\left[\hat{p}-\frac{z_{\frac{1}{2}\alpha}}{2\sqrt{n-1}}<p<\hat{p}+\frac{z_{\frac{1}{2}\alpha}}{2\sqrt{n-1}}\right]=1-\alpha\,.$$

This number ¼ is the maximum value of the variance therefore it is a conservative estimate of the interval.

4.4.2 Example – Yes or No

A constitutional republic is having a referendum to decide on an issue, yes or no. To investigate the possible result, a sample of 2,500 have been asked for their intention to vote, of which 1239 have said they will vote ‘Yes’. We will analyse this data in here ways:

- Carry out a statistical test on the value $p = 0.5$, using $\alpha = 0.01$.
- Construct a 99% confidence interval for the true probability p of the Bernoulli variable modelling the voters intentions,
- Using the sample value of this parameter to estimate the probability the electorate vote ‘Yes’.

We will address the three options in order.

The statistical test is done straightforwardly, with two options. First note that with α = 0.01, the two-tailed critical values were $z_{0.005}$ = ±2.576. With the null hypothesis for a value of p, our test statistic is

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim \mathrm{N}(0,1).$$

With p = ½, it then follows that the variance is ¼ giving a standard deviation of ½. With n = 2,500 this becomes

$$Z = 100(\hat{p} - ½).$$

The sample value was $\hat{p}$ = 0.4956, so Z = 0.44, well away from the critical values. We do not reject the Null Hypothesis. It is worth noting that this is an identical test to that on a single fair die.

It can be argued that we have a second option to set aside the knowledge that the variance is given by $\sigma^2 = p(1-p)$ and use the sample value for the variance, in other words $\sqrt{\hat{p}(1-\hat{p})}$. In fact, carrying out the calculation in this instance yields ½ for the standard deviation for B to 5 decimal places, so the debate makes o practical difference.

The confidence interval is calculated from the equations quoted above. Using sample information, this is

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

For a 99% confidence interval, α = 0.01, so $z_{0.005}$ = 2.576. With the sample value $\hat{p}$ = 0.4956,

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = 0.010.$$

The bounds for the confidence interval are then [0.470, 0.521]. It is worth noting that using $\sqrt{\hat{p}(1-\hat{p})} \cong \sqrt{p(1-p)} \cong \frac{1}{2}$ would have yielded almost identical answers given how close $\hat{p}$ was to ½.

To address the final part of the question, we calculate the probability the electorate vote 'Yes' with the sample value $\hat{p}$ = 0.4956. This is the event that the binomial variable $Y = \sum B_i$ (Y is used since this variable is the number of yes votes!) gives up a value greater than ½N, where N is the total voting population. Therefore we want

$$P[Y > ½N].$$

We use the normal approximation to the binomial distribution, with $N\hat{p}$ as the mean and $\sqrt{N\hat{p}(1-\hat{p})}$ as the standard deviation:

$$Y \sim \mathrm{N}(N\hat{p}, \sqrt{N\hat{p}(1-\hat{p})}).$$

We are of course covering very similar ground computationally. Indeed if we standardise and do a bit of algebra we are calculating the probability

$$P[Z > \sqrt{\frac{N}{\hat{p}(1-\hat{p})}}\left(0.5-\hat{p}\right), \text{ where } Z \sim \mathrm{N}(0,1).$$

We have the value $\hat{p}$ = 0.4956, so our probability is

$$P[Z > 0.0088\sqrt{N}].$$

The striking result here is that the probability is that it depends on the number N, though not linearly. The probability of getting a yes vote, with a given value of the parameter, does depend on the population. Once the total electorate goes past 250,000 then the probability is that of Z being above 4, which is very low. For example, the total electorate in Ireland is approximately 3.4 million people. Why are pollsters not definitively sure how an election will go? People change their minds, frequently.

4.5 The Chi square Distribution and Tests

The next group of statistical tests are based on a particular continuous distribution known as the χ^2 statistic (pronounced 'ki' squared). Broadly speaking, this class of tests are used when seeing if a set of data, particularly a frequency distribution, can be matched with the distribution of a particular random variable, such as a uniform distribution or the normal distribution. The first test, which uses this variable and distribution, is a test on a variance. We will look at the distribution first, where we need a bit of notation.

4.5.1 The Gamma function

The quantity $\Gamma(x)$ is defined by the equation

$$\Gamma(x+1) = \int_0^\infty u^x e^{-u}\,\mathrm{d}u, \text{ for any number } x > -1.$$

This function is known as the gamma function. Using integration by parts, it is a simple matter to set up the following recursion relation:

$$\Gamma(x+1) = x\Gamma(x) \text{ , for any } x > -1.$$

It is a trivial matter to show that $\Gamma(1) = 1$. This leads us to the case where the argument is a positive integer n; these two results mean that

$$\Gamma(n+1) = \int_0^\infty u^n e^{-u}\,\mathrm{d}u = n!$$

4.5.2 Chi-square Distribution

The χ^2 distribution is now defined as follows. Let Z_i, for $i = 1$ to k, be a set of independent (sic) standard normal random variables. [We have only studied independence so far in terms of events; we will look at this concept later in the context of random variables and correlation.] So $Z_i \sim \mathrm{N}(0,1)$ for $i = 1$ to k. Let X be the random variable given by

$$X = Z_1^2 + Z_2^2 + \ldots + Z_k^2.$$

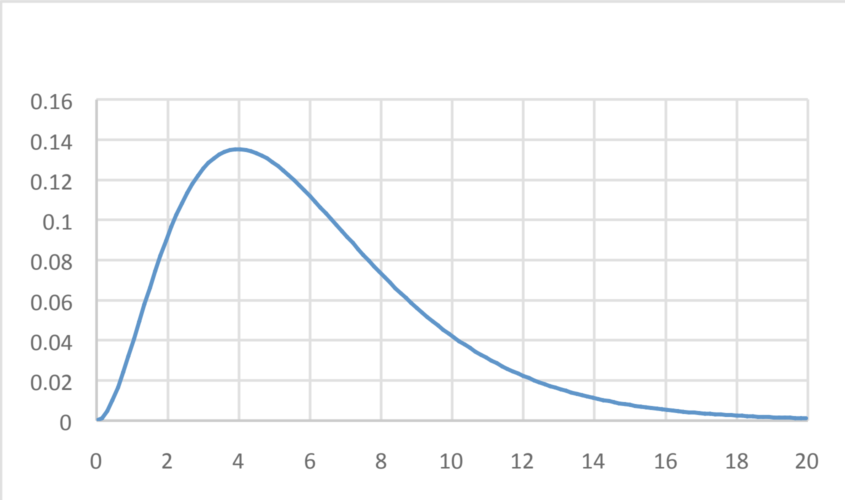
It can be proved that the density function for this variable is

$$f(x) = \frac{1}{2^{\frac{1}{2}d}\,\Gamma(\frac{1}{2}d)} x^{\frac{1}{2}d-1} e^{\frac{1}{2}x}.$$

The variable X is then said to follow the χ^2 distribution with k degrees of freedom. Integration by parts gives the distribution parameters:

$$E[X] = d \text{ and } Var[X] = 2d.$$

The graph below shows the probability density function for this distribution, for a degrees of freedom number of 6.



Take note of how the distribution is not symmetric and the χ^2 variable only takes on positive values. The area to the right of a given value a then gives us the probability $P[\chi^2 > a]$.

The first type of test used with this variable is a test on the variance, and by extension the standard deviation, the analogue of the tests on means which we started with.

4.6 Comparing variances

Consider again the case where a sample of n values of a normally distributed random variable X is given. The values are:

$$X_1, X_2, X_3, \ldots X_n.$$

Let σ^2 be the suspected or claimed value of this true variance. Let S be the *sample standard deviation*, and so S^2 is the *sample variance*:

$$S^2 = \frac{\sum_i \left(X_i - \overline{X}\right)^2}{n-1} = \frac{\sum_i X_i^2 - n\overline{X}^2}{n-1}.$$

We start with the following observation; consider the following sum of random variables constructed from X:

$$Q = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2.$$

Each element in this sum is a standard normal variable, but they are not independent since they all have the mean involved. Consider the following bit of algebra:

$$Q = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 = \sum_{i=1}^{n} \left(\frac{(X_i - \mu) - (\overline{X} - \mu)}{\sigma} \right)^2.$$

If we set

$$Z_i = \frac{X_i - \mu}{\sigma}, \text{ then } Z_i \sim \mathrm{N}(0,1) \text{ for } i = 1 \text{ to } k.$$

Then Q is written as

$$Q = \sum_{i=1}^{n} \left(Z_i - \overline{Z} \right)^2.$$

A bit of work with the summations here yields the further version:

$$Q = \sum_{i=1}^{n} Z_i^2 - n\overline{Z}^2.$$

Now the first term is the summation of n standard normal variables, but the last term is one standard normal variable that is fully linked to the first sum; it will not be proved here, but this means we can say that we can perform a linear transformation of the variables so that the quantity Q is therefore equivalent to the sum of n – 1 standard normal variables, that is, it follows the χ^2 distribution with n – 1 degrees of freedom. We can now write the quantity Q as

$$Q = \frac{\sum_i \left(X_i - \overline{X}\right)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}.$$

As observed, it has a χ^2 distribution with n – 1 degrees of freedom, so a test can be done with this as statistic. The tests go as follows.

- The Null Hypothesis H_0 is that the value σ^2 claimed for the variance of a normally distributed random variable is correct.
- The Alternative hypothesis H_A is that the true value for the variance is higher/lower/not equal to the σ^2 claimed.
- Under the Null Hypothesis, the quantity

  $$\frac{\sum_i \left(X_i - \overline{X}\right)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

  follows the χ^2 distribution with n – 1 degrees of freedom.
- In exactly the same way as we have seen with the test on a mean, the Null Hypothesis is rejected if the sample value of this statistic has a low probability; we set up a level of significance, critical values and identify the 'number of tails'.

We must look at what happens when we are carrying out a left-hand one tailed test or a two-tailed test. We need critical values for the lower end of a range, in other words, given a probability α, we have a method of producing a where

$$P[\chi^2 > a] = \alpha.$$

We must now find the value a so that

$$P[\chi^2 < a] = \alpha.$$

Because the χ^2 distribution is not symmetric we use the simplest law of probability for the converse, so that:

$$P[\chi^2 < a] = \alpha \text{ means } P[\chi^2 > a] = 1 - \alpha.$$

So we simply extend the table, from values 0.05, 0.025 down to 0.005, to include values 0.95, up to 0.995.

4.6.1 Example: A Test on the Variance

A company is machining engine parts with a particular component diameter X, measured in millimetres. The variance of this diameter should be no larger than 0.02mm^2. A random sample of 10 diameters gave a sample variance s^2 = 0.03mm^2. Carry out a one-tailed test on this claim.

Before embarking on this solution, note that the information available includes the sample variance; prime data snooping. Since we see before doing the test that the sample value is in fact higher than 0.02, we do a one-tailed test. The only question left is if it is high enough to reject the claim of 0.02mm^2.

- The null hypothesis is H_0: $\sigma^2 = 0.02$,
- The alternative is H_a : $\sigma^2 > 0.02$, a one-tailed test.
- Take a level of significance of 0.05.

- For this level of significance, at 9 degrees of freedom, the critical value is $\chi^2 = 16.92$; $P[\chi^2 > 16.92] = 0.05$.
- The sample value of the variable is: $\frac{(n-1)S^2}{\sigma^2}$.
- Variable s is the sample variance, $s^2 = 0.03$ mm^2. The value of the true variance σ^2, under the Null Hypothesis is the claimed value, 0.02 mm^2. The test value is 9x0.03/0.02 = 13.5

The null hypothesis is therefore not rejected.

4.6.2 Another Test on Variance

In the same situation, with a variance of $\sigma^2 = 0.02$mm^2, a random sample of 14 diameters gave a sample variance $s^2 = 0.015$mm^2. Carry out an appropriate test on the claim.

We are carrying out the same test, but now the alternative is different, because we know that $s^2 = 0.015$ mm^2. The solution is:

- The hypotheses are: H_0: $\sigma^2 = 0.02$, H_a : $\sigma^2 < 0.02$.
- At probability 0.95 at 13 degrees of freedom, the left-hand critical value is $\chi^2 = 5.89$; $P[\chi^2 < 5.89] = 0.05$.
- We have $s^2 = 0.015$ mm^2. The value of the true variance under the Null Hypothesis is $\sigma^2 = 0.02$ mm^2.
- The value calculated is 13x0.015/0.02 = 13x0.75 = 9.75.

The null hypothesis is therefore not rejected.

4.6.3 A Two-tailed Example with Data

A component manufacturer taken a random sample of 13 components, which are tested to destruction and their lifetimes recorded in units of 10^5 seconds. Carry out an appropriate statistical test on each of the two statements given here, using the data given below.

a) The mean lifetime of the components is 108.

b) The variance of the lifetimes is 5.2.

Denote the values by L_i, with $n = 13$. The available data is:

$$\sum_i L_i = 1{,}410.8, \text{ and } \sum_i L_i^2 = 153{,}123.25.$$

The first test will be a two-tailed test on the mean and the second will be a two-tailed test on the variance, since we have no prior information about the value of s^2. We will take 0.01 as the level of significance throughout.

For part (a), the Null Hypothesis is that the mean is 108. The alternative Hypothesis is that it is not 108. To phrase these mathematically; let μ be the mean of the radius of the discs.

- The Null Hypothesis H_0 is that $H_0 : \mu = 108$.
- The Alternative hypothesis H_A: $\mu \neq 108$.

Given the alternative hypothesis, the test will be a two-tailed t-test on the mean, because there is no expectation on which direction the mean goes.

We have 0.01 as the level of significance and divide this by 2 for a two-tailed test. In this instance, we are dealing with a small sample size; the method for dealing with this issue will be covered next. For the time being, we take it that the critical values are ±3.055, for the same test statistic for the mean. The sample mean is:

$$1410.8/13 = 108.523.$$

The sum of squares is given as 153,123.25, therefore the sample standard deviation is:

$$s^2 = (153{,}123.25 - 13\text{x}108.523^2)/12 = 18.893/12 = 1.574$$

So : $s = 1.255$, from this,

$$(108.523 - 108)/(1.255/\sqrt{13}) = 1.503.$$

This is well inside the range of the critical values so the null hypothesis is not rejected; the value of $108\text{x}10^5$s is a valid figure for the mean lifetime.

For part (b), the variable used will be the χ^2 variable. The Null Hypothesis is that the variance is 5.2, and the alternative Hypothesis is that it is not 5.2. To phrase these mathematically; let σ^2 be the underlying variance of the radius of the discs.

- Then the Null Hypothesis is $H_0 : \sigma^2 = 5.2$
- The Alternative hypothesis $H_A : \sigma^2 \neq 5.2$.

The test will be a two-tailed χ^2-test, because there is no expectation on which direction the variance goes.

- With 0.01 as the level of significance, divide this by 2 for a two-tailed test, giving 0.005.
- The number of degrees of freedom is $n - 1 = 12$. This means the upper critical value is 28.3, from the tables.

- From the tables, looking up the value for 0.995, the lower critical value is then 3.074.

With the critical values established, the test statistic is given by

$$\frac{(n-1)S^2}{\sigma^2}.$$

The top line from the previous calculation for *s* gives

$$(n - 1)s^2 = 18.893.$$

The variable σ^2 is the supposed true underlying value of the variance, here 5.2. With this information, the sample value is

$$\chi^2 = 18.893/5.2 = 3.779.$$

This is within the range so the Null Hypothesis is not rejected. The claim of 5.2 is not invalidated by the data we collected.

Note that if we had taken a level of significance of 0.05, the lower bound would be 4.404, and therefore in that case the Null Hypothesis is rejected.

4.6.4 Confidence Interval Approach

We restate the concept of a confidence interval for a test on a variance. For a level of significance value α, with degrees of freedom k, define a number $\chi^2(k, \alpha)$, the critical value, such that

$$P[\chi^2 > \chi^2(k, \alpha)\ ] = \alpha.$$

The bracket notation here is used rather than subscripts, as there are two quantities that the critical value depends on. Then from our study of this variable applied to a data set,

$$\frac{(n-1)S^2}{\sigma^2},$$

we have seen that for a probability β we can say

$$P\left[\frac{(n-1)S^2}{\sigma^2} > \chi^2(n-1,\beta)\right] = \beta.$$

Then apply $\beta = \frac{1}{2}\alpha$ and rearrange to get

$$P\left[\frac{(n-1)S^2}{\chi^2(n-1,\frac{1}{2}\alpha)} > \sigma^2\right] = \tfrac{1}{2}\alpha.$$

Apply $\beta = 1 - \frac{1}{2}\alpha$ and the probability of the converse to get

$$P\left[\frac{(n-1)S^2}{\sigma^2} < \chi^2(n-1,1-\tfrac{1}{2}\alpha)\right] = \tfrac{1}{2}\alpha.$$

Rearranging means that

$$P\left[\frac{(n-1)S^2}{\chi^2(n-1,1-\frac{1}{2}\alpha)} < \sigma^2\right] = \tfrac{1}{2}\alpha.$$

The confidence interval is then

$$\frac{(n-1)S^2}{\chi^2(n-1,\frac{1}{2}\alpha)} \text{ to } \frac{(n-1)S^2}{\chi^2(n-1,1-\frac{1}{2}\alpha)}.$$

4.7 The t-Distribution

The next step is to introduce a distribution to deal with the case of small sample sizes; it is called the *t*-distribution, also known as Student's *t*-distribution.

In English the test takes its name from the work of William Sealy Gosset at the Guinness Brewery here in Dublin, who worked on the problems of small samples of the ingredients they used.

One story on the origin of the pseudonym is that Guinness preferred staff to use pen names when publishing scientific papers instead of their real name, so he used the name "Student" to hide his identity. Another version is that Guinness did not want their competitors to know that they were using the *t*-test, or indeed statistics in the first place, to test the quality of raw material.

This distribution is constructed as follows. Let Z be a standard normal variable, with k such variables listed as Z_i, for $i = 1$ to k. All these variables are independent. So Z, $Z_{ip} \sim N(0,1)$ for $i = 1$ to k. Let X be the random variable given by

$$Q = Z_1^2 + Z_2^2 + \ldots + Z_k^2.$$

As we have seen, Q follows the χ^2 distribution with k degrees of freedom. The variable Q will be independent of Z. Consider now the variable T constructed as

$$T = \frac{Z}{\sqrt{Q/k}}.$$

The distribution for this variable is the Student's *t*-distribution with k degrees of freedom. The density function f_k for the variable with k degrees of freedom is

$$f(x) = C\left(1 + \frac{x^2}{k}\right)^{-\frac{1}{2}(k+1)}, \text{ where } C = \frac{\Gamma(\frac{1}{2}(k+1))}{\sqrt{\pi k}\,\Gamma(\frac{1}{2}k)}.$$

It can be readily shown that as k becomes very large, this becomes the expression for the standard normal variable. Firstly, set $g(x)$ to be the function defined as

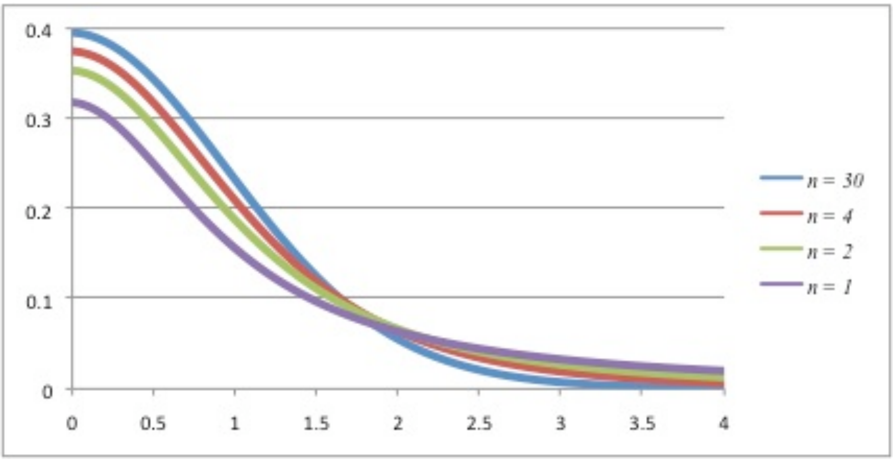
$$g(x) = \frac{\Gamma(\frac{1}{2}(x+1))}{\sqrt{x}\Gamma(\frac{1}{2}x)}.$$

It can be shown using the properties of the gamma function, specifically $\Gamma(x + 1) = x\Gamma(x)$ for any $x > -1$, that $g(x)$ heads to 1 as x becomes large. Apply this to $\frac{1}{2}(n + 1)$ and $\frac{1}{2}n$ and it is seen that for large k,

$$\pi C^2 \sim \tfrac{1}{2}k.$$

This result, combined with the limit definition of the exponential, gives the density function as that of the standard normal distribution.

The t-distribution is very similar to the standard normal distribution once the degrees of freedom parameter becomes large. The right-hand side of the distribution is shown here for several values of this parameter.



The distribution therefore has some very similar properties to N(0,1):

1. The two distributions are symmetric. This is an extremely important and useful property of both distributions, and we will use this frequently.
2. The expected value is $E[T] = 0$, in the same way as the z. The variance of the distribution is $Var[T] = n/(n - 2)$, for $n > 2$. It is notable that this number is not defined when $n = 1$ or 2.
3. For a variable with the normal or z-distribution, the probability of getting a number close to 0 is very high, and the probability of getting a large positive or negative number is low. The t-distribution is the same, but with the t-distribution, the lower the value of n, the more likely the higher values become. It has a longer 'tail' than N(0, 1).

We will now see how this distribution is used in a test.

4.8 The *t*-test on a Mean

In the context of a Hypothesis test, the statistic T has an immediate application for tests on a mean for sample values of a normal variable. Consider the variable given by the equation:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}},$$

where the sample parameter values are the familiar functions of a data set as before. Consider the following rearrangement of the term, done to bring in known variables into the expression:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{S}\sqrt{n} = \frac{\dfrac{\bar{X}-\mu}{\sigma}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}}\sqrt{n} = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}}.$$

The quantity above the line is standard normally distributed. The quantity under the square root below the line may be written as

$$\frac{(n-1)S^2}{\sigma^2(n-1)} = \frac{\chi^2_{n-1}}{(n-1)},$$

where the variable $\chi^2{}_{n-1}$ follows the χ^2 distribution with $n - 1$ degrees of freedom. We therefore have the correct ratio for the Student t-distribution. It remains to observe, unproven, that S is independent of Z. Therefore the quantity T given by

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}},$$

identical to the expression for the Z variable in a test on a mean, where all symbols are as before, follows the t-distribution, with $n - 1$ degrees of freedom.

The properties listed above mean that carrying out a z-test or a t-test or constricting confidence intervals are essentially the same process, when the variable X is normally distributed. Confidence Intervals calculated for small sample sizes using critical values of the distribution are exact. The t-distribution is robust against this condition on X being weakly observed.

4.8.1 Tables for the t-distribution

Because the Student's *t*-distribution is almost exclusively used in the context of a statistical test, we use a table of critical values such as the one shown here.

	Probability:			
Degrees of freedom	**0.050**	**0.025**	**0.010**	**0.005**
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
…	…	…	…	…
39	1.685	2.023	2.426	2.708
…	…	…	…	…

This is the simplest version of the table, with the following structure.

- The values in bold across the top are the probabilities,
- The integers in bold down the first column are the degrees of freedom.
- The values in the table are the critical values.

So for a probability α from the top, leading to a number a from the table, it means that

$$P[T > a] = \alpha,$$

for the particular number of degrees of freedom. For example, for a probability of 0.025, for 5 degrees of freedom, the table gives the number 2.571. This is shown in the orange coloured cell. This result means that, for 5 degrees of freedom:

$$P[T > 2.571] = 0.025.$$

The table shown is used directly for a one-tailed test. For a two-tailed test, the level of significance is divided by 2. It is possible to get a separate table for two-tailed tests, but this is unnecessary.

4.8.2 Rainfall example

A weather station has recorded daily rainfall amounts at a particular location on the east coast of Ireland. A statistical test is to be carried out to see if this data supports a claim that the mean rainfall for this region is 700ml. The amounts are listed here, measured in millilitres:

706	705	746	695	691
729	693	723	705	704
663	679	719	675	647
717	740	736	721	717
711	716	718	750	760
716	722	697	716	724
700	699	708	683	715
695	693	680	709	725

The steps are essentially the same as those for the z-test:

1. Framing the hypotheses: the Null hypothesis for a test on a mean states the claim for the true mean is correct: in this case, $\mu = 700$. The alternative hypothesis states that it is incorrect: $\mu \neq 700$.
2. Decide on a level of significance; we will take 0.01.
3. Decide if the test is one- or two-tailed. Since the alternative hypothesis does not specify any direction, the test is two tailed and so we divide the 0.01 by 2 and use 0.005 for the table.
4. Now find a critical value of the distribution. The number of data values given was 40, so we deal with 39 degrees of freedom. This gives the value 2.708, so our critical values are ±2.708.
5. Calculate the sample mean $\bar{x}$ and the sample standard deviation from the data; in this case they are:

   $$\bar{x} = 708.8,\ s = 22.858.$$

   From this, calculate the value of t:

   $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = (708.8 - 700)/(22.858/\sqrt{40}) = 2.407.$$

6. The value found for the test statistic is compared to the critical values; since our result for t, 2.407, is inside the range –2.708 to 2.708, the Null hypothesis is not rejected.

For this case, it is worth noting that the number of data points, 40, is quite high. Most statisticians would use critical values from the *z* distributions for this test (which would be 2.241), do the same calculations, and ignore the degrees of freedom.

4.8.3 Example – Engineering salaries

A claim has been made in a national newspaper that the mean salary of Computer engineers, five years after graduation, is €58,000 or higher. The claim is to be tested, at a significance level of 0.05. The following data was collected to investigate this claim; it is the salaries of 15 people in this position, in thousands of euro.

51.7	50.6	50.1
60.6	55.0	52.3
44.1	61.7	63.4
46.8	56.0	61.1
47.3	59.2	53.2

The test being carried out here is on the claim that the true mean of these salaries is 58K or higher. Therefore

- The Null Hypothesis is: $\mu = 58$.
- The alternative hypothesis is: $\mu < 58$.

Because the question has a direction, it means it is a one-tailed test. The level of significance is given as 0.05 and the number of degrees of freedom is 15 – 1 = 14; the tables give a value of 1.761. Recall the symmetry of the *T* distribution so that, for 14 degrees of freedom:

$$P[T < -1.761] = 0.05.$$

The critical value is –1.761. Now do the calculations to find t. From the data, the sum of the values is 813.1, giving a sample mean of 54.207. The sum of the squares is 44,578.79, and using this in the equation for the standard deviation gives s = 5.996. Finally, putting these values into the equation for t gives:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = (54.207 - 58)/(5.996/\sqrt{40}) = -2.450.$$

Now this value is below the critical value of –1.761, therefore we can reject the null hypothesis; the mean salary after 5 years for the Computer Engineers is indeed less than 58K.

4.8.4 The Confidence Interval for Small Samples

We restate here the concept of a confidence interval for small sample sizes, using the t-distribution. For a level of significance value α, define a number $t(n–1,\alpha)$ such that

$$P[T > t(n-1,\alpha)\ ] = \alpha,$$

where the variable T follows the t-distribution, with n – 1 degrees of freedom. This is of course the critical value for a level of significance value α, for the t-distribution. Then for the usual meanings of the symbols used for sample parameters from a normal random variable, the $(1 - \alpha)$% confidence interval is given by

$$P\left[\overline{X} - \frac{t(n-1,\frac{1}{2}\alpha)}{S/\sqrt{n}} < \mu < \overline{X} + \frac{t(n-1,\frac{1}{2}\alpha)}{S/\sqrt{n}}\right] = 1-\alpha \cdot$$

The confidence interval is then

$$\overline{X} - \frac{t(n-1,\frac{1}{2}\alpha)}{S/\sqrt{n}} \text{ to } \overline{X} + \frac{t(n-1,\frac{1}{2}\alpha)}{S/\sqrt{n}}.$$

If the assumption about the variable X being normally distributed is satisfied, then this is an exact confidence interval.

4.9 Differences

We now extend the t-test on a mean to another form of statistical test that measures results in before/after or with/without tests or questions. These very important tests crop up throughout science and engineering. Here is an example from the field of health.

4.9.1 Example – Smoking is Bad For You

A test is conducted to see if smoking a cigarette has an immediate detrimental effect on a person's aerobic capacity. To answer this question, the following data was collected from a group of 12 people:

- For each person, their peak flow lung function (PFLF) value was measured,
- The subject then smoked a cigarette,
- A second peak flow value was measured 20 minutes after the subject smoking.

The values are shown in the following table:

Before	*After*
619.92	616.20
633.24	632.47
674.64	673.43
622.44	621.28
677.52	677.51
672.84	668.14
633.96	630.22
615.60	611.72
668.16	663.22
660.24	658.00

Carry out a statistical test to see whether there has been a decrease in peak flow value, at a level of significance of 0.005.

In this case study, the values for the *same* subjects are being compared, before and after. To see if there has been a significant change, the changes in PFLF for each subject are tested to see if they are significantly different from zero.

This type of test, which is essentially a t-test on a mean applied to some differences, is called a paired-sample t-test.

For the n subjects, let R_1 to R_n be the values recorded before the cigarette and let S_1 to S_n be the values recorded after. Then the differences are simply

$$D_i = S_i - R_i$$

If there is a change in PFLF for a given subject *i* the difference D_i will be non-zero. If there has been a change overall, then the mean of the differences should be non-zero.

This points to the way to decide if there has been a change overall; carry out a similar test to the ones we have looked at on the random variable of the difference *D*, to see if the mean is zero. Let μ_D be the true underlying mean of the random variable of the difference D_i. The Null and alternative hypothesis are:

- The Null Hypothesis is that there is no change; the mean of the differences is zero. Mathematically, $\mu_D = 0$.
- The Alternative hypothesis is that the resistances have changed; the mean of the differences is not zero. Mathematically, $\mu_D \neq 0$.

Now let $\overline{D}$ be the sample mean of the differences, let s_D be the sample standard deviation of the differences and let *n* be the sample size. Then the *T* variable for this test will be:

$$T = \frac{\overline{D} - \mu}{S_D / \sqrt{n}}.$$

This is the same expression as before for *T*, it is just specifically referring to the variable in question, the difference *D*. We are of course assuming that the variables *R* and *S* and their difference are normally distributed. Here are the details. The Hypotheses are framed first:

- The Null Hypothesis is that the mean difference is zero.

- $H_0 : \mu_D = 0$,

- The Alternative hypothesis H_A is that the mean is above 0:
  - $H_A : \mu_D > 0$.

In this case, we are investigating a decrease in the peak flow, so this will be a one-tailed test. The critical value at a significance of 0.005 for 9 degrees of freedom is 3.250, for a one tailed test. In other words, for the variable T,

$$P[T > 3.25] = 0.005$$

From the data collected, the differences D_i are calculated and:

- The sum of the differences is 26.37, the sum of squares is 97.794.
- The mean is $\overline{D} = 2.64$, and $s_D = 1.772$.

The variable t is found to be 4.77, from this data. The value of t exceeded the critical value of 3.250, and so we reject the Null hypothesis. The data suggests smoking does lead to a decrease in aerobic capacity.

The alternative to choosing this direction of subtraction, ‘before minus after’, would have been to proceed in the normal manner of ‘after minus before’. In this context, a decrease in Peak flow values would appear as a negative value for the mean, so we would then use –3.250 as the critical value, since, for nine degrees of freedom, $P[T > 3.25] = 0.005$ means that $P[T < -3.25] = 0.005$.

4.10 Goodness-of-Fit of a Frequency Distribution

A question that frequently comes up in statistical analysis is that of the distribution underlying a given data set. If we are told that a random variable, for example, heights, is normally distributed, then given a set of data, regarded as a sample from a random variable, we wish to carry out a test on whether the data came from a particular distribution. More precisely, we will test whether the underlying random variable, from which the samples are taken, has a particular distribution with certain parameters.

Like in previous cases, it will be instructive to see where the test comes from. Typically we will construct a frequency distribution from our data; the frequencies will be a set of observed values O_i. Under the Null Hypothesis that the underlying random variable has a particular distribution, we will generate corresponding expected values E_i, for each group. The difference between the two sets of values will be measured and tested by the quantity given by the equation

$$\sum_i \frac{\left(O_i - E_i\right)^2}{E_i},$$

where the summation is done over every pair of corresponding values. Let us see where this originates.

Consider the case where a distribution assigns probabilities $p_1, p_2, \ldots, p_k$, to k possible outcomes. Naturally it must be the case that $p_1 + p_2 + \ldots + p_k = 1$. This case is an extension of the binomial distribution of $k = 2$. The distribution that handles this more general case is called the multinomial distribution. Let $(r_1, r_2, \ldots, r_k)$ be set of counts with total n, that is, ways in which n outcomes of the experiment are assigned to the k outcomes. So

$$r_1 + r_2 + \ldots + r_k = n.$$

The probability that this particular set of counts $(r_1, r_2, \ldots, r_k)$ come up is then given by the multinomial expression

$$\frac{n!}{r_1! r_2! \ldots r_k!} p_1^{r_1} p_2^{r_2} \ldots p_k^{r_k} = \frac{n!}{r_1! r_2! \ldots r_k!} \prod_{i=1}^{k} p_i^{r_i}.$$

We saw in the case of the binomial distribution that the count r for one of the outcomes, in other words the 'successes', eventually forms a normal distribution with mean np and variance $np(1 - p)$, so that the quantity

$$\frac{r - np}{\sqrt{np(1-p)}} \sim \mathrm{N}(0,1).$$

In a similar vein, if we looks at the vector of similar quantities

$$\left(\frac{r_1 - np_1}{\sqrt{np_1}}, \frac{r_2 - np_2}{\sqrt{np_2}}, \ldots, \frac{r_k - np_k}{\sqrt{np_k}} \right),$$

we can show that they are k standard normal variables, but not independent; one can be constructed from the others from the relations

$$p_1 + p_2 + \ldots + p_k = 1, \text{ and } r_1 + r_2 + \ldots + r_k = n.$$

This is reflected in the fact that the denominators are not quite 'correct'. In fact we have $k - 1$ independent standard normal variables in this vector, as n increases. With this in mind, we can look at the length of this vector:

$$\left(\frac{r_1 - np_1}{\sqrt{np_1}}\right)^2 + \left(\frac{r_2 - np_2}{\sqrt{np_2}}\right)^2 + \ldots + \left(\frac{r_k - np_k}{\sqrt{np_k}}\right)^2.$$

This is the sum squares of $k - 1$ independent standard normal distributions, so therefore it follows the χ^2 distribution with $k - 1$ degrees of freedom. The quantity is then

$$\frac{(r_1 - np_1)^2}{np_1} + \frac{(r_2 - np_2)^2}{np_2} + \ldots + \frac{(r_k - np_k)^2}{np_k}.$$

Noting that the r_i are the observed values and the np_i are the predicted values from the Null hypothesis that the distribution assigns probabilities p_1, p_2, …, p_k, to k possible outcomes, we have arrived at the idea that

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2 \text{ with } k - 1 \text{ d.o.f.}$$

It is worth noting assumptions that were made along the way.

1. The samples of the random variable were all taken from the same distribution and were equally likely to be chosen.
2. The value of n is reasonably large; in particular, it can be shown that the need for convergence to standard normal variables within the derivation means that it is important that no value of np_i is less than 5, or that for large k, no more than one or two are below 5.

3. There was a reduction in one degree of freedom from the relation $p_1 + p_2 + \ldots + p_k = 1$. If a number p of parameters have to be calculated from data, there is a corresponding reduction in the degrees of freedom.

It is instructive to repeat the above work for the case $k = 2$, so that we are dealing with one independent cell and the full algebraic work can be done. This does, however, reduce to the case of a z-test on a proportion, since a chi-square distribution with 1 degree of freedom is N(0,1).

We now restate this theory as a Null-hypothesis based test, keeping in mind the assumptions we noted above.

4.10.1 The Test for a Frequency Distribution

Let X be a random variable, which is suspected of having a particular distribution with p parameters. Let $X_1, X_2, \ldots X_n$, be a set of n values sampled from this variable. Assume also that the number n is reasonably large. The Hypothesis test that the variable follows a given distribution goes as follows:

- Framing the hypotheses: The *Null hypothesis* states that the variable X follows the suspected distribution, the *alternative hypothesis* states that it does not.
- Decide on a level of significance; as before, this is usually 0.05 or 0.01.

- The degrees of freedom for the χ^2 distribution is $k - p - 1$, where p is the number of parameters for the proposed distribution calculated from the data and k is the number of groups into which the data will be divided.
- From this, identify the critical values for variable χ^2. This will be a two-tailed test, since the occurrence of the observed values all being too close to the expected values would be improbable.
- Create a frequency distribution from the data $X_1, X_2, \ldots X_n$, in other words, divide it into k groups. Let O_i be the observed values of the frequencies in each group.
- Now calculate the corresponding expected values E_i using the proposed distribution for variable X under the Null Hypothesis.
- With these values, calculate the sample value of

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

- If the value found for the statistic is found to be higher than the critical value, then the Null hypothesis is rejected.

4.10.2 Example – Testing for a Normal Distribution

The random variable H of Height in Men is often described as having a Normal distribution. Carry out a statistical test to see whether the following frequency distribution is consistent with this claim. Here is the data:

Height	Frequency
1.55m to 1.60m	4
1.60m to 1.65m	12
1.65m to 1.70m	37
1.70m to 1.75m	51
1.75m to 1.80m	24
1.80m to 1.85m	9
1.85m to 1.90m	1

Recall our convention that the group '1.65m to 1.70m' means those heights that are above or equal to 1.65m and below 1.70m, and so on. The Hypothesis test goes as follows:

- Framing the hypotheses: The *Null hypothesis* states that the variable X follows the normal distribution, the *alternative hypothesis* states that it does not.
- We will take a level of significance as 0.05.
- The normal distribution has two parameters and the total number of observed values we have is 7. However, when we calculate the probabilities for these groups, their sum will be less than 1. We therefore need to include another group for 'outside the range'. The degrees of freedom is then $8 - 2 - 1 = 5$.
- From the tables, the critical values are 0.831, 12.833.

The required frequency distribution is already available, in other words, the data X_1, X_2, … X_n has been divided it into groups. Let O_i be the observed values of the frequencies in each group.

We must now come up with expected values for the groups; work out the two parameters of the distribution, the sample frequency mean and the sample frequency distribution:

$$\bar{x} = 1.715,\ s = 0.058.$$

Now calculate the expected values; start with the first group, 155 to 160. The probability that a height will fall in this range is

$$P[1.55 < H < 1.60] = P[H < 1.60] - P[H < 1.55].$$

Using

$$P[H < 1.55\text{m}] = 0.002,\ P[H < 1.60\text{m}] = 0.024.$$

This means that the probability a value of the height is in this group is

$$P[1.55 < H < 1.60] = 0.024 - 0.002 = 0.022,$$

and so the expected number of values in the group is

$$138\text{x}0.022 = 3.036.$$

We will continue in this way, calculating a probability for each group, and multiplying by n to see the expected value for that group. The details are set out below in a table.

Height group	Probability for group	Expected Value
1.55m to 1.60m	0.022	2.980
1.60m to 1.65m	0.108	14.890
1.65m to 1.70m	0.267	36.873

1.70m to 1.75m	0.329	45.377
1.75m to 1.80m	0.201	27.769
1.80m to 1.85m	0.061	8.437
1.85m to 1.90m	0.009	1.268
1.90m		

The last expected value is required, that of those outcomes not part of the range of the data. This is easily calculated because its observed value is 0, therefore the term is E_8 itself and this is simply the remaining probability times n, or $n - \sum E_i$. In this case it is 138 – 137.594 = 0.406. Now that all the expected values are available, the value of the statistic can be calculated. The equation is:

$$\sum_i \frac{(O_i - E_i)^2}{E_i} =$$

$$= \frac{(4-2.98)^2}{2.98} + \frac{(12-14.89)^2}{14.89} + \ldots + \frac{(1-1.268)^2}{1.268} + 0.406 =$$

$$= 0.349 + 0.561 + 0.0 + 0.697 + 0.512 + 0.038 + 0.057 + 0.406 = 2.619.$$

This figure is within the range of critical values, so we do not reject the Null Hypothesis; there is no evidence to suggest, from this data set, that the height of men is not Normally distributed.

4.10.3 Another Test for a Normal Distribution

Carry out a test on the following diameters of discs produced by a milling machine to see if the variable diameter follows a normal distribution.

Diameter Group	***Frequency***
200 to 205	12
205 to 210	7
210 to 215	16
215 to 220	30
220 to 225	35
225 to 230	30
230 to 235	12
235 to 240	5
240 to 245	2
245 to 250	1

Solution:

Firstly frame the hypotheses:

- The *Null hypothesis* states that the variable X follows the normal distribution, the *alternative* states that it does not.
- We will take a level of significance as 0.05.
- The normal distribution has two parameters, and the number of observed/expected values we have is $k = 10 + 1$, so the degrees of freedom is $11 - 2 - 1 = 8$.
- From the tables, the critical values are therefore 2.180 and 17.535.

The required frequency distribution is already available, in other words, the data D_1, D_2, … D_n has been divided it into groups. Let O_i be the observed values (i.e. frequencies) in each group.

To carry out the calculations, start by finding the frequency sample mean and frequency standard deviation:

$$\bar{x} = 220.867, s = 9.278.$$

Now calculate the expected values.

Diameter Group	***Frequency***	Expected Value	$(O_i - E_i)^2/E_i$
200 to 205	12	4.7042	11.3154
205 to 210	7	11.5694	1.8047
210 to 215	16	21.4255	1.3739
215 to 220	30	29.8812	0.0005
220 to 225	35	31.3862	0.4161
225 to 230	30	24.8289	1.0770
230 to 235	12	14.7923	0.5271
235 to 240	5	6.6363	0.4035
240 to 245	2	2.2416	0.0260
245 to 250	1	0.5700	0.3244
250			

When the figures in the last column are totalled up, the result is 17.2685. Add in the last quantity, which is $150 - \sum E_i$, which exceeds the critical value. The Null hypothesis is therefore rejected.

To see why, observe the first value; it is a high value compared to the next, instead of tailing off as would be expected in a Normal distribution. It is much higher than the expected value.

4.10.4 Categories and Tests for Independence

Consider the following example – a lecturer suspects that mature students taking a particular module are doing better in examinations than the other students. In order to study this question, the students are divided into three groups:

A. Those coming straight from school,
B. Those who waited up to 5 years before attending college,
C. Those who waited more than 5 years before attending college.

The exam results are also divided into three categories; fail, pass and merit. For a particular class and year, the results of the examinations are counted for each group; this information is given in the following table. These numbers are called the observed values.

Age grp.	Merit	Pass	Fail
A	11	16	11
B	5	6	1
C	7	2	0

Deciding whether or not mature students are doing better overall means comparing whether these results are different from what would be expected if the grades were distributed in proportion. This immediately suggests a test on the counts in each table following a distribution based on the Null Hypothesis that they are assigned in proportion.

The proportion of students in each category will have to be calculated from the data; specifically the sub-totals across and down the categories:

Age grp.	Merit	Pass	Fail	
A	11	16	11	38
B	5	6	1	12
C	7	2	0	9
	23	24	12	59

Rather than directly calculate the proportions, the following quick method yields the expected values. For example, the number of students expected to fail from group A is the 12 fails by 38/59, or alternatively the 38 in group A by 12/59. Either way, the expected number of Group A fails would be

$$\frac{12 \text{ x } 38}{59}$$

The table of expected values is shown here:

Age grp.	Merit	Pass	Fail
A	14.81	15.46	7.73
B	4.68	4.88	2.44
C	3.51	3.66	1.83

As before, call each of the observed values O_i and call the expected values E_i:

$$\sum_i \frac{(O_i - E_i)^2}{E_i}.$$

The distribution of this variable will be χ^2; we must identify the degrees of freedom. If r is the number of rows and c is the number of columns, then using the data we have clalculated $r - 1$ and $c - 1$

parameters. The total number of groups is rc so the degrees of freedom is given by

$$rc - (r - 1) - (c - 1) - 1 = (r - 1)(c - 1).$$

In this instance we have 4 degrees of freedom, so our critical values are 0.484 and 11.143. The value of the test statistic is 9.56, within the range. We do not reject the null hypothesis. However, it should be noted that several of the cells had expected values below 5, therefore we would not regard this as a rigorous result.

4.10.5 The Framework of a χ^2 Test for Independence

In this type of hypothesis test, a possible link or connection is being investigated between two quantities or categories. Appropriate data will have been collected, counts in various categories in a contingency table.

- Framing the hypotheses: The first step is to identify, and then state, the situation where there is no link or connection between the two variables or categories; in other words they are independent. The *Null hypothesis* states that the variables involved are independent. The *alternative hypothesis* states that there is a link.
- Decide on a *level of significance*.
- Identify the degrees of freedom – for a contingency table of values for a test of independence, it is $(r - 1)(c - 1)$.
- Find the critical values of the distribution.

- Calculate the expected values if all the counts of passes and fails of students in each category were in proportion.
- If the value found for the statistic is found to be an unlikely value, then the Null hypothesis is rejected and the alternative is accepted.

4.10.6 Example – Production Processes

A company is producing components according to three different processes in 4 separate factories. The number of components which fail is recorded during one week. Determine, using a chi-squared test, whether or not the three processes are being used differently in the four factories. The values are given in the contingency table shown:

Process	*A*	*B*	*C*
Factory 1	12	21	24
Factory 2	23	34	31
Factory 3	32	54	17
Factory 4	41	45	13

The test proceeds as follows:

- Framing the Hypotheses: The null hypothesis is that there is no link between the process used and the factory it is used in. The alternative hypothesis is that there is a link.
- We will take the level of significance as 0.01.

- The number of degrees of freedom is (4 – 1)(3 – 1) = 6, and so using the tables, the critical values for these parameters are 0.676 to 18.548.
- The value found is 28.10, well in excess of the upper limit. The null hypothesis is therefore rejected, and we conclude there is a connection between the manufacturing processes and how they are used in each factory.