Statistical Testing

Damian Cox

April 3, 2019

Contents

1 Introduction

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. Before we look at hypothesis testing and how it is done, we must look again at the idea of a statistic.

1.1 The idea of a Statistic

So far we have been using the word 'statistics' to mean 'descriptive statistics'; calculating certain numbers from a set of data in order to describe or summarise that data. Examples include the mean and standard deviation, median and quartiles.

Another important point that arises here is 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, coming from the idea of an expected value. If we treat the values listed in a dataset as being actual values of the random variable, obtained when the experiment is run many times, then the calculated values of the parameter may be regarded as an estimate of a true underlying figure. We say therefore that the set of actual values is a sample from the random variable.

We saw an example of this when we studied the normal probability distribution; we use figures for the mean and standard deviation that would come from the calculations we saw when we first looked at descriptive statistics. For example, if we measure the height of several hundred Irish men and come up with a mean of 1.71m, we can regard this figure of 1.71m as an estimate of the true underlying mean $\bar{H}$ for the height of the Irish male population. Therefore the figure is an estimate, one based on a random sample of heights.

This is an extremely important observation and is the basis of this topic of hypothesis testing. In the case of heights, each measurement of the variable height is based on a random selection of subjects, and so each measurement of height is a random variable. Because a figure such as the mean is calculated from a set of random variables, that is, the actual heights recorded, then so also is the estimate of the mean.

For example, say we calculate an estimate for the mean of the variable height H based on n values. When 20 subjects are selected and their height measured, this is 20 values of the random variable H. Then the mean is given by the familiar equation

$$\bar{H} = \frac{H_1 + H_2 + \ldots + H_n}{n}.$$

This means that the value of the mean $\bar{H}$ is based on the values $H_1, H_2, \ldots, H_n$, of the random variable H, so in fact H is itself a random variable. Alternatively, we could view the values $H_1, H_2, \ldots, H_n$ as being n random variables, but all with the exact same normal distribution with the same mean and standard deviation. This kind of random variable, calculated from other random variables, is the real meaning of the word *statistic*.

1.1.1 Definition: 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 sample values of a random variable. 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. 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.

1.2 Hypothesis Testing

Hypothesis testing is the process of making decisions based on statistics; a statistic is calculated and its distribution is known, therefore it is known whether the value obtained is unusual or not and decisions are made on this basis. Hypothesis testing is extremely important in scientific, engineering and business studies; the concept that a decision can be made, from numerical data, in the probability context, is a powerful idea. You can be sure that almost every time you see a result quoted with a 'margin of error' or referred to as a 'significant result', there was a hypothesis test behind it. It is also the reason statistics is so important; while descriptive statistics is quite useful, it is the use of the parameters such as means and standard deviations in hypothesis testing that make it so vital. It essentially underpins the correct use of the scientific method.

1.2.1 Two (Re)definitions

Before moving on to an example, take note of the following re-definitions and terminology. Let X be a random variable, for which μ is the true underlying mean and σ is the true underlying standard deviation.

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

  $$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

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

- Let the symbol S be the estimate of the standard deviation

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

  This is now called the *sample standard deviation.*

We will now look at an example of a test on a mean.

1.3 A Starting Example

The simplest statistic to be used in a Hypothesis test comes directly from 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.

1.3.1 First Example

The following are the measurements of the resistance, in kilo-Ohms, of a sample of 55 components being produced by a manufacturing process. The measurements are all in kilo-Ohms.

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

It is proposed that the value of the true underlying mean is 24 kilo-Ohms. The question to be addressed is how well does this result for the mean represent the true mean of the wider overall population. We will firstly calculate the sample mean for this set of data.

Let $n = 55$ be the number of data values in the list. The sample mean of these figures is then

$$\bar{X} = \frac{\sum_i X_i}{n} = \frac{1375.62}{55} = 25.0113.$$

This is close to the proposed figure of 24 kilo-Ohms for μ, but we must decide if it close enough to stick with 24 kilo-Ohms or different enough that we should change our estimate. 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 the value $\mu = 24$ is correct. This is done with a particular random variable Z given by the equation:

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

Because Z is calculated from the random variables $\bar{X}$ and s, which in turn were calculated from the data we have collected, it is a random variable. It is therefore a *statistic*, in the sense defined above.

This variable Z is used because it is in fact the standard normal variable, when the number n is very large; it has a normal distribution with mean 0 and standard deviation 1. The basis for this claim is the *Central Limit Theorem*, which we will see later.

Working out the value of Z is straightforward. We first need S, the standard deviation. The calculation is shown here:

$$\sum_i X_i^2 = 34,905.44.$$

Then S is given by

$$S^2 = \frac{34,905.44 - 55 \times 25.0113^2}{54} = 9.2488, \Rightarrow s = 3.0412.$$

Then our value of Z is:

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{25.0113 - 24}{3.0412/\sqrt{55}} = 2.466.$$

Now the question arises whether 2.466 is an unusual value to come up in for a standard normal distribution.

Let us start by choosing a probability that we consider to be low, in other words, representing an unusual event. The value of 0.05, which is equivalent to 1 in 20, is frequently used in this context. We will now look at an event involving Z with this probability, bearing in mind that for a continuous variable, we only talk in terms of ranges.

The true mean could be higher or lower than the proposed value of 24 kilo-Ohms, in other words, 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 of 0.05. Recall that this means finding a number a such that (remembering that Z is symmetric)

$$P[Z < -a] = P[Z > a] = 0.025.$$

This makes our test a *two-tailed test*; we will review this idea later. From the tables the values are ± 1.96; in other words, we can note the probability of two events:

$$P[Z < -1.96] = P[Z > 1.96] = 0.025.$$

In summary, the probability of getting a Z value between the two numbers ± 1.96 is 0.95, whereas the probability of a value either above 1.96 or below -1.96 is 0.05. So therefore a z-value outside the range -1.96 to 1.96 qualifies as an unusual event, inconsistent with the idea that the true mean is 24 kilo-Ohms.

These values of a test statistic, which quantify what is regarded as an unusual event for that statistic, are called *critical values*. The value we found for z, 2.466, 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 kilo-Ohms calculated from the data for the mean, is unusual enough to set aside our claim of 24 kilo-Ohms.

2 The Structure of Hypothesis Tests

The analysis carried out in the previous section on the estimate for the mean is an example of a Hypothesis Test. More specifically, it is a Hypothesis test on an estimate for a population parameter, in this case the mean. The reasoning behind the analysis and conclusion will now be put on a more formal basis; this section will cover the fundamental ideas and steps behind this method.

2.1 The Framework of a Hypothesis Test

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 arises, given that the sample will be finite and sometimes quite small, 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. It is the default starting point. 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 number between 0 and 1 which is regarded as an appropriately low probability. In most cases this is taken as 0.05, that is, one chance in 20.

3. Decide if the test is one- or two-tailed. If the claim of the population parameter can only be invalidated by an unusually high value of the sample parameter or alternatively only by an unusually low value, then the test is one-tailed. If it could be invalidated by values either to high or too low, the test is two-tailed.

4. We now measure, with a statistic with a known distribution, 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 will then allow us to decide whether the value calculated for the statistic is unlikely or not. This means that the full title of our method here is a *Null Hypothesis Based Statistical Test.*

5. Now find a critical value of the distribution; this is a value with which we can compare the test 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. 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. This type of test is called a z-test, because it uses the standard normal distribution to decide on whether a proposed value for a mean can be accepted.

We will now recast the two versions of the analysis of the resistances example, firstly as a two-tailed and secondly as a one-tailed test.

2.1.1 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 kilo-Ohms for the true mean, or not.

1. 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 situation here is that the true mean is not 24 kilo-Ohms, this is the Alternate Hypothesis:

$$H_0 : \mu \neq 24.$$

2. The test statistic is the variable

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

   which follows the standard normal distribution.

3. We will choose a level of significance of 0.05.

4. Since the Null Hypothesis could be invalidated by a higher or lower value of the sample mean, then we use a two-tailed test. This means we must divide the 0.05 significance level by 2, to give 0.025. From the z-tables, we know that

$$P[Z < -1.96] = P[Z > 1.96] = 0.025.$$

   The values ± 1.96 are our critical values. Then if we get a Z value in this range, it means that nothing unusual has happened, whereas if we get a value outside this range, the data was unusual if the Null Hypothesis is true.

5. Carrying out the calculations, the value of Z was 2.466.

6. 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.

2.1.2 A One-Tailed test

For the second version of the test on the resistances, we now ask the question if the data collected supports the claim that the true mean is at most 24 kilo-Ohms, or if the value of μ is higher. This type of analysis may arise if we have reason to think that there is no possibility of the true mean being lower. For example, if we are coming to this analysis having already seen the sample mean, we can rule out the possibility that $H_0 < 24$. Checking the data before setting up the framework is called *data-snooping.*

Here is the test redone as a one-tailed test.

1. The Null Hypothesis H_0 is that the true mean for the resistances of the components is 24:

$$H_0 : \mu = 24.$$

The alternative situation here is that the true mean is greater than 24 kilo-Ohms, this is the Alternate Hypothesis:

$$H_0 : \mu > 24.$$

2. The test statistic is the variable

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

which follows the standard normal distribution.

3. We will choose a level of significance of 0.05.

4. Since the Null Hypothesis could be invalidated by a higher sample mean, then we use a one-tailed test. From the z-tables, we know that

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

The value 1.65 is our critical value. Then if we get a Z value below 1.65, it means that nothing unusual has happened, whereas if we get a value above this, the data was unusual if the Null Hypothesis is true.

5. Carrying out the calculations, the value of Z was 2.466.

6. This value is above the critical value, 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.

2.1.3 Another Test on a Mean

The following are the lifetimes, in units of 10^6 seconds, of a selection of components produced by a manufacturing facility. Test the claim that the true mean is 25×10^6 seconds.

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	26.50	26.93	20.75	29.85	25.65	24.56	28.99
28.17	28.23	28.71	34.40	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		

Here is the relevent test.

1. The Null Hypothesis H_0 is that the true mean for the lifetimes of the components is 25×10^6 seconds.

$$H_0 : \mu = 25,\ H_A : \mu \neq 25.$$

2. The test statistic is the variable

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

which follows the standard normal distribution.

3. We will choose a level of significance of 0.05.

4. Use a two-tailed test. This means we must divide the 0.05 significance level by 2, to give 0.025. From the z-tables, we know that

$$P[Z < -1.96] = P[Z > 1.96] = 0.025.$$

The values ± 1.96 are our critical values.

5. Carrying out the calculations. The sums are:

$$n = 50,\ \sum_i X_i = 1,303.77,\ \sum_i X_i^2 = 34,545.445.$$

With this, we find that

$$\bar{X} = 26.075,\ s = 3.348.$$

The value of z is then

$$z = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{26.075 - 25}{3.348/\sqrt{50}} = 2.272.$$

6. 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 lifetimes is 25×10^6 seconds.

2.1.4 Summary: The Steps in a 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 Null Hypothesis H_0 is that the true mean for the lifetimes of the components is correct. The Alternate is that it is not correct.

$$H_0 : \mu =, \;\; H_A : \mu \neq ?, \;\; \mu > ?, \;\; \mu < ?.$$

2. The test statistic is the variable

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

which follows the standard normal distribution.

3. Choose a level of significance, e.g. 0.05 or 0.01. This is often denoted as α.

4. If H_A is $\mu \neq ?$, then the test is two-tailed. This means we must divide the level of significance level by 2. From the z-tables, critical values $\pm a$ is identified so that

$$P[Z < -a] = P[Z > a] = \alpha/2.$$

The values $\pm a$ are our critical values. If the test is one-tailed, so the Alternate Hypothesis is $\mu >?$ or $\mu <?$, then do not divide by 2. The tables give a value

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

Then a or $-a$ is the critical value, whichever is required given the direction of the question.

5. Carrying out the calculations for the test value of Z.

6. If the value of Z is outside the range of the critical value(s), reject the Null Hypothesis.

2.2 The Central Limit Theorem

It is worth giving a brief explanation why this test works for the mean. Indeed, in these tests, which use a variety of statistics, it is worth knowing how the distributions are found. The basis of the z-test comes from the following fundamental theorem, called the Central Limit Theorem. The distributions for the other tests, such as the chi-square test and the t test, come from similar analysis of sums of variables with some knowledge of their distribution and that of their means. The theorem is stated here, with two more refinements that make it increasing powerful and useful.

2.2.1 Central Limit Theorem: Part 1

Let $X_1, X_2, \dots X_n$ be n random variables, all following a normal distribution with true underlying mean of μ and with standard deviation σ. Then the random variable of the mean of these n numbers is normally distributed, with underlying mean of μ and with standard deviation $\sigma/\sqrt{n}$.

This version of the theorem means that the variable Z given by

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

follows the standard normal distribution. As it stands, this theorem is only useful if we are dealing with data values we knew to be normally distributed and if we had their standard deviation. However, the next part of the theorem addresses this limitation.

2.2.2 Central Limit Theorem: Part 2

Let $X_1, X_2, \ldots X_n$ be n random variables, all following a normal distribution with true underlying mean of μ and with standard deviation σ. Then the random variable of the mean of these n numbers is normally distributed, with underlying mean of μ and with standard deviation converging on $s/\sqrt{n}$ when n becomes large.

This version of the theorem means that the variable Z given by

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

follows the standard normal distribution once n is reasonably large. Typically we will take $n = 30$ as being 'large'.

2.2.3 Central Limit Theorem: Part 3

Let $X_1, X_2, \ldots X_n$ be n random variables, all following a distribution with true underlying mean of μ and with standard deviation σ. Then the random variable of the mean of these n numbers is normally distributed, with underlying mean of μ and with standard deviation converging on $s/\sqrt{n}$ when n becomes large.

This final version of the theorem means that the variable Z given by

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

follows the standard normal distribution once n is large. Typically we will take $n = 30$ as being 'large'. We have also dropped the requirement that the variables $X_1, X_2, \ldots X_n$ follow the normal distribution.

A consequence of the increasing strength of these result is that the variable Z converges on the standard normal distribution relatively quickly as n increases if the distribution for the X_i are similar to a normal distribution, but more slowly if the distribution is not similar.

3 The Student T Test

The next step is to adjust the statistical test on means for cases where the sample size n is not large. We will first introduce a new distribution to deal with this situation.

3.1 The T Distribution

The statistic used for small sample sizes (with the sample standard deviation) is called the t-distribution, also known as Students 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 very similar to the standard normal distribution with one significant difference; it depends on an integer parameter called the *degrees of freedom* . The right-hand side of the distribution is shown here for several values of the degrees of freedom parameter. The standard normal distribution is included for comparison.

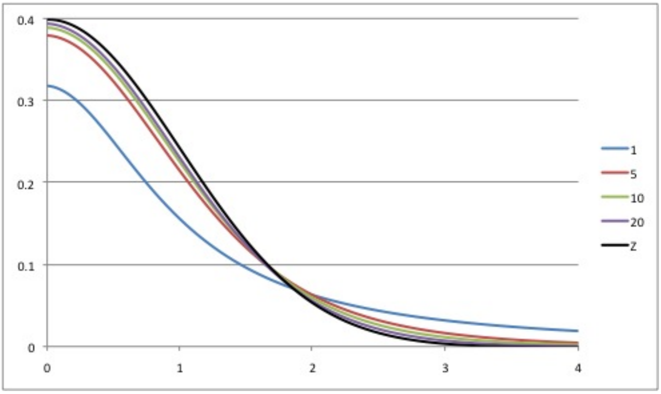

Figure 1: The PDF of the T- and Standard Normal Distributions

3.1.1 Comparing the T Distribution and the Normal

This distribution is very similar to the standard normal distribution with the significant difference of the dependence on the degrees of freedom parameter. For its use in a hypothesis test, this parameter is directly connected to the size of the sample, n. Specifically, when used in a test on a mean, the degrees of freedom parameter is $n-1$, where the mean is calculated from n values. The similarities and differences are listed here:

- The mean of a set of numbers with a t-distribution is 0, in the same way as the standard normal distribution.
- The distribution is symmetric. This is an extremely important and useful property of the distribution and we will use this frequently as is the case with the standard normal distribution.
- For a variable with the standard normal 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 the lower the value of n, the more likely the higher values become.

- For large values of n, the t distribution is very similar to the normal distribution. Indeed, as the second and third parts of the Central Limit Theorem tell us that for large values of n, this variable becomes the standard normal variable.

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

3.1.2 Comparison of the Z and T Tests

In the context of a Hypothesis test, the statistic T given by the equation:

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

where all symbols are as before, follows the t distribution, with $n - 1$ degrees of freedom. This is exactly the same expression as the one for the z-test; the adjustment made here is the distribution.

The properties listed above mean that for all intents and purposes, carrying out a z-test or a t-test are essentially the same process. In this document, a t-test will be used for sample sizes as high as the low 40s, though in practice, most scientists or engineers using statistics will use the z-tables for a sample size above 30.

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

- The values in bold across the top are the probabilities, that is to say, the level of significance
- 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.

3.2 The Test on a Mean for Small Sample Sizes

The distribution we have discussed is used, as suggested already, for a test on the mean when the sample size is small. We will set out the nature of the variable used for this test first and then give an example of the test.

3.2.1 The T Test on a Mean

For a normally distributed random variable X, we take a sample of n values. This means we are dealing with n normally distributed random variables, call them X_1, X_2 to X_n. Then the sample mean and sample standard deviation are given by the usual equations:

$$\bar{X} = \frac{\sum_i X_i}{n}, \; S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1} = \frac{\sum_i X_i^2 - n\bar{X}^2}{n-1}.$$

Then the variable T defined by

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

Degrees of freedom	Probabilities			
	0.05	0.025	0.01	0.005
1	6.314	12.706	31.821	63.657
2	2.92	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
16	1.746	2.120	2.583	2.921
...	...	...	...	...
39	1.685	2.023	2.426	2.708
40	1.684	2.021	2.423	2.704
41	1.683	2.02	2.421	2.701
42	1.682	2.018	2.418	2.698
43	1.681	2.017	2.416	2.695
44	1.680	2.015	2.414	2.692
45	1.679	2.014	2.412	2.690
46	1.679	2.013	2.410	2.687

follows the Student T Distribution at $n-1$ degrees of freedom.

Take note of the fact that, because we are dealing with a small sample size, we require that the original data values came from a normally distributed variable. We will now and carry out a test on rainfall data using this test.

3.2.2 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	693	723	705
717	740	663	679	719	675	729	704
695	693	680	709	647	736	721	717
700	699	708	711	716	718	750	760
725	716	722	697	716	724	683	715

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

- Frame 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; $\mu \neq 700$.

- Decide on a level of significance; we will take 0.01.

- 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.

- 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.

- Now do the calculations for the test value of T. Calculate the sample mean 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:

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

- The value found for the test statistic is compared to the critical values; since our result $t = 2.407$ is within 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 distribution for this test (which would be ± 2.241), do the same calculations, and ignore the degrees of freedom.

For this test it is also worth noting that if we had taken a level of significance of 0.05, the critical values would be ± 2.023 and therefore the Null Hypothesis would have been rejected.

3.2.3 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 euro or higher. The claim that the starting salaries are 58K or higher 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 steps in the test are :

- 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 alternate hypothesis is: $\mu < 58$.

- Because the question has a direction, it means it is a one-tailed test; the claim can only be called into question by a sufficiently low sample mean, not by a very high sample mean.

- 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. Keep in mind that this means that, for 14 degrees of freedom:

$$P[T > 1.761] = 0.05.$$

  We will be watching out for a value of $\bar{X}$ that is too low, in other words, a negative value of of T. We therefore need the negative critical value. Since the T Distribution is symmetric like the standard normal distribution, we can say that

$$P[T > 1.761] = 0.05 \Rightarrow P[T < -1.761] = 0.05.$$

  The critical value is therefore -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, so using this in the equation for the standard deviation gives s = 5.996. Finally, putting these values into the equation for t gives:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{54.207 - 58}{5.996/\sqrt{15}} = -2.450.$$

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 according to this data.

3.3 Tests on Differences

We now move on 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. As we will see, the test is simply just an extension of the test on a mean. Here is an example from the Cold War.

3.3.1 Before and After EMP

It is suspected that an Electromagnetic Pulse (EMP) of a certain strength will affect the workings of an electronic component. It has been decided that measuring their resistance R will provide a way of assessing this. A set of components were chosen and their resistances measured both before and after being subjected to the EMP.

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

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

For the set of components, let R_1 to R_n be the values recorded for the n components, before the pulse. Let S_1 to S_n be the values recorded after the pulse. Then the differences are simply

$$D_i = S_i - R_i.$$

If there is a change in resistance, for a given component 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: $\mu_D = 0$.
- The Alternative hypothesis is that the resistances have changed, the mean of the differences is not zero: $\mu_D \neq 0$.

Now let $\bar{D}$ 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:

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}.$$

This is the same expression as before for T, it is just specifically referring to the variable in question, the difference D. Here is a worked example.

3.3.2 Example: EMP

For the set of components described above, let R_1 to R_n be the values recorded for the n components, before the electromagnetic pulse. Let S_1 to S_n be the values recorded after the pulse. Here are the values from the experiment:

$$\text{Before: } 13.73,\ 12.47,\ 13.95,\ 13.75,\ 12.71,\ 13.68,\ 12.33.$$

$$\text{After: } 13.57,\ 12.48,\ 14.81,\ 13.62,\ 12.23,\ 13.12,\ 11.54.$$

The differences are

$$D_i = S_i - R_i.$$

Let μ_D be the true mean of the variable D. Frame the hypotheses:

- The Null Hypothesis is: $\mu_D = 0$.
- The Alternative hypothesis: $\mu_D \neq 0$.

Continuing the steps in the test: Choose a level of significance of 0.01 and dividing by 2 for a two-tailed test gives 0.005. From the tables, the critical value for T at 6 degrees of freedom is 3.707. Therefore our critical values are ± 3.707. Now calculate the differences; they are, in the same order as above,

$$0.16,\ -0.01,\ -0.86,\ 0.13,\ 0.48,\ 0.56,\ 0.79.$$

For this list of 7 numbers, their sum is 1.25 and the sum of squares is 1.950. Therefore $\bar{D} = 0.18$ and $s_D = 0.537$. Then the t value for this test will be:

$$t = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} = \frac{0.18 - 0}{0.537/\sqrt{7}} = 0.881.$$

Thus the Null hypothesis is not rejected; there is no significant increase or decrease in the values of the resistance.

3.3.3 Example: Changes in Capacitance

It is suspected that a particular treatment changes the capacitance and therefore affects the working of a metal component in an aircraft engine. 10 components were given the treatment, with their capacitances measured before and after. The values are given here:

Before:

123.45, 120.14, 125.22, 120.39, 124.30, 121.28, 121.53, 121.60, 122.15, 122.63

After:

124.97, 120.53, 124.27, 122.25, 125.97, 122.59, 123.37, 120.37, 123.19, 124.97

Test the hypothesis that the capacitance has changed, using 0.05 as the level of significance.

Let μ_D be the true mean of the variable D.

- Frame the hypotheses; the Null Hypothesis is: $\mu_D = 0$, the Alternative hypothesis: $\mu_D \neq 0$. This is a two-tailed test.

- The level of significance is 0.05, dividing by 2 for a two-tailed test gives 0.025.

- From the tables, the critical value for 9 degrees of freedom is ± 2.262.

- The sample mean of the differences is 0.98. The sample standard deviation is

$$s_D^2 = \frac{\sum_i D_i^2 - \bar{D}^2}{n-1} = \frac{22.79 - 10 \times 0.98^2}{9} = 1.47.$$

  Thus $S_D = 1.21$. The value for t is then

$$t = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} = \frac{0.98 - 0}{1.21/\sqrt{10}} = 2.56.$$

- This is outside the range of the critical values so we reject the Null Hypothesis; there is a change.

3.3.4 The Direction of Change

When carrying out a two-tailed test, the order of subtraction does not matter as it simply changes the sign of the difference variable D; the Z and T variables are symmetric.

However, when dealing with one-tailed tests, we must be careful that the order of subtraction and the choice of the sign of the critical value are consistent with the Null and Alternative Hypotheses. This is illustrated in the next example.

3.3.5 Example: Smoking is Bad For You

A test is conducted to see if smoking a cigarette has an immediate detrimental effect on the aerobic capacity of a subject. To answer this question, the following data was collected from a group of 10 people:

- For each person, their peak flow lung function 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:	619.92	633.24	674.64	622.44	677.52
After:	672.84	633.96	615.6	668.16	660.24
Before:	616.2	632.47	673.43	621.28	677.51
After:	668.14	630.22	611.72	663.22	658.00

Let variable D represent the differences between corresponding values. Let P_i be peak flow before the cigarette and let Q_i be the peak flow afterwards: then

$$D_i = P_i - Q_i.$$

Carrying out the subtraction in this order means that if there is an overall drop in peak flow values, then this will show up as a positive value of $\bar{D}$.

We now carry out the test. Let μ_D be the true mean of the variable D.

- Frame the hypotheses; the Null Hypothesis is: $\mu_D = 0$, the Alternative hypothesis: $\mu_D > 0$. This is a one-tailed test.
- The level of significance is 0.005. From the tables, the critical value for 9 degrees of freedom is 3.250.
- The sample mean of the differences is 2.64 and $s_D = 1.772$. The value for t is then found to be 4.77.
- This is above the critical value so we reject the Null Hypothesis; there is a positive change, meaning 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. This is coming from the symmetry property of the T distribution:

$$P[T > 3.25] = 0.005 \Rightarrow P[T < -3.25] = 0.005.$$

4 Confidence Intervals

In our discussion of the tests on a mean, we have been using a test statistic to carry out a hypothesis test on a proposed value of the true mean μ of a population. This approach is based on having a value of the true mean, in advance, which we wish to test. The test statistic is either the Z or T variables, depending on the sample size n, and they are functions of the sample parameters coming from a data set. We will see in due course that the same approach is taken with other parameters such as the variance and the correlation coefficient. This approach can be restrictive, so we now look at an alternative.

4.1 Definition and Calculation of a CI

The alternative approach to the Null-Hypothesis based test is to calculate the sample mean from the data available and then use our knowledge of the behaviour of the sample mean as a variable to comment on how likely it is to be incorrect by a certain amount. This is equivalent to deciding on a range in which we expect to find the true mean, to a certain probability. This is phrased in the following definition.

4.1.1 Definition: Confidence Interval

A confidence interval at a given probability p for the mean of a population is a pair of numbers a and b such that

- The probability the interval (a, b) holds the true mean is p and
- The true mean is equally likely to be above the upper limit as below the lower limit.

4.1.2 Calculation of a Confidence Interval

To see how a confidence interval can be calculated, look again at what the Central Limit Theorem tells us and at the properties of the standard normal distribution. Let α be a level of significance, in other words a low probability. Let the number z_α be the one-tailed critical value for the probability α, so if Z is the standard normal variable then

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

So for a confidence interval for a probability p, we are looking for two numbers a and b such that

$$P[a < \mu < b] = p.$$

This range must also be symmetric, so in fact we are dealing with the idea of a middle range. From this we see that

$$P[\mu < a] = (1-p)/2 \text{ and } P[\mu > b] = (1-p)/2.$$

Let us set $\alpha = (1-p)/2$ and look at the second equation:

$$P[\mu > b] = \alpha.$$

Now take advantage of what we know about the behaviour of the variable

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

and try and bring it in to the equation for b. It will be true that

$$P[\mu > b] = \alpha \Rightarrow P[-\mu < -b] = \alpha \Rightarrow P\left[\frac{\bar{X}-\mu}{s/\sqrt{n}} < \frac{\bar{X}-b}{s/\sqrt{n}}\right] = \alpha.$$

The random variable on the left of the inequality sign in the event follows the standard normal distribution, so we can say that

$$P\left[Z < \frac{\bar{X}-b}{s/\sqrt{n}}\right] = \alpha.$$

Recalling that we set z_α as the one-tailed critical value for the probability α:

$$P[Z > z_\alpha] = \alpha \Rightarrow P[Z < -z_\alpha] = \alpha.$$

This must mean that

$$\frac{\bar{X}-b}{s/\sqrt{n}} = -z_\alpha.$$

Solving this equation shows that

$$b = \bar{X} + z_\alpha \frac{s}{\sqrt{n}}.$$

The same logic also shows that

$$a = \bar{X} - z_\alpha \frac{s}{\sqrt{n}}.$$

It will be the case that if the sample size n is not high, then we will use the corresponding critical values for the T distribution instead.

We will now illustrate this idea using data from previous examples of tests on a mean.

4.1.3 Example: A confidence Interval

Recall the example of rainfall data for the East Coast of Ireland. We will calculate a 99% Confidence interval for the rainfall from this data.

The sample mean and the sample standard deviation were calculated from the data;

$$\bar{X} = 708.8, \;\; s = 22.858.$$

For a 99% CI, the level of significance is

$$\alpha = \frac{1 - 0.99}{2} = 0.005.$$

For a sample size of $n = 40$ we could use the critical value from the T tables for 39 degrees of freedom:

$$P[T_{39} > 2.708] = 0.005,$$

or we could use the Z table critical value:

$$P[Z > 2.576] = 0.005.$$

Working with the T value, the confidence interval is:

$$\bar{X} \pm 2.708\frac{s}{\sqrt{n}} = 708.8 \pm 2.708\frac{22.858}{\sqrt{40}}.$$

The two numbers are: 699.013 and 718.587. Therefore the probability that the true mean is in this range is 0.99.

Recall that when we first looked at this data, we carried out a two-tailed test on the Null Hypothesis that $\mu = 700$. We did not reject the Null Hypothesis in that case. This was a very close result. This is consistent with the result here that the value of 700ml is just inside this confidence interval. For a 95% CI, we would use the critical value for 0.025 at 39 degrees of freedom.

$$708.8 \pm 2.023\frac{22.858}{\sqrt{40}} \equiv (702.475, 715.125).$$

Notably the value 700ml is not in this range, reflecting the fact that we would reject the Null Hypothesis in a two-tailed test with a level of significance of 0.05.

5 Tests on a Variance

The first few tests we looked at were for testing a Hypothesis on the mean of a list of data; it may also be required to test the standard deviation of a sample, to see if it differs from a suspected value. We will show here a test on the variance, which is the square of the standard deviation. In order to be able to do this, we introduce a new distribution.

5.1 The Chi Square Distribution

The χ^2 (pronounced 'ki square') statistic follows a distribution that is quite distinct from the standard normal and t-distribution that we have considered so far. Like the t-distribution, it depends on an integer parameter which, in the context of a test, will come up as a *degrees of freedom* value. The distribution is given in the third table available to you.

The graph in figure 2 shows the probability density function for the chi square distribution, for a degrees of freedom number of 6. Take note of how the distribution is not symmetric - the χ^2 variable only takes on positive values. The areas under the curve to the right of a given value a then gives us the probability $P[\chi^2 > a]$, as shown with the grey area.

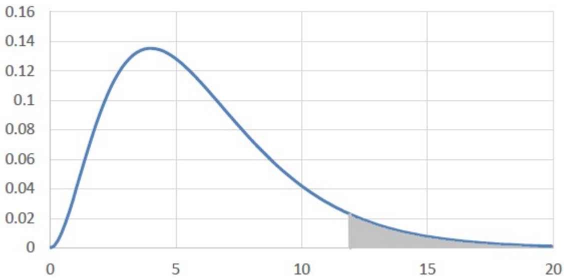

Figure 2: The PDF of the Chi square Distribution

The tables show the critical values for the distribution. Thus for a given probability α, a figure a may be found across from the degrees of freedom value, so that

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

The figures below the probabilities 0.95, 0.975 etc. will be important when a two-tailed or left-tailed test is to be done.

5.2 The Test on a Variance

We will now state the details of the test on the variance.

Consider again the case of a sample of n values of a random variable X. The values are denoted as:

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

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

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

It can be shown that the quantity Q given by

$$Q = \frac{\sum_i (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

D.o.f	Probabilities							
	0.05	0.025	0.01	0.005	0.995	0.99	0.975	0.95
1	3.841	5.024	6.635	7.879	0	0	0.001	0.004
2	5.991	7.378	9.21	10.597	0.01	0.020	0.051	0.103
3	7.815	9.348	11.345	12.838	0.072	0.115	0.216	0.352
4	9.488	11.143	13.277	14.860	0.207	0.297	0.484	0.711
5	11.07	12.833	15.086	16.75	0.412	0.554	0.831	1.145
6	12.592	14.449	16.812	18.548	0.676	0.872	1.237	1.635
7	14.067	16.013	18.475	20.278	0.989	1.239	1.690	2.167
8	15.507	17.535	20.090	21.955	1.344	1.646	2.180	2.733
9	16.919	19.023	21.666	23.589	1.735	2.088	2.700	3.325
10	18.307	20.483	23.209	25.188	2.156	2.558	3.247	3.94
11	19.675	21.920	24.725	26.757	2.603	3.053	3.816	4.575
12	21.026	23.337	26.217	28.300	3.074	3.571	4.404	5.226
13	22.362	24.736	27.688	29.819	3.565	4.107	5.009	5.892
14	23.685	26.119	29.141	31.319	4.075	4.66	5.629	6.571
15	24.996	27.488	30.578	32.801	4.601	5.229	6.262	7.261
16	26.296	28.845	32.000	34.267	5.142	5.812	6.908	7.962
17	27.587	30.191	33.409	35.718	5.697	6.408	7.564	8.672
18	28.869	31.526	34.805	37.156	6.265	7.015	8.231	9.39
19	30.144	32.852	36.191	38.582	6.844	7.633	8.907	10.117
20	31.410	34.170	37.566	39.997	7.434	8.260	9.591	10.851
21	32.671	35.479	38.932	41.401	8.034	8.897	10.283	11.591
22	33.924	36.781	40.289	42.796	8.643	9.542	10.982	12.338
23	35.172	38.076	41.638	44.181	9.260	10.196	11.689	13.091
24	36.415	39.364	42.98	45.559	9.886	10.856	12.401	13.848
25	37.652	40.646	44.314	46.928	10.520	11.524	13.120	14.611

follows the χ^2 distribution with $n-1$ degrees of freedom. A test can therefore be done with this statistic. The tests go as follows.

- A value σ^2 for the true underlying variance of the variable X is put forward.
- The Null Hypothesis H_0 is that this value σ^2 is correct. The alternate Hypothesis H_A is that it is not.
- Depending on the alternate Hypothesis, decide if the test is one or two tailed. If the level of significance is α and a left-hand, lower, critical value is required, it is calculated by using the complementary event:

  $$P[\chi^2 > a] = \alpha \Rightarrow P[\chi^2 < a] = 1 - \alpha.$$

  We need this as the distribution is not symmetric.
- Critical values are chosen from the tables and the value of test statistic Q are calculated. The Null hypothesis is rejected or not as usual.

5.2.1 Example: A Test on a Mean and Variance

A component manufacturer has taken a random sample of 13 components and tested them to destruction. Their lifetimes were recorded in units of 10^6 seconds and the following information produced:

$$\sum_i L_i = 1,410.8, \ \sum_i L_i^2 = 153,123.25.$$

Carry out an appropriate statistical test on each of the two statements given here, using the data given below.

1. The mean lifetime of the components is 108.
2. The variance of the lifetimes is 5.2.

The first test will be a two tailed t-test on the mean and the second will be a two-tailed test on the variance. We will take 0.01 as the level of significance throughout.

For the test on the mean, the Null Hypothesis is $H_0 : \mu = 108$. The alternate Hypothesis is $H_A : \mu \neq 108$.

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; divide this by 2 for a two-tailed test. At 12 degrees of freedom, this means the critical values are ± 3.055, from the tables. The sample mean is

$$\frac{1410.8}{13} = 108.523.$$

The sample standard deviation is:

$$s^2 = \frac{153,123.25 - 13 \times 108.523^2}{12} = \frac{18.893}{12} = 1.574 \Rightarrow s = 1.255.$$

From this,

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{108.523 - 108}{1.255/\sqrt{13}} = 1.503.$$

This is well inside the range of the critical values so the null hypothesis is not rejected; the value of 108×10^5 seconds is a valid figure for the mean lifetime according to this data.

For the test on the variance, the Null Hypothesis is $H_0 : \sigma^2 = 5.2$. The alternate Hypothesis is $H_A : \sigma^2 \neq 5.2$.

Given the alternative hypothesis, the test will be a two-tailed test on the variance, 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; 0.005. The number of degrees of freedom is $n - 1 = 12$. This means the upper critical value is 28.300, from the tables. Thus the upper tail is Thus we are watching out for the event

$$P[\chi^2 > 28.3] = 0.005.$$

We find the other critical value using the complementary event:

$$P[\chi^2 < a] = 1 - \alpha.$$

From the tables under 1 - 0.005 = 0.995,

$$P[\chi^2 > 3.074] = 0.995.$$

Thus 3.074 is the lower critical value.

The test statistic is

$$Q = \frac{\sum_i (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}.$$

The top line of the variable expression has been calculated already, so

$$Q = \frac{18.893}{5.2} = 3.779.$$

This is within the range of our critical values, therefore the Null Hypothesis is not rejected. The claim of $\sigma^2 = 5.2$ is not invalidated by the data we collected.

It is worth noting that if we had taken a level of significance of 0.05 the lower bound would be 4.404 so the Null Hypothesis wold be rejected.

5.2.2 Example: Data Snooping

A company is machining engine parts with a particular component diameter D, 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 $s^2 = 0.03$mm^2. Carry out a one-tailed test on this claim.

The test must be one-tailed because the information available for this example is the sample variance itself and the sample value is in fact higher than 0.02. The only question left to answer is if it is high enough to reject the claim of 0.02mm^2.

Here is the test:

- The Null Hypothesis is $H_0 : \sigma^2 = 0.02$. The alternate Hypothesis is $H_A : \sigma^2 > 0.02$.
- Take 0.05 as the level of significance and the number of degrees of freedom is $10 - 1 = 9$. The critical value is 16.92.
- The sample value of the test statistic is

$$Q = \frac{9 \times 0.03}{0.02} = 13.5.$$

- The null hypothesis is therefore not rejected.

5.2.3 And Again...

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

- The Null Hypothesis is $H_0 : \sigma^2 = 0.02$. The alternate Hypothesis is $H_A : \sigma^2 < 0.02$. This is still a one-tailed test, but now the direction has changed.
- Take 0.05 as the level of significance. Sine we need the left-had, lower critical value we look at the values under 0.95. The number of degrees of freedom is $14 - 1 = 13$. From the tables, at probability 0.95 at 13 degrees of freedom, the critical value is 5.89.

$$P[\chi^2 < 5.89] = 0.95.$$

- The sample value of the test statistic is

$$Q = \frac{13 \times 0.015}{0.02} = 9.75.$$

- The Null Hypothesis is therefore not rejected.

6 Correlation and Regression

The subject of correlation and regression are extremely important to statistics and the sciences in general. As the word 'correlation' itself suggests, the core of this subject is the idea of investigating how two quantities are linked.

6.1 Recap: The idea of Correlation

The starting point is that data for two variables x and y is available of the form:

$$X_1 \dots X_n \text{ and } Y_1 \dots Y_n.$$

Thus for every value X_i there is a corresponding value Y_i. Our data may therefore be regarded as a list of points:

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n).$$

6.2 Correlation as a statistic

We saw correlation treated as a simple geometric idea with the calculation of a parameter to show how close to a line the set of points listed were. The parameter was the correlation coefficient r given by

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}.$$

Since we are now treating lists of data, such as the points invovled here, as random samples from variables, the correlation number r which is a function of this data is itself a random variable. We will therefore treat this parameter as a statistic.

The value of r is a function of the data points (X_i, Y_i) and so is in fact a value of a statistic. The number r should be therefore viewed as the sample value for the quantity ρ, the true underlying correlation between the two variables. We will set up the structure of the test here and then look at the interpretation of what results for r mean.

6.2.1 The Structure of a Test on a Correlation

The first and most important, question is to decide on a Null Hypothesis. Look at this in terms of the investigation we are carrying out: is there a correlation? The answer *yes* can come from a non-zero value of r, whereas the concept of *no correlation* can be identified as the statement $\rho = 0$. This will therefore serve as our Null Hypothesis.

If $\rho = 0$, it can be proved that the variable given here

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}, \tag{1}$$

where r is the correlation coefficient and n is the number of data points, follows the T Distribution with $n-2$ degrees of freedom.

Therefore for a test for a correlation between two variables, with the data as noted above in equation 1, the Null Hypothesis is $\rho = 0$ and the test statistic is t as given above.

6.2.2 An Alternative Approach

An alternative approach is to generate a table of critical values for the statistic r itself by solving equation 1 for r in terms of t. Starting with the equation:

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \Rightarrow r = \frac{t}{\sqrt{t^2+n-2}}.$$

Therefore if we have a level of significance and degrees of freedom and therefore a critical value from the t distribution, this version of the equation gives us a critical value for r. This means one could draw up a table of critical values for the statistic r, derived from the t-table. We will not take this approach.

6.2.3 Example: A test on a Correlation

The following are a list of corresponding values of two variables x and y:

$$x : 1.3, 1.9, 2.4, 3.1, 3.8, 4.5, 5.1$$

$$y : 1.9, 7.1, 7.3, 7.2, 11.9, 15.1, 12.8.$$

We will carry out a test to see if there is a positive correlation between the variables x and y.

Let ρ be the true underlying correlation.

- The null hypothesis: there is no correlation: $H_0 : \rho = 0$.

  The alternative: there is a positive correlation: $H_A : \rho > 0$.

- We will take a level of significance of 0.01.

- The test is one-tailed, so 0.01 remains as our probability. This is because we are only investigating a positive correlation.

- There are $7 - 2 = 5$ degrees of freedom.

- From the t tables, the critical value for a probability of 0.01 at 5 degrees of freedom is 3.365.

Recall that r can be found with the following equation:

$$r = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_i X_i^2 - n\bar{X}^2}\sqrt{\sum_i Y_i^2 - n\bar{Y}^2}}.$$

Some important calculations are shown here:

$$\sum_i X_i^2 = 81.37, \sum_i Y_i^2 = 692.61 \text{ and } \sum_i X_i Y_i = 234.25.$$

The calculation for r is then:

$$r = \frac{25.234 - 7 \times 3.157 \times 9.043}{\sqrt{81.37 - 7 \times 3.157^2}\sqrt{692.61 - 7 \times 9.043^2}} =$$

$$= \frac{34.403}{\sqrt{11.597}\sqrt{120.197}} = 0.92.$$

Now calculate the value of t:

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} = \sqrt{5}\frac{0.92}{\sqrt{1-0.92^2}} = 5.248.$$

The value found is well above the critical value; therefore the null hypothesis is (emphatically!) rejected. There is a positive correlation.

In fact, seeing such a high value of t even for a low number of points tells us we are likely to reject the Null Hypothesis.

6.3 The Regression Equations

On studying correlation, we also looked at the the linear regression coefficients a and b which were values of linear parameters which attempted to fit a line

$$Y = \alpha + \beta X$$

to our data, which minimised the squared error

$$\sum_i (a + bX_i - Y_i)^2.$$

The equations were

$$b = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{\sum_i X_i^2 - n\bar{X}^2}, \quad a = \bar{Y} - b\bar{X}.$$

These two parameters are now treated as sample values of the two coefficients for the regression line, so depend on quantities calculated from the original data points, that is, the values of the two variables x and y. This means that these coefficients are themselves statistics, in the same way that a sample mean, a sample standard deviation, or quantities like t and r are statistics. Distributions can be found for them, leading to very useful and powerful tests for the idea of dependence of one variable on another. The details are beyond the scope of this study.

7 Categorical Tests

The final set of tests we will look at concern tests that are not on a particular parameter, but rather the link between categories. These types of tests invariably rely on the χ^2 distribution, with a test variable of the form

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

where the symbols O_i are observed values, usually of counts of subjects in a particular class, and the E_i are the expected values, usually counts in those classes under the Null Hypothesis.

When we study categorical tests, they will be first introduced in the context of tests for the independence or otherwise of two categorical variables. The second major and related test is to see if a frequency distribution is following a particular distribution for a random variable, for example the normal distribution.

We will start with the test for independence.

7.1 Test for Independence

Consider the following motivating example.

7.1.1 Example: Pass Rates

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 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 and deciding whether these results are different from what would be expected if there were no connection, so that the passes, merits and fails were distributed evenly, that is, in proportion.

From looking at the table of results, there does seem to be a higher proportion of mature students doing well, for example, students in category C gaining merits. To see what the results would be if they were coming up in proportion to the numbers in each category, we will firstly have to calculate the totals in each category. The subtotals are recorded in the table as shown.

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

To see what would be expected if no group is doing better than the other, divide up the results in proportion. For example, it was observed that 11 students from group A failed. To see how many we would expect if the number was in proportion, note that 12 students failed overall, and 38 are in group A. Both of these numbers are out of a total of 59 students. Then the number of students *expected* to fail from group A is found by multiplying 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

$$\frac{12 \times 38}{59}.$$

This number is the expected value for the cell for fails from group A. A table of such expected values can be calculated for this situation, 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

The next step is to provide some measure of how different these figures are from the observed values.

7.1.2 The Test for Independence

The value of the test variable is calculated from the data in the following way. Call each of the observed values O_i and call the expected values E_i. The index i covers all the cells in the table. The sample value for this statistic chi square is is given by equation

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

We must have a degrees of freedom number; if r is the number of rows and c is the number of columns, this is given by

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

In this case, the degrees of freedom is $(3 - 1)(3 - 1) = 4$. We will take a level of significance of 0.05. The critical value is found in the tables, 9.48, so that at 4 degrees of freedom it is the case that

$$P[\chi^2 > 9.48] = 0.05.$$

We now have everything in place to carry out the test. We calculate the value of $(O_i - E_i)^2/E_i$ for each cell and sum the results:

$$\begin{array}{cccc} & \text{Merit} & \text{Pass} & \text{Fail} \\ A & \dfrac{(11-14.81)^2}{14.81} & \dfrac{(16-15.46)^2}{15.46} & \dfrac{(11-7.73)^2}{7.73} \\ B & \dfrac{(5-4.68)^2}{4.68} & \dfrac{(6-4.88)^2}{4.88} & \dfrac{(1-2.44)^2}{2.44} \\ C & \dfrac{(7-3.51)^2}{3.51} & \dfrac{(2-3.66)^2}{3.66} & \dfrac{(0-1.83)^2}{1.83} \end{array}$$

The sum of these values is 9.56.

This result is above the value of 9.48, and the probability of such an event happening was 0.05, our level of significance. Therefore we call into question the idea that there is no link between the groups of students and how well they do; in other words, we reject the null hypothesis and we can conclude that the mature students are indeed doing better.

Having finished this test, we will now formally set out the framework of the test for independence.

7.1.3 The Framework of a Chi-square Test for Independence

In this type of hypothesis test, a possible link or connection is being investigated between two quantities or variables. Rather than a contiguous range of values, the variables are divided into categories. Appropriate data will have been collected, consisting of counts in each combination of the categories of the two variables. This table of counts is known as a contingency table. In the students example, the two categorical variables are Type of student and Type of result.

- Framing the Hypothesis: 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 are independent, the alternate states that there is a link.

- Decide on a level of significance; as before, this means choosing a number between 0 and 1 which is regarded as a low probability. Usually this is 0.05 or 0.01.

- Identify the degrees of freedom; for a contingency table with r rows and c columns, it is $(r-1)(c-1)$.

- Find a critical value of the distribution, that is, a value associated with the level of significance. The test statistic we will use will follow the chi-square distribution with the $(r-1)(c-1)$ degrees of freedom. Thus for the 4 degrees of freedom of this case,

$$P[\chi^2 > 9.48] = 0.05.$$

- The next step is to calculate the expected values from the data for the case where the null hypothesis is true. This effectively means drawing up a second table of values where each value is the product of the two corresponding sub-totals, divided by the overall total. In the mature students example, this is equivalent to calculating the expected values as if all the counts of passes and fails of students in each category were in proportion.

- The value of Q, given by

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

  is calculated from these observed and expected values.

- The test value of Q is compared to the critical value; if the value found for the statistic is found to be an unlikely value, in other words if it is higher than the critical value, then the Null hypothesis is rejected and the alternative is accepted.

7.1.4 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-square test, whether or not the three processes are being used differently in the four factories. The contingency table for this question is shown here.

	Factory 1	Factory 2	Factory 3	Factory 4
Process A	12	23	32	41
Process B	21	34	54	45
Process C	24	31	17	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 alternate hypothesis is that there is a link.

- We will take the level of significance as 0.01.

- The number of degrees of freedom is $(4-1)(3-1) = 6$.

- From the tables, the critical value for these parameters is 16.81; in other words,

$$P[\chi^2 > 16.812] = 0.01.$$

- The expected values under the Null Hypothesis are calculated:

	Factory 1	Factory 2	Factory 3	Factory 4
Process A	17.741	27.389	32.058	30.813
Process B	25.297	39.055	45.712	43.937
Process C	13.963	21.556	25.231	24.251

For example, the expected value for Factory 3 with Process B is

$$\frac{103 \times 154}{347} = 45.712.$$

- The test statistic is calculated:

$$Q = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(12 - 17.741)^2}{17.741} + \ldots \frac{(13 - 24.251)^2}{24.251} = 28.099.$$

  The value found is 28.099, well in excess of 16.812.

- The null hypothesis is therefore rejected, so we conclude there is a connection between the manufacturing processes and how they are used in each factory. In fact, the Null Hypothesis would be rejected at a level of significance of 0.0001 based on this data, the critical value at this level would be 27.856.

7.2 Goodness-of-Fit of a Distribution

A question that frequently comes up in statistical analysis is the distribution underlying a given data set. If we are told that a random variable, for example, heights, is normally distributed, we expect that most of the numbers will be quite close to the mean value and very few of them will be far away from the mean, with 'far away' being defined by the standard deviation. We will now look at the reverse question - given a set of data, regarded as a sample from a random variable, carry out a test on whether the data looks like it 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. In fact, what we are looking for is a more general version of the test we have already studied for use on categories. The categorical test we used concerned a set of observed values O_i, which were compared to corresponding expected values Ei, which were calculated under a Null Hypothesis of 'occurring in proportion'. The difference between the two sets of values was measured and tested by the variable given by the equation

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

where the summation is done over every pair of corresponding values. The statement 'occurring in proportion' will now be broadened for any frequency distribution, leading to the following Hypothesis Test.

7.2.1 The Structure of a 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 that the number n is reasonably large. The Hypothesis test that the random variable X follows a given distribution goes as follows:

- Framing the Hypothesis: The Null hypothesis states that the variable does follow the proposed distribution. The alternate states that it does not.

- Decide on a level of significance; as before, this means choosing a number between 0 and 1 which is regarded as a low probability. Usually this is 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 distribution proposed in the Null Hypothesis calculated from the data and k is the number of groups into which the data will be divided.

- Find a critical value of the distribution; the test statistic we will use will follow the chi-square distribution.

- The next step is to calculate the expected values E_i from the data for the case where the null hypothesis is true. This effectively means drawing up a frequency distribution of expected counts of the number of values of X in each group.

- The observed values of the actual distribution O_i are taken from the data.

- The value of Q, given by

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

  is calculated from these observed and expected values.

- The test value of Q is compared to the critical value; if the value found for the statistic is found to be an unlikely value, in other words if it is higher than the critical value, then the Null hypothesis is rejected and the alternative is accepted.

7.2.2 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.

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 H 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, the mean and the standard deviation, which will be calculated from the data. The total number of observed or expected values we have is $k = 7$, so the degrees of freedom is $7 - 2 - 1 = 4$.
- From the tables, the critical value is therefore 9.488.
- The required frequency distribution is already available, in other words, the data $H_1, H_2, \ldots H_n$ 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. This is done on the basis of the Null Hypothesis, in other words, the assumption that the grouped data follows the normal distribution. Firstly, work out the two parameters of the distribution, the sample frequency mean and the sample frequency distribution:

  $$n = \sum_i f_i = 138.$$

  The mean is:

  $$\bar{h} = \frac{\sum_i m_i f_i}{n} = \frac{236.65}{138} = 1.715.$$

  $$s^2 = \frac{\sum_i m_i^2 f_i - n \times \bar{h}^2}{n-1} = \frac{406.2813 - 138 \times 1.715^2}{137} = \frac{0.4601}{137}.$$

  So $s = 0.058$.

- With the parameters for the normal distribution, it will now be possible to calculate the expected values, in other words, how many of the data values we expect in each group. We will show the calculations for the first group, 1.55m to 1.60m. The probability that a height will fall in this range is

$$P[1.55\text{m} < H < 1.60\text{m}] = P[H < 1.60\text{m}] - P[H < 1.55\text{m}].$$

Standardise and use the tables:

$$P[H < 1.60\text{m}] = P\left[Z > \frac{1.6\text{m} - 1.715\text{m}}{0.058\text{m}}\right] = P\left[Z < -2.84\right].$$

Using the symmetry of the standard normal distribution, this is

$$P\left[Z < -2.84\right] = P\left[Z > 2.84\right] = 0.0022.$$

For the lower limit:

$$P[H < 1.55\text{m}] = P\left[Z > \frac{1.55\text{m} - 1.715\text{m}}{0.058\text{m}}\right] = P\left[Z < -1.98\right].$$

Using the symmetry of the standard normal distribution, this is

$$P\left[Z < -1.98\right] = P\left[Z > 1.98\right] = 0.0238.$$

With these results,

$$P[1.55\text{m} < H < 1.60\text{m}] = P[H < 1.60\text{m}] - P[H < 1.55\text{m}] = 0.0216.$$

- With the probability calculated, the expected value for that group is now

$$E_1 = 0.0216 \times 138 = 2.981.$$

The term for the test statistic is then

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(4 - 2.981)^2}{2.981} = 0.3485.$$

Continuing in this way for the other groups, the test statistic is 2.213.

- This is well below the critical value, so the Null Hypothesis is not rejected. Based on this data, the random variable H is indeed normally distributed. Alternatively, there is no evidence to suggest, from this data set, that the height of men is not Normally distributed.

During the calculations for each group, we can take advantage of the fact that the upper bound of each group is the lower bound of the next group. We will need to standardise the upper and lower bounds before finding the difference in probabilities, unless we are using a spread-sheet, where the normal distribution probabilities can be calculated directly.

7.2.3 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	*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

We set out the steps in the test here. Let us use D to refer to the random variable of the diameters of these machined discs.

- The Null hypothesis states that the variable D 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 $k = 10$, so the degrees of freedom is $10 - 2 - 1 = 7$.
- From the tables, the critical value is therefore 14.067.
- To find the expected values, start by finding the frequency sample mean and frequency standard deviation:

$$n = \sum_i f_i = 150.$$

The mean is:

$$\bar{h} = \frac{\sum_i m_i f_i}{n} = \frac{33,130}{150} = 220.867.$$

$$s^2 = \frac{\sum_i m_i^2 f_i - n \times \bar{h}^2}{n-1} = \frac{7,330,137.5 - 150 \times 220.867^2}{149} = \frac{12824.833}{149}.$$

So $s = 9.278$. The following table set out the results of our calculations.

Group	Frequency	Stand. L	$P[D < L]$	E_i	$\frac{(O_i - E_i)^2}{E_i}$
200 to 205	12	-2.249	0.0123	4.704	11.315
205 to 210	7	-1.710	0.0436	11.569	1.805
210 to 215	16	-1.1713	0.121	21.426	1.374
215 to 220	30	-0.633	0.264	29.881	0.0005
220 to 225	35	-0.093	0.463	31.386	0.416
225 to 230	30	0.446	0.672	24.829	1.077
230 to 235	12	0.985	0.838	14.792	0.527
235 to 240	5	1.523	0.936	6.636	0.403
240 to 245	2	2.062	0.980	2.242	0.026
245 to 250	1	2.601	0.995	0.560	0.324

When the figures in the last column are totalled up, the result is 17.2685, which exceeds the critical value. The Null hypothesis is therefore rejected.

To see an example why this might be the case, observe the first value; it is a high value compared to the next value, instead of tailing off as would be expected in a Normal distribution. It is much higher than the expected value. This gives a very large value of the term $\frac{(O_1 - E_1)^2}{E_1}$.

7.2.4 Example: The Uniform Distribution

A nursery is growing garden plants that come in 15 different varieties, labelled by a number 1 to 15. The selection of plant varieties for trays for selling is random. For a given months production of trays, the number of times each plant variety was selected is given in the list shown below.

A Null Hypothesis-based statistical test is to be done to determine if the frequency distribution of counts follows the uniform distribution, at a level of significance of 0.05 and then at 0.01. The test is to be two-tailed.

Plant Type	Count	Plant Type	Count
1	15	9	16
2	14	10	13
3	13	11	11
4	9	12	15
5	17	13	18
6	18	14	14
7	16	15	21
8	12		

We set up the framework of the test first.

- The Null Hypothesis is that the counts follow the uniform distribution, that is, they are all equal. The alternative is that they do not.
- There are no parameters calculated from the data, therefore the number of degrees of freedom is $15 - 0 - 1 = 14$.

- The test should be two-tailed, as observed values too close to the expected value would be very unusual. It would suggest the variable was not truly random.

- The critical values for a two-tailed test are: For 0.05, they are 26.119 and 5.629. For 0.01, they are 31.319 and 4.075.

The next step is to calculate the sample value of the test statistic for a goodness-of-fit test for a distribution:

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

This can be simplified for the uniform distribution. Since the expected values are all the same, E, then

$$Q = \sum_i \frac{(O_i - E)^2}{E} = \frac{1}{E}\sum_i (O_i - E)^2.$$

Look at the summation:

$$(O_i - E)^2 = O_i^2 - 2EO_i + E^2.$$

Carry out the summation:

$$\sum_i (O_i - E)^2 = \sum_i O_i^2 - 2E\sum_i O_i + kE^2.$$

Use $\sum_i O_i = kE$ so that

$$\frac{1}{E}\sum_i (O_i - E)^2 = \frac{1}{E}\left[\sum_i O_i^2 - 2EkE + kE^2\right] =$$

$$= \frac{1}{E}\left[\sum_i O_i^2 - kE^2\right] = \frac{1}{E}\sum_i O_i^2 - kE.$$

Some calculation shows that

$$\sum_i O_i = 222.$$

$$\sum_i O_i^2 = 15^2 + 14^2 + 13^2 + \ldots = 3416.$$

If the distribution is uniform, then the expected value of the counts is

$$\frac{1}{15} \sum_i O_i = \frac{222}{15} = 14.8.$$

Now the sample value Q can be calculated: $E = 14.8$, so the sample value is

$$Q = \frac{3416}{14.8} - 222 = 8.811.$$

So in both cases the Null Hypothesis is not rejected. Thus this data suggests a uniform distribution for the variable is consistent with the data.