Introduction to inferential statistics using Microsoft Excel

Contents

1 Introduction ........................................................................................................................................... 1
2 Calculating confidence intervals in Excel ......................................................................................... 2
3 Carrying out tests of difference ........................................................................................................... 7
4 Carrying out tests of association ......................................................................................................... 15
5 Linear regression ................................................................................................................................... 22

1 Introduction

This guide introduces the use of Microsoft Excel (hereafter: Excel) for calculating inferential statistics and carrying out selected hypothesis tests. It shows how some of the techniques discussed in Chapter 13 can be applied in Excel. Please refer to Chapter 13 for more details on the specific techniques and their interpretation; the focus here is on how to carry them out in Excel. It covers four main topics:

1. Calculating confidence intervals in Excel
2. Carrying out tests of difference in Excel
3. Carrying out tests of association in Excel

The guide is not written for a specific version of Excel although it includes screenshots for Excel 2010. Most of the functionality referred to in the guide is also available in earlier and later versions, although the user interface has changed somewhat.

The guide assumes that you have entered your data and prepared it for analysis as described in the guide Introduction to using Microsoft Excel for quantitative data analysis. It also assumes that you are familiar with basic Excel functionality, including creating and editing charts (for more information on how to use functions and the Data Analysis ToolPak see the other guides to use Excel on the companion website).
The techniques presented in this guide are mainly those for which Excel provides built-in functions. For more information on carrying out other statistical tests in Excel, see:


There are also many websites which discuss the use of Excel for statistical analysis. Some of these are very helpful on more advanced techniques for those with a working knowledge of both Excel and the statistics being applied.

## 2 Calculating confidence intervals in Excel

Recall from Chapter 13 that the general form for a confidence interval is:

\[
\text{Point estimate} \pm \text{margin of error}
\]

Excel can be used to generate confidence intervals for a population mean and population proportion based on sample data by calculating point estimates and margins of error.

### 2.1 Confidence interval for the mean of a metric variable

To calculate the confidence interval for the mean of a metric variable you need to know:

- Sample mean. This can be calculated using the AVERAGE function or via the Descriptive Statistics routine in the Data Analysis ToolPak.
- The margin of error. There are two ways this can be calculated in Excel, either using the CONFIDENCE.T function or the Descriptive Statistics routine in the Data Analysis ToolPak.

We illustrate this by calculating the confidence interval for the mean of the variable Satisfaction in the customer satisfaction dataset.

### 2.2 Using the Descriptive Statistics routine

To use the Descriptive Statistics routine to do so, carry out the following steps:

1. Select Data > Data Analysis > Descriptive Statistics > OK to open the Descriptive Statistics dialogue box.
2. In the dialogue box, enter the range for the Satisfaction variable in the Input Range box. If you have included the column header, tick the Labels in the first row box. Choose the destination for the output (New Worksheet Ply is the default and recommended).

3. Now select Summary statistics and Confidence Level for Mean (Figure 1). The default for the latter is 95%; as explained in Chapter 13 this is common in management research but you can change it if desired.

Figure 1 – Descriptive statistics dialogue box

4. Click OK to generate the output (Figure 2). Excel’s Descriptive Statistics output includes the mean (4.25) and what it rather confusingly labels as the Confidence Level (95%). This is actually the margin of error for the confidence interval for the mean (0.565606).

Figure 2 – Descriptive statistics output
5. This information can now be used to calculate the confidence interval for the mean (rounded to 2 decimal places) as follows:
   
   a. Lower Confidence Level = mean – margin of error = 4.25 – 0.565606 = 3.68.
   
   b. Upper Confidence Level = mean + margin of error = 4.25 + 0.565606 = 4.82.

6. The results can be incorporated in a table along with other relevant statistics for the variable or they can be reported in text along with the mean, using the format\(^1\): 95% CI [Lower level, Upper level]. In our example this would be 95% CI [3.68, 4.82].

2.2.1 Using the CONFIDENCE.T function

As an alternative, you can use the CONFIDENCE.T statistical function to calculate the desired margin of error for a confidence interval for the mean. To use this method you will need to have already calculated the sample standard deviation and the sample size as well as the sample mean, for example by using the relevant statistical functions or the Descriptive Statistics routine.

To calculate the confidence interval using the CONFIDENCE.T function:

1. Select a destination cell in the worksheet for the output. Add a caption so that you can remember what the output means.

2. Select Formulas > More Functions > Statistical > CONFIDENCE.T > OK to open up the Function Arguments dialogue box.

---

\(^1\) This is the style used by the American Psychological Association and provides a useful format if you have not been advised to use an alternative style.
3. In the dialogue box, enter the desired Alpha (\(\alpha\)). Recall from Chapter 13 that this is the significance level and is commonly set at 5% (0.05) in business and management research. Setting it to 0.05 will give a margin of error for a 95% (i.e. \(1 - \alpha\)) confidence interval.

4. Enter the sample standard deviation and the sample size in their respective boxes or references to the cells in which the data are located (Figure 3).

Figure 3 – CONFIDENCE.T Function Arguments dialogue box

5. Click OK. The output is created in the destination cell (Figure 4). The answer is 0.565605706 which rounds to 0.565606 as in the previous example.

6. The confidence interval can now be calculated and reported as per steps 5 and 6 in the previous example.

Figure 4 – Output from the CONFIDENCE.T function (cell caption added)
Using the CONFIDENCE.T function may seem long-winded compared to using the Descriptive Statistics routine but if, for example, you have created a table of means, standard deviations and sample sizes for different groups it is an easy way of calculating the confidence intervals for the means of each group.

(Note: if you have large sample sizes (≥ 30), you can use the CONFIDENCE.NORM function instead of the CONFIDENCE.T. This will give a different answer to the one calculated by the Descriptive Statistics routine which uses the same method as the CONFIDENCE.T function so ensure you are consistent if calculating and reporting confidence intervals.)

2.2.2 Confidence interval for a proportion

You can also calculate confidence intervals for proportions. Suppose, for example, that out of a random sample of 120 respondents, 90 indicated that they intend to buy your new product. The proportion intending to buy is therefore is 0.75 (= 90/120) or 75%. This is a point estimate. To calculate an approximate 95% confidence interval for a proportion for sample size of ≥ 30, you can use the following procedure:

1. Calculate the population proportion of interest and the sample size (in this case 0.75 and 120 respectively).
2. Use the following formula to calculate the margin of error for an approximate 95% confidence interval:

$$\text{margin of error} \approx 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where \(\hat{p}\) = the sample proportion and \(n\) = the sample size, therefore:

$$\text{margin of error} \approx 1.96 \times \sqrt{\frac{0.75(1-0.75)}{120}} = 0.0775$$

3. We can now use the margin of error to calculate the confidence interval for the population proportion (to 2 decimal places) as follows:
   a. Lower confidence level = 0.75 – 0.0775 = 0.67
   b. Upper confidence level = 0.75 + 0.0775 = 0.83.
4. An approximate 95% confidence interval for the population proportion is therefore [0.67, 0.83]: between 67% and 83% of customers are likely to buy your new product.
Excel can, of course, be used to perform the above calculations. (Hint: use the SQRT mathematical function to help.)

3 Carrying out tests of difference

Excel’s Data Analysis ToolPak contains a number of statistical tests of difference. Selected ones are shown in Table 1:

Table 1 – Selected statistical tests of difference available in Excel

<table>
<thead>
<tr>
<th>Excel function name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anova: Single Factor</td>
<td>Performs one-way analysis of variance (ANOVA)</td>
</tr>
<tr>
<td>F-Test Two Sample for Variance</td>
<td>Performs a test of whether or not the variances of two samples are equal (note: the p-value returned by the test is one-tailed and should be doubled to calculate the two-tailed p-value)</td>
</tr>
<tr>
<td>t-Test: Paired Two-Sample for Means</td>
<td>Performs a t-test to compare the means of a paired sample</td>
</tr>
<tr>
<td>t-Test: Two-Sample Assuming Equal Variances</td>
<td>Performs a t-test to compare the means of two independent samples, assuming equal variances</td>
</tr>
<tr>
<td>t-Test: Two-Sample Assuming Unequal Variances</td>
<td>Performs a t-test to compare the means of two independent samples, assuming unequal variances</td>
</tr>
</tbody>
</table>

In this section we will illustrate their use with two examples: the t-test assuming equal variance and the single-factor ANOVA.

3.1 Two-sample t-test assuming equal variances

Excel provides two versions of the t-test for independent samples: with and without the assumption of unequal variances. We will demonstrate the test using the customer satisfaction dataset to test the following hypotheses:

H₁: the satisfaction levels of male and female customers are different.

H₀: there is no difference in the satisfaction level of male and female customers.
3.1.1 Are the variances equal?

Prior to running the t-test, you can test the assumption of equality of variance using the F-Test Two-Sample for Variance routine in the Data Analysis ToolPak. The test is run in a similar way to the t-test. Note that the test returns a one-tailed $p$-value; you should double to get the two-tailed $p$-value which is required. If the two-tailed $p$-value is less than 0.05 (or if you cannot test the assumption) you should assume the variances are unequal and use the unequal variances version of the test. In this case we will use the test assuming equal variances because running the F-Test in Excel returned a one-tailed $p$-value of 0.37. This gives a two-tailed $p$-value of 0.74. Since 0.74 > 0.05 the assumption of equal variances can be made.

3.1.2 Running the t-test

For the test to run you have to enter the range for the two variables (i.e. the responses for male and female customers) into the test dialogue box. Your dataset must be in a suitable format for you to be able to do this. If it is not, there are two options:

1. Option 1 is to sort the dataset so that the rows containing cases from each group (i.e. all males or all females) are adjacent. To do this:
   a. Click on a column header cell in the dataset.
   b. Select Home > Sort & Filter > Filter to apply column header filters to the dataset.
   c. Click on the Filter arrow in the relevant column header (in this case Gender).
   d. Click on Sort A to Z or Sort Z to A. We have chosen Sort Z to A to sort in the order Males – Females. This rearranges the dataset so that it is sorted by the categories in the variable Gender. All the Male and Female cases are now grouped together (Figure 4).

Figure 5 – Dataset sorted by categories in variable Gender
e. You can now enter the range for each group by entering the range of the relevant data in the Input Range dialogue box.

f. (Hint: to return the dataset to its original layout when you have carried out the test, you can sort using the ID variable, selecting Sort A to Z. This will recover the original order, based on the ID variable.)

2. Option 2 is to create a copy of the data in which the data for each group to be compared appears in separate columns. You can do this by sorting the data as in option 1 and then selecting the relevant data from your main dataset and pasting it into a new worksheet to create two new columns (one for males, one for females in this case). Give each a column header (Figure 6). This method can be useful if you have a very large dataset and want to keep data and test outputs together. (Hint: give the new worksheet a suitable name to make it easier to manage the workbook.)

Figure 6 – Worksheet showing satisfaction levels for male and female respondents

In our example we will be using option 2 for clarity of presentation.
To run the test, carry out the following steps:

1. Select Data > Data Analysis > t-Test: Two Sample Assuming Equal Variances > OK to open the test dialogue box.

2. In the dialogue box:
   a. Enter the ranges of the data for the two variables in Input. If both columns have column headers (as here) you can include them in the range if desired.
   b. Enter 0 in the Hypothesized Mean Difference box. This tells Excel that this is a test of the null hypothesis (H₀) that there is no difference between the group means.
   c. If you have included column headers, tick the Labels box.
   d. Set the desired Alpha (i.e. significance level) for the test. The default is 0.05.
   e. Choose the Output option. If you have followed option 1 to prepare your data, use the default of New Worksheet Ply to insert the output in a new worksheet. If you have chosen option 2 (as here), you can select a suitable Output Range in the active worksheet (Hint: make sure it does not include the dataset or the test output will overwrite it). The completed dialogue box is shown in Figure 7.
f. Click OK. The resulting output is shown in Figure 8 with the column spacing adjusted to make it easier to read.

3. Next step is to interpret the output. Pay particular attention to the following:
   a. Means of each group (this allows you to see the magnitude of the difference between the groups).
   b. The group sizes (=10 for each group); you can check this to make sure no cases have been accidentally omitted from the sample.
   c. The test statistic (t Stat = 2.240186).
   d. Degrees of freedom (df = 18).
   e. Two-tailed p-value (= 0.037935).

4. Compare the p-value to your desired significance level (Alpha), in this case 0.05. Since 0.037935 < 0.05, we would reject $H_0$, accept $H_1$ and conclude that there is a statistically significant difference between the two groups. As noted in Chapter 13,
you should also comment on the practical significance of the difference (this is where
inspection of the means is useful).

5. The results can be reported in a simplified version of the table or if referred to in text
using the format $t(df) = t$ Stat to 2 decimal places, $p = p$-value to 2 decimal places, so:
$t(18) = 2.24, p = 0.04$.

(Note: our example uses a very small sample size for illustrative purposes. In reality, such
small sample sizes may be too small to detect differences in the underlying population. In
addition, it is important to ensure that the samples do not violate assumptions that the data are
normally distributed.

3.2 One-way analysis of variance (ANOVA)

Excel’s ANOVA routine also required the data to be in a suitable format. In this case the data
for each group to be included in the analysis should be in adjacent columns (or rows). You
can use the option 1 described in preparing for the $t$-test to create a suitable dataset.

Here we will illustrate how to use Excel’s ANOVA: Single Factor function in the Data
Analysis ToolPak on a set of data reporting the engagement levels of three groups of
employees from different regions, measured on a 1–7 scale (where 7 indicates a high level of
engagement). We will test the following hypotheses:

$H_1$: the engagement level of employees from different regions is different.

$H_0$: there is no difference in the engagement level of employees from different
regions.

3.2.1 Setting up your data for one-way ANOVA in Excel

Excel’s ANOVA routine requires that your dataset is laid out with the data for each group in
separate but adjacent columns (or rows) as shown in the example in Figure 9. Note that the
group sizes do not need to be equal when using Excel’s single-factor ANOVA function. You
can create this layout using option 2 as described in the $t$-test example.
3.2.2 Running ANOVA

To run the test, carry out the following steps:

1. Select Data > Data Analysis > ANOVA: Single Factor > OK to open the test dialogue box.

2. In the dialogue box:
   a. Enter the range of the data for the variables in the Input box. If all columns have column headers (as here) you can include them in the range if desired.
   b. Check that the Grouped By Columns button is selected.
   c. If you have included column headers, tick the Labels box.
   d. Set the desired Alpha (i.e. significance level) for the test. The default is 0.05.
   e. Choose the Output option. If you have space in your worksheet (as here), you can select a suitable Output Range in the active worksheet (Hint: make sure it does not include the dataset or the test output will overwrite it). If your dataset is very large, you may prefer to put the output in a new worksheet, in which case select New Worksheet Ply. The completed dialogue box is shown in Figure 10.
3. Next step is to interpret the output which Excel presents in two tables. Pay particular attention to the following:

4. In the summary table:
   a. Means of each group (this allows you to see the magnitude of the difference between the groups).
   b. The group sizes (=20 for each group); you can check this to make sure no cases have been accidentally omitted from the sample.

5. In the ANOVA table:
a. Between groups and within groups degrees of freedom \((df = 2 \text{ and } 57\) respectively).

b. The test statistic; for ANOVA this is an \(F\)-statistic \((F = 18.04695)\).

c. \(p\)-value (= 8.48E-07). Note that Excel reports very small \(p\)-values using scientific notation. In standard notation this is equal to 0.000000884. When reporting such a small \(p\)-value you can simply state \(p < 0.001\) rather than give the exact figure.

6. Compare the \(p\)-value to your desired significance level (Alpha), in this case 0.05. Since 0.000000848 < 0.05, you would reject \(H_0\), accept \(H_1\) and conclude that there is a statistically significant difference between the means of the three groups. As noted in Chapter 13, you should also comment on the practical significance of the findings.

7. The results can be reported in a simplified version of the table or if referred to in text using the format \(F\)(between groups \(df\), within groups \(df\)) = \(F\) statistic to 2 decimal places, \(p = p\)-value to 2 decimal places, so: \(F(2, 57) = 18.05, p < 0.001\).

3.2.3 Limitations of Excel’s ANOVA routine

There are some limitations to Excel’s ANOVA routine which also illustrate some of the limits of Excel as a statistical package. Firstly, it does not include a routine allowing you to test the assumption of equality of variance amongst the three groups. Secondly, an ANOVA test allows you to infer that there is a statistically significant difference between the groups but not to pinpoint which particular group differences are causing that difference. For this, further tests (known as post hoc tests) are needed. Such tests are not available in standard Excel (they can be run in specialist packages such as IBM SPSS or be generated in Excel with some manipulation).

4 Carrying out tests of association

Excel can be used to carry out two important tests of association introduced in Chapter 13, the chi-squared test and linear regression. It will also calculate Pearson’s \(r\) as described in Data exploration with Microsoft Excel: analysing more than one variable. Other tests can also be carried out but require more extensive manipulation of the data so are not covered in this short guide (see Davis, G. and Pecar, B. (2013). Business statistics using Excel. 2nd ed. Oxford: Oxford University Press for additional guidance).
4.1 Chi-squared test of association

A chi-squared test of association can be carried out using either the CHISQ.DIST.RT function or CHISQ.TEST function. Both require some manipulation of the data; we will show both, starting with the former as it illustrates the principle of the test more clearly and generates important additional information.

4.1.1 Chi-Squared test of association using the CHISQ.DIST.RT

We will illustrate this test using the same dataset that was used in Chapter 13 (Research in practice 13.4). This dataset reports the frequency of car driving by household income and is in the form of a contingency table (Figure 12). We will test the following hypotheses at the 5% (Alpha = 0.05) significance level:

H₁: there is an association between household income and frequency of car driving.

H₀: there is no association between household income and frequency of car driving.

To carry out this test your data must be in a contingency table containing counts (not per cents). If your data are not already in this format, see Data exploration with Microsoft Excel: analysing more than one variable for guidance on creating contingency tables.

Figure 12 – Frequency of car driving by household income

As noted in Chapter 13, the chi-squared test works by comparing the observed frequencies (the ones in the contingency table) with the expected frequencies to derive a test statistic which is then compared to an appropriate chi-squared distribution to calculate the required p-value. Excel’s CHISQ>DIST>RT will calculate the p-value but requires you to calculate the test statistic and the degrees of freedom for the test. We will show each step in turn.

4.1.1.1 Calculating the expected frequencies

The expected frequencies (E) for each cell in the table can be calculated using the formula:
To do this in Excel, simply copy and paste a version of the original contingency table into the worksheet and use the above formula to calculate the expected value for each cell as shown in Figure 12. (Hint: use Paste Special > Values so that you only paste the number values and not formulae when creating your copy, then delete the observed frequency cells, leaving the row and column totals for the calculation.)

You should check that the expected frequency table does not violate any of the restrictions identified in Chapter 13 (i.e. no expected value below 5 in a 2x2 table and not more than 20% below 5 and none below 1 in larger tables).

**Figure 13 – Observed and expected frequencies**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed frequencies (O)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Driving frequency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every day</td>
<td>At least 3 times per week</td>
<td>1-2 times per day</td>
<td>Less than once per week</td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up to £15000</td>
<td>205</td>
<td>110</td>
<td>15</td>
<td>23</td>
<td>353</td>
<td></td>
<td></td>
</tr>
<tr>
<td>£15000 to £25000</td>
<td>228</td>
<td>72</td>
<td>36</td>
<td>13</td>
<td>349</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over £25000</td>
<td>250</td>
<td>82</td>
<td>36</td>
<td>10</td>
<td>478</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>783</td>
<td>264</td>
<td>127</td>
<td>46</td>
<td>1220</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expected frequencies (E)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Driving frequency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every day</td>
<td>At least 3 times per week</td>
<td>1-2 times per day</td>
<td>Less than once per week</td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up to £15000</td>
<td>252</td>
<td>85</td>
<td>41</td>
<td>15</td>
<td>393</td>
<td></td>
<td></td>
</tr>
<tr>
<td>£15000 to £25000</td>
<td>124</td>
<td>78</td>
<td>36</td>
<td>13</td>
<td>240</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over £25000</td>
<td>107</td>
<td>103</td>
<td>50</td>
<td>18</td>
<td>478</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>783</td>
<td>264</td>
<td>127</td>
<td>46</td>
<td>1220</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.1.1.2 **Calculating the chi-squared test statistic**

The chi-squared test stat is given by the following equation:

\[
\text{chi-squared test statistic} = \sum \frac{(O - E)^2}{E}
\]

where \(O\) = observed frequencies for each cell and \(E\) = the corresponding estimated frequencies. The \(\Sigma\) is instructing you to sum all the calculations for each of the cells.

The easiest way to do this is to create a table containing the value of \((O-E)^2/E\) for each cell and then sum all of the cells to give the final answer as shown in Figure 14. This gives a test
statistic of 43.685 in our example. (Hint: \( ^2 \) can be used in the Excel formula bar to calculate the square of a value.)

**Figure 14 – Calculating the test statistic**

Next you need to work out the degrees of freedom \((df)\) for the test. This is given by the equation:

\[
\text{df} = (r - 1) \times (c - 1)
\]

Where \(r\) = the number of rows in the contingency table and \(c\) = the number of columns. In our example there are 3 rows and 4 columns (row and column totals are not counted) so:

\[
\text{df} = (3 - 1) \times (4 - 1) = 2 \times 3 = 6
\]

So the number of degrees of freedom for the test is 6.

### 4.1.1.3 Determining the p-value using the CHISQ.DIST.RT function

We now have the information needed to use the CHISQ.DIST.RT function to calculate the \(p\)-value for the test.

1. Select a destination cell for the function output.
2. Select Formulas > More Functions > Statistical > CHISQ.DIST.RT to open up the Function Arguments dialogue box.
3. In the Function Arguments dialogue box, enter the chi-squared test statistic (or the reference of the cell containing the statistic) into the box marked X (that is probably supposed to be the letter \(\chi\)) and the number of degrees of freedom (or the cell reference) in the box marked Deg_freedom (Figure 15).
4. Click OK to get the output (Figure 16). Note that as with the ANOVA test the p-value is given using scientific notation (8.53401E-8).

5. Next step is to interpret the output. Compare the p-value to your desired significance level (Alpha), in this case 0.05. Since the calculated p-value is less than < 0.05, we would reject H₀, accept H₁ and conclude that there is a statistically significant association between household income and frequency of car driving. As noted in Chapter 13, you should also comment on the practical significance of the findings. 

6. The results can be reported using a simple table showing the test statistic, degrees of freedom, sample size and p-value or if reporting in text using the format \( \chi^2 \) (degrees of freedom, sample size) = chi-squared statistic to 2 decimal places, \( p = p \)-value to 2 decimal places, so in this example: \( \chi^2 \) (6, 60) = 43.69, \( p < 0.001 \).
4.1.2 Chi-Squared test of association using the CHISQ.TEST function

You can carry out the same test using the CHISQ.TEST function as an alternative. To do this you need to create a table containing the expected cell frequencies in the same way as when using the CHISQ.DIST.RT function but this time you do not need to go and calculate the chi-squared statistic or degrees of freedom.

To use the CHISQ.TEST function:

1. Prepare your data in a contingency table showing observed frequencies.
2. Create a table of expected frequencies.
3. Select a destination cell for the function output.
4. Select Formulas > More Functions > Statistical > CHISQ.TEST to open the Functions Argument dialogue box.
5. In the dialogue box enter the range of observed frequencies from the contingency table in the box marked Actual_range and the expected frequencies in the box marked Expected_range (remember not to include column and row totals in the range). See Figure 17.

Figure 17 – CHISQ.TEST Function Argument dialogue box

6. Click OK. The output is shown in Figure 18. It is the same as for the CHISQ.DIST.RT function and should be interpreted in the same way.
The CHISQ.TEST function is slightly quicker to apply but for reporting test results it is standard practice to include details of the test statistic and degrees of freedom. For this reason, we would suggest using the CHISQ.DIST.RT function or at least calculate the necessary details as shown in the example for that function. You can always run both functions as a check for possible mistakes in preparing the chi-squared statistic.

5 Linear regression

Excel’s Regression routine in the Data Analysis ToolPak can be used to perform bivariate and multivariate linear regression. We will illustrate its use for bivariate regression by analysing the relationship between satisfaction and commitment in the customer satisfaction database. The researcher has developed the conceptual model shown in Figure 19:

We will test the following hypotheses, derived from Figure 19, at a 5% significance level:

H$_1$: Customer satisfaction has an impact on customer commitment.

H$_0$: Customer satisfaction has no impact on customer commitment.
5.1.1 Exploring the relationship using scatterplot and Pearson’s r

An assumption of linear regression is that the relationship between the two variables is linear. A scatterplot can be used to test this assumption (see Data exploration with Microsoft Excel: analysing more than one variable for further details). Figure 20 shows a scatterplot of satisfaction against commitment. The plot indicates that the relationship is fairly linear and is positive, as suggested by the conceptual model.

**Figure 20 – Scatter plot of satisfaction against commitment**

![Scatterplot of satisfaction and commitment](image)

You can also use Excel’s chart functions to display a regression line, the regression equation and the $R^2$ value. To do so, right click on one of the data points in the scatter plot. From the menu select Add Trendline to open up the Format Trendline dialogue box. In the Trendline Options view, select Linear from the Trend/Regression Type and tick Display Equation on chart and Display R-squared value on chart boxes (see Figure 21).

**Figure 21 – Format Trendline dialogue box**

![Format Trendline dialogue box](image)
Click Close. The output is shown in Figure 22. This is a quick way of inspecting the regression equation and $R^2$ but although these values will be the same as those produced by a bivariate regression test (for an interpretation see Chapter 13 and below), they do not allow us to say anything about the statistical significance of the findings. To do this you will need to use the Regression routine in the Data Analysis ToolPak.

Figure 22 – Regression line, regression equation and $R^2$ fitted to the scatter plot

5.1.2 Running linear regression

To run the test, carry out the following steps:

1. Select Data > Data Analysis > Regression > OK to open the test dialogue box.

2. In the dialogue box:
   a. Enter the range for the dependent variable (in this case Commitment) in the Input Y Range box.
   b. Enter the range for the independent variable (in this case Satisfaction) in the Input X Range box. (Note: for multiple regression you can add multiple independent variables if they are in adjacent columns by selecting all of them in the range.)
   c. If you have included column headers in the Input ranges (as here) tick the Labels box.
   d. Choose the Output option. The default is New Worksheet Ply which is recommended. The completed dialogue box is shown in Figure 23.
e. Click OK. The resulting output is shown in Figure 24 with the column spacing adjusted and all output set to three decimal places to make it easier to read (as a result the p-values no longer appear in scientific notation). (Note: Excel’s output duplicates Lower and Upper confidence interval columns. The duplicates have been deleted for clarity.)

![Figure 23 – Regression dialogue box](image)

![Figure 24 – Regression analysis output](image)

3. Next step is to interpret the output which Excel presents in three tables. Pay particular attention to the following:

4. In the regression statistics table:
   a. $R^2$ (R Square = 0.639).
b. The sample size (Observations = 20).

5. In the ANOVA table which tests the significance of the $R^2$:
   a. Regression and residual degrees of freedom ($df = 1$ and $18$ respectively).
   b. The test statistic; for ANOVA this is an $F$-statistic ($F = 31.849$).
   c. The $p$-value ($= 0.000$). Compare the $p$-value to your desired significance level (Alpha), in this case 0.05. Since $0.000 < 0.05$, you would conclude that the $R^2$ is statistically significant.

6. In the coefficients table:
   a. The $b$ coefficient (Satisfaction = 0.892).
   b. The test statistic (t Stat = 5.643).
   c. The $p$-value ($= 0.000$). Compare the $p$-value to your desired significance level (Alpha), in this case 0.05. Since $0.000 < 0.05$, you would conclude that the $b$ coefficient is statistically significant.
   d. The 95% confidence interval (0.560, 1.224) which indicates a range of plausible values for the $b$ coefficient.

7. On the basis of this analysis you would reject $H_0$ at the 5% level and conclude that customer satisfaction impacts on commitment. As with other tests, you should also comment on the practical significance of your findings.

8. The results of regression analysis are most easily reported in tabular form.

5.1.3 Testing the assumptions of regression analysis

As noted in Chapter 13, linear regression analysis requires a number of assumptions to be met. In addition to using a scatterplot to check the linearity of the relationship as discussed above, Excel can help test other assumptions. Two of these are:

- Normality of errors. The difference between the observed values and the values predicted by the regression model are known as errors or residuals. The test assumes that these errors are normally distributed for each value of the independent variable. You can test this assumption using a normal probability plot of the residuals. To create this, tick the Normal Probability Plot box in the Regression dialogue box when you run the test. The output is shown in Figure 25 (a straight line has been added using the Add Trendline option described above). For the assumption to hold, the points should lie in a reasonably straight line as here.
Fig. 25 – Normal probability plot with added trendline

- Constant variance of errors. The assumption of homogeneity of variance (homoscedasticity) requires that the variance of the errors is constant across all values of the independent variable. You can check this visually using a residual plot. To create such a plot, tick the Residual Plot box in the Regression dialogue box when you run the regression analysis. The output is shown in Fig. 26. For the assumption to hold, the plot should show no observed pattern, for example if the error value changes considerably as the value of the independent variable changes. That does not seem to be the case in this example, so the assumption of constant variance is reasonable.

Fig. 26 – Residual plot

5.1.4 Limitations of Excel’s Regression routine

Excel’s regression routine provides a useful way of carrying out bivariate regression but, although it can run a multiple regression analysis, it lacks the supporting tools and diagnostics for in-depth analysis of multivariate problems. For this reason, it is probably better to use specialised statistics software such as SPSS for multiple regression problems.