Data exploration with Microsoft Excel: analysing more than one variable

Contents

1 Introduction........................................................................................................................................1
2 Comparing different groups or different variables .................................................................2
3 Exploring the association between categorical variables......................................................12
4 Exploring the association between two metric variables.......................................................16
5 Plotting time series.......................................................................................................................20

1 Introduction

This guide covers the use of Microsoft Excel (hereafter: Excel) for exploring more than one variable at the same time (bivariate and multivariate analysis). It shows how 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. Comparing different groups or different variables
2. Exploring the association between nominal or ordinal variables
3. Exploring the association between two metric variables
4. Plotting time series

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 information on how to use functions and the Data Analysis ToolPak see the guide Introduction to using Microsoft Excel for quantitative data analysis).
2 Comparing different groups or different variables

If you have used metric variables as part of your data collection, you may want to compare them in terms of their means or other descriptive statistics. Alternatively, you may wish to compare different groups in terms of their means or other descriptive statistics for relevant metric variables. Here we will show how you can use Excel to assist you in carrying out such comparisons for your sample data.

2.1 Comparing means and other descriptive statistics for different variables

You can use Excel’s statistical functions or the Descriptive Statistics function in the Data Analysis ToolPak to generate relevant statistics for comparison. It is usually helpful to calculate a range of measures, including central tendency, dispersion and skewness/kurtosis to give you a good understanding of the different variables. Use tables both to compare and present your comparisons. Figure 1 shows such a table created from a small dataset on customer satisfaction using the statistical functions listed in Table 1 and reporting a range of descriptive statistics for the two metric variables in the dataset. See Introduction to using Microsoft Excel for quantitative data analysis for more details on using functions.

Figure 1 – Tabular summary of descriptive statistics for customer satisfaction and customer commitment (n = 20)

<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>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>Gender</td>
<td>Store format</td>
<td>Store location</td>
<td>Satisfaction</td>
<td>Commitment</td>
<td>Variable</td>
<td>Mean</td>
<td>Median</td>
<td>Std Dev</td>
<td>Min</td>
<td>Max</td>
<td>Skewness</td>
<td>Kurtosis</td>
</tr>
<tr>
<td>2</td>
<td>Female</td>
<td>Magistore</td>
<td>North</td>
<td>3</td>
<td>3</td>
<td>Customer satisfaction</td>
<td>4.25</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>-0.34</td>
<td>-0.34</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>Convenience</td>
<td>Central</td>
<td>4</td>
<td>3</td>
<td>Customer commitment</td>
<td>4.6</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>-0.34</td>
<td>-0.34</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Female</td>
<td>Magistore</td>
<td>South</td>
<td>5</td>
<td>6</td>
<td>Customer satisfaction</td>
<td>4.25</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>-0.34</td>
<td>-0.34</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>Convenience</td>
<td>Central</td>
<td>5</td>
<td>5</td>
<td>Customer commitment</td>
<td>4.6</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>-0.34</td>
<td>-0.34</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 – Excel statistical functions used in Figure 1 (in order of use, left to right)

<table>
<thead>
<tr>
<th>Function name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>Returns the arithmetic mean (average) of the given numbers</td>
</tr>
<tr>
<td>MEDIAN</td>
<td>Returns the median of the given numbers</td>
</tr>
</tbody>
</table>
### Function name | Description
--- | ---
STDEV.S | Returns the standard deviation of the given numbers, based on a sample
MIN | Returns the minimum value of the given numbers
MAX | Returns the maximum value of the given numbers
SKEW | Returns the skewness of a dataset
KURT | Returns the kurtosis of a dataset

#### 2.2 Comparing means and other descriptive statistics for groups

You can compare means and other descriptive statistics for different groups (such as male/female) within your dataset by manually sorting the data by group (Hint: use the sort and filter commands in the Home tab) and then calculating the relevant statistics for each group using functions or the Descriptive Statistics routine as before. This can be tedious and error prone, particularly if the dataset is very big. Pivot tables, on the other hand, provide a quick way of carrying out such analysis. Figure 2 shows such a pivot table, based on the dataset used in Figure 1, that compares male and female customers in terms of their satisfaction levels. It includes the mean, standard deviation and sample size for each group.

**Figure 2 – Pivot table showing customer satisfaction by customer gender (n = 20)**

To create a pivot table like this, carry out the following steps:

1. Click on any cell in the dataset to select it.
2. Select Insert > PivotTable > PivotTable to open up the Create PivotTable dialogue box (see *Data exploration with Microsoft Excel: univariate analysis* for more details on how to do this).
3. In the dialogue box select the data range and choose where you want the PivotTable report to be placed (New Worksheet is the default and recommended).
4. Click OK to create an empty PivotTable report and PivotTable Field List in a new worksheet.
5. To create the table, start by dragging and dropping the field ‘Gender’ into the Row Labels area. Gender will act as the grouping variable and give your row headers. Next, drag and drop a copy of the field ‘Satisfaction’ into the Values area. This will create a column showing the sum of the Satisfaction variable for each group and for the grand total as shown in Figure 3 (Note: Satisfaction is a metric variable so Excel’s default action is to sum it). We will need three columns of satisfaction-related data in our table so repeat this twice to get three columns all showing the same output (Figure 3).
Figure 3 – PivotTable Field List and pivot table with three columns showing Sum of Satisfaction by gender

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Row Labels</td>
<td>Sum of Satisfaction</td>
<td>Sum of Satisfaction</td>
<td>Sum of Satisfaction</td>
</tr>
<tr>
<td>4</td>
<td>Female</td>
<td>37</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>6</td>
<td>Grand Total</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
</tbody>
</table>

6. The next step is to change the data displayed to something more useful for our purposes. We want the three columns to show the mean score, the standard deviation and the sample size. To do this:

7. Click on the down arrow of the first Sum of Satisfaction fields in the Values area of the Pivot Table Field List and choose Value Field Settings from the menu that opens.

8. Select the Summarize Values By tab if this is not already selected. Choose Average (this is the Mean in Excel). Type Mean in the Custom Name box (Figure 4).

Figure 4 – Value Field Settings dialogue box
9. If desired, you can change the number format by clicking on the Number Format button to open up a new dialogue box. In this case we have set the number format to Number, with 2 decimal places.

10. Click OK. This applies the new settings to the first column in the pivot table report (Figure 5).

**Figure 5 – Pivot table with first column showing mean by group**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Row Labels</td>
<td>Mean</td>
<td>Sum of Satisfaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Female</td>
<td>3.70</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>4.80</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Grand Total</td>
<td>4.25</td>
<td>85</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11. Now repeat the procedure for column 2 by following steps 7 to 9 but this time choosing StDev (this is the standard deviation for a sample) in the Summarize Values By box, naming it Std Dev and again setting the number format to 2 decimal places. Click OK and the settings will be applied to the second column.

12. Repeat steps 7 to 9 once more, this time choosing Count Numbers in the Summarize Values By box, allocating the custom name n (for sample size). There is no need to change the number settings as this will return a straight count of how many in each group. Click OK to apply the settings. The finished pivot table is shown in Figure 6 along with its PivotTable Field List.

**Figure 6 – Completed pivot table showing customer satisfaction by customer gender (n = 20)**
Having created the comparison table, it is easy to see that female customers are less satisfied than male customers in the sample data. The spread of ratings appears similar (as measured by the standard deviation) and the sample size of the two groups is equal (i.e. 10 per group).

2.3 Visualising differences

Excel’s chart tools can be used to create graphs of means or other statistics to allow visual comparison. If your data are already in a suitable format or you have created tables of means (as in Figure 1), you can create charts of these using the charts available in the Insert tab.

Pivot charts can also be used; they are especially useful when you wish to visually compare groups. You can create them from an existing pivot table or directly from the PivotChart command. We will demonstrate their use by creating a vertical bar (Excel: column) chart comparing the mean satisfaction level by customer according to store location from the customer satisfaction database.

1. Click on any cell in the dataset to select it.
2. Select Insert > PivotChart to open the Create PivotChart with PivotTable dialogue box. Choose the data table/range if not already entered and select the location for the output (New Worksheet is the default). Click OK.
3. This creates a blank PivotTable report, PivotChart and PivotTable Field List (Figure 7).
4. To populate the chart, drag and drop a copy of the ‘Store location’ field into the Axis Fields (Categories) area.

5. Next, drag and drop a copy of the ‘Satisfaction’ field into the Values area.

6. A PivotChart in the form of a bar (Excel: column) chart is created along with a PivotTable report of the data (Figure 8). The chart shows the Sum of Satisfaction which is the default for metric data.
7. To change the chart to show the mean satisfaction level by store location, click on the down arrow of the Sum of Satisfaction field in the Values area of the Pivot Table Field List and choose Value Field Settings from the menu that opens.

8. Select the Summarize Values By tab if this is not already selected. Choose Average (this is the Mean in Excel).

9. Click OK. The chart now shows the mean satisfaction levels by store location (Figure 9). The data show small differences in satisfaction levels between the three locations in the sample data.

If desired you can now format the chart using the PivotChart tools as we have done in Figure 10. (Hint: the grey Field Buttons [Average of Satisfaction and Store Location] can be hidden by selecting PivotChart tools > Analyse > Field Buttons > Hide All.)
2.4 Adding additional variables to pivot tables and pivot charts

You can add additional variables to pivot tables and pivot charts to produce cross-tabulations that allow you to explore the differences between groups in greater detail. Figure 2 showed a difference in satisfaction level between male and female customers in the sample. We can further explore potential differences between customers of different gender by introducing a third variable. The pivot table in Figure 11 does that by comparing satisfaction levels by gender by store location. Notice that the difference between males and females in the sample almost disappears if we control for where they shop. Both males and females show similar levels of satisfaction when they shop in the same store format (both higher in convenience stores than megastores). Adding an additional variable in this way has given us an insight into the data that the original analysis did not reveal.

Figure 11 – Customer mean satisfaction level by gender by store type (n = 20)

To create a cross-tabulation of this kind, create a pivot table showing mean satisfaction level by gender. Then:

1. In the PivotTable Field List, drag a copy of the field ‘Store format’ into the Column Labels area.
2. This will create column headers for both the store formats (Convenience and Megastore) and Grand Total (Figure 12).

Figure 12 – PivotTable report and PivotTable Field List for cross-tabulation of gender and store format

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average of Satisfaction</td>
<td>Column Labels</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Male</td>
<td>Convenience</td>
<td>Megastore</td>
<td>Grand Total</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Female</td>
<td>3.00</td>
<td>3.06</td>
<td>3.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>3.13</td>
<td>3.50</td>
<td>4.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Grand Total</td>
<td>5.10</td>
<td>3.40</td>
<td>4.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The same logic can be applied to creating a pivot chart to display the data. Create a new pivot chart as before then:

1. In the PivotTable Field List, drag and drop a copy of the Gender field into the Axis Fields (Categories) area and the Satisfaction field into the Values area.
2. Use the Value Field Settings dialogue box to set the Satisfaction values to Average (i.e. the mean).
3. Now in the PivotTable Field List, drag and drop a copy of the Store format field into the Legend Fields (Series) to create a clustered bar (Excel: column) chart showing satisfaction levels by gender by store format (Figure 13).
4. You can now edit the chart using the PivotChart tools as required.

Figure 13 – Clustered bar chart of customer mean satisfaction levels by gender and store format ($n = 20$)
You can add more variables for still greater resolution but in practice it becomes increasingly hard to interpret and unless the dataset is very large the sample sizes in individual cells can get very small (or non-existent).

3 Exploring the association between categorical variables

Pivot tables can also be used in Excel to create contingency tables of the type introduced in Chapter 13 for exploring the association between categorical (non-metric) variables. Similarly, pivot charts can be used to visualise such associations.

3.1 Contingency tables

We will use the customer satisfaction database to create a contingency table showing shopping frequency by gender and store location. To create the table, carry out the following steps:

1. Click on any cell in the dataset to select it.
2. Select Insert > PivotTable > PivotTable to open up the Create PivotTable dialogue box.
3. In the dialogue box select the data range and choose where you want the PivotTable report to be placed (New Worksheet is the default and recommended).
4. Click OK to create an empty PivotTable report and PivotTable Field List into a new worksheet.
5. To create the contingency table, start by dragging and dropping the field ‘Gender’ into the Row Labels area. Gender will act as the grouping variable and give your row headers. Next, drag and drop another copy of the field ‘Gender’ into the Values area. This will create a frequency table showing the count of Gender for each group and for the grand total as shown in Figure 14 (Note: Gender is a nominal variable so Excel’s default action is to count the number of times it occurs).
6. To create the contingency table, drag and drop a copy of the field ‘Store location’ into the Column Labels area. This creates column headers for each of the store locations (North, Central and South) and the grand total. The cells show the number of customers by gender for each store location (Figure 15).

Figure 15 – Contingency table of customer store location by gender (counts) (n = 20)

7. As explained in Chapter 13, contingency tables containing counts can be difficult to interpret, especially if group/category sizes are very different or the dataset is very large, so it is common to convert the counts to per cents. To do so:

8. Click on the down arrow of the Count of Gender field in the Values area of the Pivot Table Field List and choose Value Field Settings from the menu that opens.
9. Select the Show Values As tab and then select % of Row Total in the Show Values as drop down box.

10. Click OK. The table now shows the per cent of each gender shopping in each store location (Figure 16). The data indicate that over half of the male customers in the sample shop in the central location, while only 10% of female customers do so.

Figure 16 – Contingency table of customer store location by gender (per cent) (n = 20)

3.2 Visualising associations between categorical variables

Pivot charts can be used to create visualisations of contingency tables. They can be generated from existing pivot tables or direct from the data using the PivotChart command in the Insert tab. To create a pivot chart from an already existing contingency table such as that in Figure 16, do the following:

- Click one of the cells in the pivot table (Note: the values have been set to counts instead of per cents).
- Select Insert > Column [chart] > 2D Column > Clustered Column. This inserts a clustered bar chart as a pivot chart into the worksheet (Figure 17).
You can change the chart type if desired by selecting PivotChart tools > Design > Change Chart Type and then selecting the type of chart you want from the Change Chart Type dialogue box. A 100% stacked bar (Excel: column) chart, for example, could be used to emphasise the difference in proportions of males and females that shop in the different store locations (Figure 18).

It is possible to include more variables in either pivot charts or pivot tables but as pointed out earlier, they become harder to interpret and unless the dataset is very large the sample sizes in individual cells can get very small (or non-existent).
4 Exploring the association between two metric variables

Excel can also be used to explore the association between two metric variables using scatterplots and by calculating Pearson’s correlation coefficient ($r$).

4.1 Creating scatterplots

Scatterplots are part of Excel’s suite of chart types. Figure 19 shows a scatterplot of satisfaction and commitment from the customer satisfaction database. It was created by selecting the range to be charted and then via Insert > Scatter > Scatter with only Markers. The default output is shown.

Figure 19 – Scatterplot of satisfaction and commitment ($n = 20$)

Once you have created the basic chart it can be formatted using the Excel Chart Tools. Figure 20 shows the same scatterplot after formatting.

Figure 20 – Scatterplot of satisfaction and commitment (formatted) ($n = 20$)
The scatterplot clearly shows a positive, fairly linear and quite strong relationship.

### 4.2 Calculating Pearson’s correlation coefficient (r)

You can use the Excel statistical function PEARSON to calculate Pearson’s correlation coefficient (r). We show it here being used to calculate Pearson’s r for the variables in the scatterplot in Figure 20.

1. Select the destination cell for the output.
2. Select Formulas > More Functions > Statistical > PEARSON to open the Function Arguments dialogue box (Figure 21).

**Figure 21 – Function Argument dialogue box for PEARSON**

3. Enter the data range for one variable into the box Array1 and the data range for the other variable into Array2.
4. Click OK. The output appears in the destination cell (Figure 22). As expected from the scatterplot the association is positive and very strong (> 0.7).

**Figure 22 – Pearson’s correlation coefficient (r)**
4.2.1 Excel’s Correlation function

An alternative way of calculating Pearson’s $r$ is to use the Correlation function in the Data Analysis ToolPak. This produces a correlation matrix for two or more variables showing Pearson’s for all combinations of pairs. It is therefore particularly useful when you have a larger number of variables.

To use the Correlation function:

1. Select Data > Data Analysis > Correlation > OK to open the Correlation dialogue box.
2. In the dialogue box, enter the data to be analysed in the Input Range box (Note: all the variables to be included in the analysis must be in adjacent columns).
3. Tick Labels in first row if you have included column headers.
4. Select an Output option. New Worksheet is the default and recommended (Figure 23).

Figure 23 – Correlation dialogue box
5. Click OK to create the correlation matrix (Figure 24). The matrix shows the correlation coefficients for all possible combinations of the two variables included in the analysis (the 1s on the diagonal arise because a variable is perfectly correlated with itself).
5 Plotting time series

Excel’s Line Charts can be used to plot time series data if required. Ensure that time is on the horizontal axis. Figure 25 shows an example of a simple line graph for a small time series dataset.

Figure 25 – Example time series plotted as a line chart