Survey Research – Data Analysis

Overview

Once survey data are collected from respondents, the next step is to input the data on the computer, do appropriate statistical analyses, interpret the data, and make recommendations pertaining to your research objectives. Steps in data analysis include: (i) editing and coding survey data, (ii) inputting them in the computer in a software-readable format, (iii) doing basic analysis such as frequency distribution, means analysis and cross-tabulation to generate insights, (iv) testing hypotheses where pertinent, and, if needed, (v) resorting to higher order analyses such as correlation and regression. This module will emphasize the practical applications of the statistical techniques and interpretation, rather than on the technical aspects or statistical details. In addition, students will learn how to perform data analysis using Excel.

Learning Objectives

Upon successful completion of this module, you will be able to:

- Write a code sheet and enter the data in an Excel spreadsheet
- Construct and interpret frequency distributions
- Compute means and proportions using Excel
- Perform statistical tests
- Analyze and interpret cross-tabulation data
- Perform correlation and regression analyses

Readings

Complete the readings for this module in the order suggested below.

1. Data Editing and Coding (handout in class)
2. Data Analysis – Basics (handout in class)
3. Data Analysis – Frequency Distribution (handout in class)
5. Data Analysis – Means and Standard Deviations (handout in class)
6. Data Analysis – Proportions (handout in class)
8. Data Analysis – Cross Tabulation (handout in class)
9. Data Analysis – Correlation (handout in class)
10. Data Analysis – Regression (handout in class)
11. Writing a Research Report (handout in class)
12. Summary (handout in class)
Data Editing and Coding

First, the completed questionnaires must be edited; potentially invalid or inaccurate questionnaires should be eliminated. Invalid, and therefore unusable, responses can be identified in different ways. If respondents take relatively very little time to complete the questionnaire or they mark the same point (say 7 on a 7-point scale) for all questions, they are indicators of invalid answers. Unusable questionnaires would be those in which respondents have failed to complete a significant portion of the questionnaire. In paper-and-pencil surveys, respondents sometimes fail to turn the page over and so fail to complete an entire page of the survey. For this reason, it is always best to point out explicitly if questions are on both sides of the sheet. In online surveys, respondents may skip screens intentionally or inadvertently. Design of an online questionnaire should try to control for this possibility.

Coding is the process of assigning numeral or character codes to all responses for each and every question in the survey. This enables us to enter data into Excel for analysis. We will illustrate coding by using the shoe purchase survey on the class website.

Question A1. First, you must know what type of scale it is – nominal, ordinal etc (refresh your memory from Module 4 – Theme 2.) This question has a nominal scale – used simply for classification. So any numbering system with Yes and No getting different numbers is a valid code. Examples: Yes = 1; No = 2; Yes = 1, No = 0; Yes = 5, No = 8 etc. As a convention, we use the code Yes = 1 and No = 0. Why might a researcher want to use this coding system? Yes = 1 and No = 2 is also commonly used.

Question A2. Recognize that this is an interval scale. So, any numbering system in which the uniform interval is maintained is acceptable. A simple coding would be to keep the numbers as is: 1=1, 2=2, …, 7=7. Another permissible and commonly used coding systems is the following: 1=-3, 2=-2, 3=-1, 4=0, 5=1, 6=2, 7=3. Note that uniform interval property is maintained. Why might a researcher want to use this coding system?

Question A3. Note that this is an open-ended question (respondents can write in their own words). There are two ways to handle this. One is to put what is said verbatim in a separate Excel column. The other is to put all open-ended statements in a separate word document. Then, analyze the open-ended statements for content and compile some general insights.

Question A4. This is an ordinal scale. There is an ordering in the range of prices that consumers would pay for dress shoes. Any numbering system that maintains the ordering is fine. Example: 1 = $1-19.99; 2=$20-39.99……9 = $160-179.99, 10 = $180 or above. You might find it useful to employ a more informative and legitimate coding system: 1= $1-19.99; 20=$20-39.99……160 = $160-179.99, 180 = $180 and above. Why is it a legitimate coding for this question and why might this coding be more informative?
In a later module, you will see how to use this type of ordinal data to approximate interval responses.

**Question A5.** This is a nominal scale. Any set of codes that uses a different number for each type of store would be valid. A simple coding would be: 1= specialty shoe store, 2=discount shoe store, 3=factory outlet, 4=department store, 5=discount store, 6=other.

**Question A6.** This is also a nominal scale. So, the temptation would be to use the same coding scheme as in Question A5: 1= Low price, 2 = Variety of shoes, and so on. This would be wrong. The big difference between Question A5 and A6 is that A5 says check one where as question A6 says check all that apply. So, in A5, one could check “Low Price,” “Variety of Shoes,” and “Good Salespeople.” A coding system such as: 1= Low price, 2 = Variety of shoes, and so on would not suffice. For any question that says check all that apply, you need to open as many columns as there are options (create that many variables) and use 0-1 (or similar) coding where 0 implies that the option was not checked and 1 implies the option was checked. For Question A6, you need to have 6 variables with 0-1 coding as follows:

- Low Price = 0 if not checked, 1 if checked
- Variety of Shoes = 0 if not checked, 1 if checked
- ....

You should be able to code the background questions **B1-B3**.

An Excel code sheet for the questionnaire is available on the course website.
Data Analysis – Basics

I will illustrate the data analysis techniques using data from the shoe questionnaire (filename: shoedata.xls) and coding sheet from the website.

Most of the statistical analysis is performed with an Excel add-in called Analysis ToolPak. To see if the Analysis ToolPak is loaded, click the “Office” button on the top left of the spreadsheet, then click on the “Excel Options” button. Select the "Add-Ins" tab on the left then click the "Go" button next to "Manage Excel Add-Ins." “Analysis ToolPak” should be checked. If not, check it and click on the "OK" button. A slightly different procedure is required for earlier versions of Excel. In either case, the Analysis ToolPak comes standard with the Microsoft Office installation.

Make sure the data are entered correctly

- Take a hard copy of the data and eyeball to make sure there are no unexpected numbers.
- Verify with the original survey to make sure that entries are correct.

In Excel, you can specify a variable by the range of cells corresponding to that variable. By assigning a name to the variables up front, you do not have to reference the range every time.

Assign a name to a range of variables

- Make sure the first row in the column has the name you want the values to be called.
- Select Range including the top name row.
- Click Insert > Name > Create
- Verify that the Top Row box is checked
- Click OK

Analyze the Data

Most data analysis for survey projects involves:
- Computing frequencies, percentages, proportions and drawing histograms or pie charts
- Computing means (averages) and standard deviations
- Descriptive statistics including mean, median, mode, minimum, maximum.
- Cross-Tabulation

You should be guided by two general rules when deciding on the nature of data analysis to be performed.
**General Rule #1.** Your analysis should be directed by your research objectives. Take each question and consider the objective(s) that relates to the question. Do the analysis so that the objective(s) can best be addressed.  

**General Rule #2.** The type of analysis you do should also be driven by the type of scale that you have used for the question. For:

- **Nominal** – Analyze using frequency distributions, proportions, mode, cross-tabulation.
- **Ordinal** – Typical analyses include frequency distributions, proportions, medians.
- **Interval and Ratio** – You can do frequency distributions (e.g., for 5 or 7-point scale), analysis of proportions plus analyses of means (averages) and standard deviations.

While doing these analyses, you may also need to sort the data or recode some of the data. The following sections describe the steps involved in performing these analyses.
Data Analysis – Frequency Distribution

The frequency distribution is a simple, but popular and very useful analysis.

It is used for:
- Detecting errors in data entry or outliers (extreme values)
- Assessing whether or not the sample is representative of the population.
- Gaining insights into attitudes and consumer behavior.

A note about outliers…. Outliers are respondents whose responses are far away from the central tendency of the sample (mean, median or mode). A negative consequence of outliers is that they can have a big impact on the mean computed from the sample. On the other hand, outliers can be particularly important in some cases. For example a company’s largest customer might be an outlier in terms of its size or purchases. Yet this customer could be very important to the company. Because of their potential impact, after having identified outliers, it is important to recheck their responses to ensure that they are not in error (entered or coded incorrectly).

To create a histogram in Excel, first you must create a “Bin.” In the bin, you set the class or value range for which frequencies have to be computed. Excel can automatically create a set of evenly distributed bins between the data’s minimum and maximum. In general, it is better to create your own bin values.

Example – suppose I want to get a frequency distribution of DSPURCH (whether you have purchased a dress shoe in the last two years – Yes (1), No (0). I want to set a bin with 0 and 1 value in some column on the Excel Table.

<table>
<thead>
<tr>
<th>bin</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>94</td>
</tr>
</tbody>
</table>

To get a frequency distribution and chart output,
- Click Data > Data Analysis > Histogram
- Specify Input Range (variable name or cell range – here b1:b103) in the Dialog box
- Specify bin range
- Click the labels box
- Click the Chart output box
- You may ask for cumulative percentage – I don’t generally find it very useful
- Specify where you want the output.

The result is a histogram like the one below.
What the “Histogram” function does is count the number of observations between two numbers in the bin. “More” indicates any value above the last number in the bin.

The above graph is not so informative and ugly! You can (and should) manipulate the graph in Excel to make it more informative and professional-looking.

You can remove the "More" bin from the chart by deleting it from the table to the left. You can activate the chart by clicking on it.

- Drag the chart wherever you want it
- Elongate by pulling on its edges
- Click on frequency legend and delete it.
- Double click on shaded portion and check none to remove the grey shading.
- Change numbers on the bin table to the left to be more descriptive; this changes the corresponding numbers on the chart.
- You can insert text by clicking on drawing on top right corner and then clicking on text like box.
- You can change the scale for the Y-axis by double clicking to the Y-axis and then clicking on scale.

Your turn – modify your histogram!

After your chart is modified, it may look something like the one below (or even better!)
Note that we have changed the raw purchase frequencies to percentages; this makes the chart much easier to read and understand. Also, use labels to make the figure as self-explanatory as possible.

What story can you tell about the chart above?

Frequency distributions of key demographic variables can be used to judge the nature and representativeness of the sample. For example:

What can you say about this annual household income distribution chart? Do you think this is representative of a big city in the United States?

You can also use the “Chart Wizard” in Excel to create various bar charts or pie charts. Pie charts sometimes look better when we are presenting percentage data. Click on the “Chart Wizard” icon and follow instructions. To use the Chart Wizard, you have to input the frequencies. That is, Chart Wizard does not compute frequencies from raw data. Only “Histogram” under data analysis tools computes the frequencies. Once frequencies are computed, the Chart Wizard enables you to create nice graphical representations.
Data Analysis – Means and Standard Deviations

Measures of Central Tendency

There are three basic measures of central tendency—mean, median and mode. Consider the scores (out of 100) for 7 students given in column B of the following table.

<table>
<thead>
<tr>
<th>Student (i)</th>
<th>Score (x_i)</th>
<th>Mean (( \bar{x} ))</th>
<th>Deviation (( x_i - \bar{x} ))</th>
<th>Deviation squared (( (x_i - \bar{x})^2 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>57.4</td>
<td>-22.4</td>
<td>501.76</td>
</tr>
<tr>
<td>2</td>
<td>65</td>
<td>57.4</td>
<td>7.6</td>
<td>57.76</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>57.4</td>
<td>17.6</td>
<td>309.76</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>57.4</td>
<td>-32.4</td>
<td>1049.76</td>
</tr>
<tr>
<td>5</td>
<td>45</td>
<td>57.4</td>
<td>-12.4</td>
<td>153.76</td>
</tr>
<tr>
<td>6</td>
<td>82</td>
<td>57.4</td>
<td>24.6</td>
<td>605.16</td>
</tr>
<tr>
<td>7</td>
<td>75</td>
<td>57.4</td>
<td>17.6</td>
<td>309.76</td>
</tr>
<tr>
<td>Total</td>
<td>402</td>
<td></td>
<td></td>
<td>2987.72</td>
</tr>
</tbody>
</table>

**Mean.** The mean, or average, is the most commonly used measure of central tendency. For the seven students in the table above, the mean score is computed by adding the student scores, then dividing by the total number of students. More generally, the mean is computed using the following equation:

\[ \text{Mean, } \bar{x} = \frac{\sum x_i}{n}, \]

where \( \sum x_i \) represents the sum (addition) of all elements of \( x_i \), \( i = 1, \ldots, n \). \( x_i \) is the value for the \( i \)th element (in our case, the \( i \)th student's score).

For the seven students, the mean = \((35+65+75+25+45+82+75)/7 = 402/7 = 57.4\)

Means are always reported for interval-scaled and ratio-scaled variables.

**Median.** The median is the value of the center-most element, when all elements are arranged in order. We can arrange the above seven numbers in ascending order of scores as 25, 35, 45, 65, 75, 75, 82. The center element is the fourth one with score of 65. Hence, median score is 65 (which was earned by Student 2). Observe that there were seven responses, an odd number. If there were an even number of responses, then there would be no one single center-most element; there would be two. In that case, the median value is the average of the two center-most elements.

Median (not mean) is the appropriate measure of central tendency to use when analyzing ordinal, or ranked, data. For example, let us say five different drivers rank
Ford cars as 1, 3, 4, 2, 3. It would be correct to report a median rank of 3; it would be incorrect to report a mean rank of 2.5.

**Mode.** Mode is the value of the element observed most frequently. In the 7-student example, the modal score is 75, because it has a frequency of two (two students earned that score). All other scores have a frequency of one (they occurred only once).

**Measures of Dispersion**

**Range.** The simplest measure of dispersion in the data is its range. Specifically: What is the *maximum* observed value? What is the *minimum* observed value?

Returning to the student scores in the previous table, we find that the minimum score is 35 (earned by student 1) while the maximum score is 82 (earned by student 6). In other words, the scores range from 35 – 82.

**Variance and standard deviation.** More informative measures of data dispersion are statistics called the variance and standard deviation. They are commonly computed in data analyses when using means. Intuitively, these statistics measure the extent to which observations are “spread” around the mean. For example, the first student scores 35 points and the mean is 57.4; so the student has deviated from the mean by 22.4 points (see table above). The deviation can be either positive or negative. To “penalize” both positive and negative deviations the same amount, we square each deviation (multiply it by itself). We then sum the squared deviations and divide by sample size (n) – 1, to get an overall estimate of the *variance*. More generally, the variance is given by the following equation:

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

In the student score example, variance = 2987.72/6 = 497.96. The *standard deviation* is the square root of variance, which equals 22.3 in our example. The smaller the variance (and standard deviation), the less the dispersion of the data (the numbers are closer to the mean). If the variance is large, there is more dispersion around the mean (numbers are all over the place!). Relating these statistics to the sources of error that we studied previously, would a high variance or standard deviation indicate a possible problem with reliability or validity?

**Computing measures of central tendency and dispersion using Excel**

You can use Excel to compute several descriptive statistics in one shot. The descriptive statistics include means, standard deviations, median, mode, minimum, maximum and a bunch of other statistics. Here’s how you use “Descriptive Statistics”:

- Click Data > Data Analysis
- Click Descriptive Statistics in the Tool Box
- Type the variable name or cell range in the Input box.
- If you type in cell range, verify *Labels in First Row* box is checked
- Check Summary Statistics.
- If you want the output to be on the same work sheet in a different place, click on output range and specify the cell where you want the output to start.
- If you want the output on a different worksheet, click on new work sheet and define a name.

Example: Descriptive Statistics for variables in the shoe data set.
DSPURCH = Whether Dress Shoe Purchased in Last Two Years
DSFIT = Importance of Dress Shoe Fit

<table>
<thead>
<tr>
<th>DSPURCH</th>
<th>DSFIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.921569</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.026751</td>
</tr>
<tr>
<td>Median</td>
<td>1</td>
</tr>
<tr>
<td>Mode</td>
<td>1</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.270177</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>0.072996</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>8.294405</td>
</tr>
<tr>
<td>Skewness</td>
<td>-3.1831</td>
</tr>
<tr>
<td>Range</td>
<td>1</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>1</td>
</tr>
<tr>
<td>Sum</td>
<td>94</td>
</tr>
<tr>
<td>Count</td>
<td>102</td>
</tr>
</tbody>
</table>

Look at minimum and maximum from the descriptive statistics and see if they are within the range you expected. The mean, median, and standard deviation (or variance) are useful for getting insights into the variable of interest and for answering some research questions. Note, however, that if the data are ordinal, you *cannot* interpret the means.

**Interpret the Descriptive Statistics**

Are the data entered correctly? Can you tell a story with the statistics?

You can also use the function wizard to compute minima, maxima, means and standard deviations.

The function wizard can be activated by an = sign.

Go to the cell where you want the statistic to appear, and type:

=MAX (column range or variable name)
for computing the maximum;
=MIN (column range or variable name) 
for computing the minimum;

=AVERAGE(column range or variable name) 
for computing the mean; and

=STDEV(column range or variable name) 
for computing the standard deviation.

Example:  =AVERAGE(C2:C103) gives the mean of DSFIT as 6.53 across 102 total observations.

=STDEV(C2:C103) gives the standard deviation of DSFIT as 0.61

**Taking means by groups.** Sometimes you may want to take means for particular subgroups, such as gender – i.e., compute the means of DSFIT for males and females separately. To do this you must first sort the data by gender.

Sorting procedure:
- Place the cursor anywhere on the spreadsheet.
- Select “Sort” from the Data menu.
- Find the appropriate sorting variable in the “Sort by” box, and choose either the ascending or descending option in the "Order" box.
- The entire data will be sorted by the variable.

Example. To sort by gender, select “Sort” from the “Data” menu, select the GENDER variable, and the “Smallest to Largest” option. The data will then be sorted with all 1’s (males) coming first in rows 2 to 45 and all 2’s (females) in rows 46 to 103. Now if you want to take averages for males, I refer to rows 2:45 of the sorted data. For example, if I want to take means of DSFIT for males, then I type

=AVERAGE(C2:C45) and get 6.5

How would you take average DSFIT for females?

**Computing Confidence Intervals and Statistical Tests of Means**

Looking at the Descriptive Statistics of the variable DSFIT, shoe fit seems to be an important criterion when purchasing dress shoes (mean =6.53 on a 7-point scale) and its standard deviation (dispersion) is relatively low (s = 0.61)

As a marketing research student, you should be able to go beyond just looking at sample means and standard deviations. Managers are not interested in the sample of 102 consumers (SMU students) who filled out the survey, but they are likely to be interested in the entire population of SMU students. We will use statistical tests to project from our
sample to the entire population. One way is to build a confidence interval around the sample mean. Typically, we use the 95% confidence level.

The 95% confidence level is given by the formula: \[ \bar{x} \pm 1.96 \frac{s}{\sqrt{n}} \]

where \( \frac{s}{\sqrt{n}} \) is called the Standard Error. Excel provides this info in the “Descriptive Statistics” analysis tool. We sometimes call \( 1.96 \frac{s}{\sqrt{n}} \) the "margin of error."

The 95% confidence interval for the importance of shoe fit is therefore:

\[ 6.53 \pm 2 \left( \frac{0.61}{\sqrt{102}} \right) = 6.53 \pm 0.12 = 6.41 \text{ to } 6.65. \]

That is, even though I have used a sample of only 102 students, I can be 95% confident that the true importance of shoe fit to the population is between 6.41 and 6.65. Isn’t that a narrow range? Isn’t that precise? Of course, what has allowed us to be so precise is the low standard error (people are very consistent in their beliefs about the importance of shoe fit). Things may be different for other variables.

I can ask a question about the population in another way. I know that in the sample, fit is considered very important (6.53 out of maximum 7). Does it mean that in the population, it is important? Statistically, I set this question up as a hypothesis test.

Because we can’t really use statistics to prove that a hypothesis is true, we use statistics to show that the opposite is not true. We call that opposite hypothesis the “null” hypothesis. In this case, we return to Question A2 and observe that a neutral response is 4. Numbers over 4 mean that the attribute is important (to a greater or lesser extent); numbers below 4 mean that the attribute is unimportant (again, to a greater or lesser extent). We use 4 as the cutoff point and set up the following null hypothesis:

\( H_0: \) The mean importance for shoe fit (in the population) is less than or equal to four (i.e. shoe fit is unimportant).

More specifically, \( \mu \leq 4 \) (or \( \mu \leq \mu_0 \) where \( \mu_0 \) is the null hypothesis mean, here \( \mu_0 = 4 \))

The Greek letter \( \mu \) generally represents population mean (\( \bar{x} \) is the sample mean). The null hypothesis is generally not what you believe to be true. You challenge the null hypothesis with the alternate hypothesis that you are betting on (the one that you do believe).

\( H_A: \) The mean importance for shoe fit (in the population) is greater than four (i.e. shoe fit is important). More specifically, \( \mu > 4 \).

This is called Test of Single Mean (Comparing the mean of one variable to a fixed value).

I can ask another question about the population. Is the importance of shoe fit greater than the importance of price? The null and alternate hypotheses would be set up as:
$H_0$: The mean importance of shoe fit (in the population) is less than or equal to the mean importance of price. That is,

$$\mu_{fit} \leq \mu_{price}$$

$H_A$: The mean importance of shoe fit (in the population) is greater than the mean importance of price. That is,

$$\mu_{fit} > \mu_{price}$$

This is called *Test of Two Means* (Comparing the means of two variables or the mean of one variable across two groups).

Tests of single mean and two mean can be accomplished by computing the $t$ statistic or using the confidence interval. Students interested in more rigorous and detailed testing procedures are referred to the Text *Marketing Research: The Impact of the Internet*, pages 523-527.

**Computing Sample Means - A Trick!**

Consider Question A4 – how much would you we willing to pay for a dress shoe. It is really an ordinal scale. So, you can analyze frequencies as shown in the figure below.

![Bar chart showing frequency distribution of willingness to pay for dress shoes](chart.png)

You can also compute the approximate mean or average price consumers are willing to pay for dress shoes by taking the mid-point of the range as shown below:
<table>
<thead>
<tr>
<th>Price Range</th>
<th>Mid-point (M)</th>
<th>Frequency (F)</th>
<th>M*F</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1-19.99</td>
<td>$10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$20-39.99</td>
<td>$30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$40-59.99</td>
<td>$50</td>
<td>18</td>
<td>900</td>
</tr>
<tr>
<td>$60-79.99</td>
<td>$70</td>
<td>16</td>
<td>1120</td>
</tr>
<tr>
<td>$80-99.99</td>
<td>$90</td>
<td>24</td>
<td>2160</td>
</tr>
<tr>
<td>$100-119.99</td>
<td>$110</td>
<td>12</td>
<td>1320</td>
</tr>
<tr>
<td>$120-139.99</td>
<td>$130</td>
<td>12</td>
<td>1560</td>
</tr>
<tr>
<td>$140-159.99</td>
<td>$150</td>
<td>10</td>
<td>1500</td>
</tr>
<tr>
<td>$160-179.99</td>
<td>$170</td>
<td>4</td>
<td>680</td>
</tr>
<tr>
<td>$180 or above</td>
<td>$200 (assumed)</td>
<td>6</td>
<td>1200</td>
</tr>
</tbody>
</table>

| TOTAL       | 102          | 10440        |

The mean price consumers will pay for dress shoes $\approx \frac{10440}{102} = \$102.35.
Note that you may have to make an assumption about the last range ($180 or above). Use your best estimate.

In general:
- Take the mid-point of the range (M). Assume some reasonable number for the last range.
- Multiply by the frequency (M*F)
- Add M*F and divide by total sample size.
Data Analysis – Proportions

Consider Question A5 (DSSTORE). Clearly, it is a nominal scale. So, you can analyze only frequencies and proportions. You can draw a bar chart or pie chart to describe the findings. You can also discuss in terms of proportions.

<table>
<thead>
<tr>
<th>Store Type</th>
<th>Proportion (p)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specialty Store</td>
<td>0.39 (39%)</td>
<td>0.29 - 0.49</td>
</tr>
<tr>
<td>Department Store</td>
<td>0.55 (55%)</td>
<td>0.45 – 0.65</td>
</tr>
</tbody>
</table>

- In the sample of 100 students (n), 39% would purchase dress shoes from specialty stores and 55% would purchase from department stores.

We want to infer something about the population. We can construct confidence interval for proportions just like we did when analyzing means.

\[
95\% \text{ Confidence Interval (CI)} = p \pm \text{margin of error} = p \pm \left( 1.96 \sqrt{\frac{p(1-p)}{n}} \right)
\]

The confidence interval is given in the above table.

With 95% confidence, we can say that the proportion of the population who purchase dress shoes from specialty stores is between 0.29 and 0.49 (29%-49%).

After computing confidence intervals, you can perform the Test of Single Proportion and Test of Two Proportions, which are just like the Test of Single Mean and Test of Two Means.
Data Analysis Using Cross Tabulation

In marketing, often we are interested in how two or more variables are related to each other. For example, is household income related to purchase of cellular phones? Is family size related to the amount of money spent on groceries? We may also want to know how strong the relationship is.

When two variables of interest are interval- or ratio-scaled measures, we can measure the correlation (to be discussed later) to infer the strength of their relationship. If the two variables are nominal or ordinal scale, we use cross-tabulation to infer the strength of their relationship. Because of the high frequency with which it is used, cross-tabulation is considered the “bread and butter” of marketing research. But many people, including professional marketing researchers, apply it incorrectly, as we will note shortly.

Simple Cross-Tabulation Analysis

Let us consider a 2x2 cross-tabulation of income by number of cars owned. Both variables have been dichotomized (partitioned into two levels) as follows: Income – Low (<$75,000), High (>=$75,000). Number of cars owned – one (1 or 0), multiple (2 or more). The following table represents a cross-tabulation of 100 consumers.

Raw Table:

<table>
<thead>
<tr>
<th>Household Income</th>
<th>Number of cars owned</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Low: Less than or equal to $75,000</td>
<td>1 or 0</td>
<td>48</td>
<td>6</td>
<td>54</td>
</tr>
<tr>
<td>High: More than $75,000</td>
<td>2 or more</td>
<td>27</td>
<td>19</td>
<td>46</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>75</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

Let us talk through the above table.

48 out of 100 respondents own 1 or 0 cars and have a low household income, 6 out of 100 respondents own 2 or more cars and have a low household income, and so on. What we want to know from this table is whether there is a relationship between income and number of cars owned. Can you make out from the data whether there is a relationship?

It is useful to create a percentage table to investigate the relationship between the two variables. There are two ways to create a percentage table. One is by taking percentages for each row, as in the following table.

Row Percentage Table:

<table>
<thead>
<tr>
<th>Family Income</th>
<th>Number of cars owned</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Low: Less than or equal to $75,000</td>
<td>1 or less</td>
<td>89%</td>
<td>11%</td>
<td>100% (54)</td>
</tr>
<tr>
<td>High: More than $75,000</td>
<td>2 or more</td>
<td>59%</td>
<td>41%</td>
<td>100% (46)</td>
</tr>
</tbody>
</table>
The other method is to take percentages for each column, as in the following table.

**Column Percentage Table:**

<table>
<thead>
<tr>
<th>Family Income</th>
<th>1 or less</th>
<th>2 or more</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low: Less than or equal to $75,000</td>
<td>64%</td>
<td>24%</td>
</tr>
<tr>
<td>High: More than $75,000</td>
<td>36%</td>
<td>76%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100% (75)</strong></td>
<td><strong>100% (25)</strong></td>
</tr>
</tbody>
</table>

One of these percentage tables is right for making causal inferences and the other one is wrong, though often used by many researchers and managers! The row percentage table is correct. In general, percentages should be taken for the causal variable. Here, we are inferring whether having higher income leads to owning multiple cars. The causal variable is income; the effect variable is car ownership. The percentage table is interpreted as follows:

11% of those with low income own multiple cars.
41% of those with high income own multiple cars.

So, the chance of owning multiple cars increases by 30% for a high-income household compared to a low-income household. Put another way, I am 30% more likely to own multiple cars if I have high income than if I have low income. This 30% is the size of the effect. Based on this effect size, it appears reasonable to say that there is a positive relationship between income and car ownership—high income families tend to own multiple cars.

You can use the test of two proportions mentioned earlier to determine whether these proportions are statistically significantly different. One could also use chi-squared statistics for testing the relationship between two variables that are cross-tabulated, though we will not consider these.

**Recoding Redux – Median Split**

For cross-tabulation, you may have to recode the data. The recoding procedure known as Median Split is illustrated for the “Income” variable (Question B3) below using Excel. Note that “Income” is an ordinal scale with 7 income ranges. Analyzing each individual income group is fairly complicated and not very meaningful if there are very few people in each group. So, we use a cutoff and combine everyone above that cutoff as “High-Income” households and everyone below that cutoff as “Low-Income” households. What should the cutoff be? Usually, the cutoff point is the median. From our earlier descriptive statistics, we found that the median income for this sample is in the group 40K-60K, and we observe from the histogram that it is probably closer to $60,000. So, we will call all those who earn below $60,000 (income code 1-3) as “Low Income” and those with household income above 60,000 (codes 4-7) as “High Income.” To achieve this separation, we have to recode the income variable and create a new variable. The recoding is done using IF statement.
• Go to a new column and write a new variable name say INCOMED on top row. In this column you will put the recoded income variable.
• Put the cursor on the second row of that column (first data row)
• Find the column ID corresponding to Income (here U2)
• Type =IF(U2<4,0,1)
• In words what this command does is: If the number in cell U2 is less than 4, give it value 0; otherwise give it value 1.
• The general syntax for this IF command is IF(CONDITION,NUMBER IF TRUE, NUMBER IF FALSE)
• Here CONDITION is U2<4

Cross-tabulation using Excel

Let us go back to the dress shoe data. Suppose we want to find out whether there is a relationship between gender and $s willing to pay for dress shoes (DSPAY). Gender is a nominal variable. DSPAY is ordinal. I am interested in knowing whether females would pay more for dress shoes than males. I divide DSPAY into high payers and low payers using the median as the cutoff. From the descriptive data below, we find that the median is between $80-100. 58% of the respondents are willing to pay less than $100 (up to 5 on the DSPAY scale) – we will call them “low payers” and 42% would pay more than $100 (high payers).

First, I create a new variable (say DSPAY_D) that takes a value 0 (“Low Payers”) if DSPAY less than or equal to 5, and 1 (High Payers”) if DSPAY is greater than 5.

The procedure is similar to how we divided income classes into high income and low income groups.
• Insert a new column
• Top row put label name (DSPAY_D)
• Second row use the following syntax: =IF(cell reference for DSPAY<6,0,1)
• Here, it would be: =IF(K2<6,0,1)
• Copy this statement by dragging to other rows of the column.
• Now you have created a variable DSPAYDUM with 0=Low Payer and 1=High Payer

The next step is to relate gender to low/high payer. The PIVOT TABLE allows you to create two-way tables to do cross-tab analysis.

1. Press "Insert" tab>Click on "PivotTable"
2. Enter the range of the data to be analyzed. Note- The data should be contiguous (in adjacent columns). So, you will need to copy and put the GENDER and DSPAY_D variables side-by-side.
3. Click "OK." A window will appear with the variables you have selected.
4. Drag the field button corresponding to the variable and drop it in the appropriate area to represent row variable and column variable. In our case, the row variable is GENDER and the column variable is DSPAY_D.
5. For counting the number of males and females who are low vs. high payers, drag DSPAY_D again and drop it in the center area.

6. In the "Summarize by" window, “Sum” will be highlighted. Highlight "Count" instead. Then click "OK."

You will get a Table that looks like the one below. (I have typed in the variable names)

<table>
<thead>
<tr>
<th>Count of DSPAY_D</th>
<th>Willingness to Pay for Dress shoes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low payer</td>
</tr>
<tr>
<td>GENDER</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>14</td>
</tr>
<tr>
<td>Female</td>
<td>42</td>
</tr>
<tr>
<td>Grand Total</td>
<td>56</td>
</tr>
</tbody>
</table>

This is the RAW table. What can you tell from this Table?

To make more meaningful interpretations, you should create a Percentage Table by taking percentage along the causal variable (Gender).

To create the percentage table, follow steps 1-6 above, then

6. Select the "Show values by" tab. "Normal" will appear in the window. Do not click OK.

7. Select the down arrow next to "Normal" to display options.

8. Highlight "% of Row" then click OK.

You will get the following percentage table.

<table>
<thead>
<tr>
<th>Count of DSPAY_D</th>
<th>Willing to Pay for Dress shoes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low payer</td>
</tr>
<tr>
<td>GENDER</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>31.8%</td>
</tr>
<tr>
<td>Female</td>
<td>75.0%</td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
</tr>
</tbody>
</table>
Data Analysis Using Correlation

Another common tool for assessing the relationship between two variables is correlation. Correlation does not require binary variables (or a small number of categories), but it should be used with proportions and/or interval variables. In other words, it can be used without dichotomizing, or splitting interval variables.

Remember that we computed the variance, or dispersion, of a single variable $x$ as:

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

Now think of a situation in which we have two variables per respondent, $x_i$ and $y_i$. Returning to our previous example, let’s say that $x_i$ is gender and $y_i$ is $ willingness to pay, DSPAY$ (note we are not using the median split of willingness to pay, DSPAY_D, here). To understand the relationship between these variables, we will determine how deviations from the mean of gender, $x_i - \bar{x}$, relate to deviations from the mean of willingness to pay, $y_i - \bar{y}$. We find this relationship by “substituting” the product of these deviations in the variance calculation above. The result is called covariance, which is computed as:

$$COV(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

If every positive deviation from the mean of gender, $x_i$, is accompanied by a positive deviation from the mean of willingness to pay, $y_i$, then multiplying the two deviations together and averaging them will give us a large positive covariance. On the other hand, if every positive deviation from the mean of gender, $x_i$, is accompanied by a negative deviation from the mean of willingness to pay, $y_i$, then multiplying the deviations together and averaging them will give us a large negative covariance. In our example, the covariance of gender and $ willingness to pay is -0.436. What does this mean?

Next we ask: How large is large? This is difficult to answer from the covariance that we calculated. However, we can create an easy-to-use scale if we divide the covariance by the standard deviations of $x$ and $y$. The resulting number is known as the correlation between $x$ and $y$, and referred to as $r_{xy}$. It is computed as:

$$r_{xy} = \frac{COV(x, y)}{\sqrt{VAR(x)\cdot VAR(y)}}$$

(recall that the square root of the variance is the standard deviation).

The range of the correlation is from -1 to +1. If the two variables have a perfect negative correlation (-1), it means that a larger value of one variable is always accompanied by a correspondingly smaller value of the other. If the variables have a perfect positive correlation (+1) it means that a larger value of one variable is always accompanied by a correspondingly larger value of the other. A correlation of 0 means that there is no discernable relationship between the two.
Computing correlations using Excel

You can use Excel to compute correlations as follows:

- Select "Data" tab > "Data Analysis Tools"
- Click on “Correlations” in the Tool Box
- Type variable name or cell range in the Input box; the ranges of the correlated variables must be side-by-side.
- If you type in the cell range, verify "Labels in First Row" box is checked
- If you want the output to be on the same work sheet in a different place, click on output range and specify the cell where you want the output to start.
- If you want the output on a different worksheet, click on new work sheet and define a name.

Example: Correlation for DSPAY and GENDER.

<table>
<thead>
<tr>
<th></th>
<th>DSPAY</th>
<th>GENDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSPAY</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GENDER</td>
<td>-0.43065</td>
<td>1</td>
</tr>
</tbody>
</table>

The correlation of -0.431 is negative, indicating that male consumers (coded 2 vs. 1 for female), are willing to pay less for dress shoes than females.

Is the correlation significant? There is no general rule of thumb regarding correlations, but those approaching 0.5 or higher in magnitude (either – or +) are usually very significant. You can assess the relationship between two variables statistically by estimating a regression. First (as in cross-tabulation), you must determine which variable is causing the other. In this case, gender is the likely cause of willingness to pay, so gender is the causal variable. We use Excel to estimate the regression as follows:

- Select "Data" tab > "Data Analysis Tools"
- Click on “Regression” in the Tool Box
- Type the variable name or cell range in the Input Y Range box. This is not the causal variable; it is the likely result of that cause.
- Type the variable name or cell range in the Input X Range box. This is the causal variable.
- If you type in the cell range, verify labels in first row box is checked
- If you want the output to be on the same work sheet in a different place, click on output range and specify the cell where you want the output to start.
- If you want the output on a different worksheet, click on new work sheet and define a name.

Example: Regression of GENDER on DSPAY.
Under the “ANOVA” heading is an F-test of statistical significance for the regression model. Under the “Significance F” heading is the probability that there is no correlation between the variables (the null hypothesis of this statistical test). In this example, the probability that there is no correlation between GENDER and DSPAY is 5.77106E-06, a very small probability. Because the probability that there is no correlation is so small, we conclude that with more than 95% confidence that GENDER is related to DSPAY.
Summary

The key take-home points from this module are the following:

- Data analysis is a crucial component of market research that translates survey responses into meaningful findings.
- Steps in data analysis include: (i) editing and coding survey data, (ii) inputting them in the computer in a software-readable format, (iii) doing basic analysis such as frequency distribution and means analysis and generating insights, (iv) testing hypotheses where pertinent, and, if needed, (v) resorting to higher order analyses such as correlation.
- Many data analyses for survey research can be conducted using Microsoft Excel. Other commonly used data analysis software programs include SAS and SPSS.
- Researchers should be guided by two general rules when deciding on the nature of data analysis to be performed:
  1. First, the research objectives should dictate what data analyses to conduct. Take each question and consider the objective that relates to the question. Does the analysis you are considering address the objective(s)?
  2. Second, the type of analysis should also be dictated by the type of scale used for the question (i.e., nominal, ordinal, interval or ratio).
- Many survey questions tend to use nominal or ordinal scales. Frequency analysis is appropriate for these data. If the data are binary (only two responses), or can be made into binary data using a median split, then we can analyze them as proportions, using means, standard deviations, hypothesis tests, etc.
- Frequency distribution can be represented using bar chart or pie chart.
- The old saying – “A picture is worth a thousand words” – holds in market research. When drawing charts, the numbers should preferably represent percentages (not raw frequencies) and the figure should be as self-explanatory as possible – i.e., should embed all important information.
- For interval and ratio scales, as well as proportions, analysis of means and standard deviations/variances is most appropriate.
- Constructing confidence intervals and conducting hypothesis tests are rigorous ways to make inferences from sample data about the population of interest.
- Cross-tabulation analysis is considered the bread and butter of marketing research because it allows us to assess the strength of relationships between two or more variables.
- For making inferences from cross-tabulation, one needs to compute percentages along the causal variable.
- After thoroughly analyzing the data, findings should be presented in a comprehensive and organized form.

While we have discussed data analysis in some depth, it is by no means exhaustive. Interested students can read Chapter 14 and 15 of the text, “Marketing Research: The Impact of the Internet,” to gain further knowledge of various statistical procedures.
Writing a Research Report

After completing your data analysis, research findings should be presented in a comprehensive and organized manner (your semester project report is expected to be presented in a powerpoint document). A research report should generally contain the following:

1. **Executive Summary** This section presents what you would tell your client if s/he did not have much time to read the report. It must clearly and concisely present the heart of the report. The objectives, methodology, findings, conclusions, and recommendations (a paragraph on each) must be presented forcefully, and briefly.

2. **Introduction and Research Objectives** Here you will briefly introduce the topic, background and comment on why the research topic is important. Then, you will state the Research Purpose (the management decision problem), and your specific research objectives/questions.

3. **Secondary Data Research** If you have collected secondary data, report key information obtained from secondary data that is pertinent to your research objectives.

4. **Qualitative Research** If you have conducted a focus group or depth interviews, report some basic information about how the quantitative research was conducted and state the key findings and insights.

5. **Survey Research Design** In this section, you will state your choice of survey method (telephone/personal/mail/email, etc.); attach a copy of the questionnaire in the Appendix.

6. **Sample Characteristics** Note how you picked the sample for survey, report sample size and sample characteristics (description of the demographics).

7. **Survey Research Findings** The data analysis and findings should be organized around the stated objectives and should be supported by meaningful tables and/or charts. Where appropriate, statistical tests should be reported. Include only important numbers, tables, and charts in the text. The less important analysis can be relegated to the appendix. *Just state the findings in the text.*

8. **Conclusions and Recommendations** State the key findings, and recommend specific actions where appropriate. Also indicate any limitations of the method (e.g., sample size, nature of data collection) that you think might have affected your findings.

9. **Appendix** should include:
   - Copy of the questionnaire
   - Coding Sheet
   - Any useful statistical tables not included in the text
   - Statistical tests/computations, if applicable