8

Quantitative Data Analysis

Chapter Summary

Introduction

Quantitative data analysis is a stage of social research designed to make sense of our findings. During data analysis, we assess whether our original theories and hypotheses formulated at the early stages of research process are supported by evidence we collected. Quantitative data analysis looks at findings expressed in numbers and statistical tables and figures. These numerical findings are usually obtained by carrying out surveys of public opinion and structured interviews.

This chapter looks at many quantitative techniques used in the process of data analysis. These techniques depend on the statistical types of variables (nominal, ordinal, interval-ratio), because the types of variables determine what kind of statistical analysis is possible. They also depend on how many variables are involved into the analysis: one, two, or three and more.

The chapter examines the most common statistical techniques, such as descriptive statistics and graphs, several measures of association between the two variables (correlation coefficient, Spearman's rho, eta), basic regression, and tests of statistical significance. In shows the application of techniques on a small student project about the use of gyms and fitness activities. The main goal of the chapter is to give an overview of the statistical techniques that are most useful in the student research projects.

Data analysis stage is the final stage of the research process. It brings the theoretical ideas formulated at the beginning of the study together with the evidence (the data) we collected in the process of research. The main task of data analysis in quantitative research is to assess whether the empirical evidence supports or refutes the theoretical arguments of the project.

However, the biggest mistake in quantitative research is to think that data analysis decisions can wait until after the data have been collected. Although data analysis *is* carried out after that stage, it is essential to be fully aware of what techniques will be used before data collection begins. Therefore our data collecting tools—questionnaire, observation schedule, and coding frame—should be designed with the data analysis in mind for two main reasons:

1. The statistical techniques that can be used depend on how a variable is measured. Inappropriate measurement may make it impossible to conduct certain types of data analysis.
2. The size and nature of the sample also imposes limitations on the kinds of techniques that are suitable for the data set.

Types of Variables

The first of these dimensions—the type of variables—is crucial for statistical analysis because it determines what kind of analysis is possible. The types of variables are also called **the level of**

measurement in statistical terms. There are several types of variables (levels of measurement), and the highest the level of measurement, the more ways of statistical analysis are possible.

The second dimension—the number of variables—also determines what statistical techniques could be applied. The **univariate analysis** includes techniques involving the examination of *only one variable*, the **bivariate analysis** looks at *two variables* at a time, and the **multivariate analysis** examines *three or more variables* at once.

Let's now consider the types of variables. There are three statistical types of variables, or three levels of measurement: **nominal**, **ordinal**, and **interval/ratio**:

- **Nominal:** These are the variables with categories which differ in name only. For example, the variable "religion" contains the name of different religions describing each respondent. The only difference that exists between respondents is in one religion or another. Categories cannot be ordered by rank and arithmetic or mathematical operations are not available (e.g., gender: nominal categories are "male," "female," "transgender").
- **Ordinal:** Variables where the categories can be rank ordered, but the distance or amount of difference between the categories may not be equal (e.g., high enthusiasm, moderate enthusiasm, low enthusiasm). Only mathematic operations of comparison are allowed for these variables; no other operations are allowed.
- **Interval/ratio:** Variables where the values are actual numbers in some actual units of measurement (e.g., age in years, say 25 years old), and the amount of difference between categories is uniform (e.g., the age of 20 is 10 years older than the age of 10; the age of 30 is exactly 10 years older than the age of 20, and there is the same distance between the age of 10 years old and 20 years old as between 20 and 30). For the interval/ratio variables, arithmetic operations are allowed as well as the operations of comparison.
    - The distinction between *interval* and *ratio* variables is that interval variables have an arbitrary zero point: for example, the temperature of 0 degrees Centigrade is arbitrary, because the temperature in a different scale, say Fahrenheit, has its own arbitrary zero point, which corresponds to 32 degrees Centigrade. By contrast, ratio variables have a fixed, non-arbitrary zero point: for example, the weight of –1 kilo is not possible, it starts from 0, or the age of –1 years is not possible, it starts from a fixed non-arbitrary zero point too.
    - Most social science variables (age, years of schooling, income) have a fixed 0 point, so they are ratio. Because of this reason, the two types are often lumped into one level, interval/ratio level of measurement. Interval/ratio variables are the highest form of measurement, because all mathematical operations are now possible with the categories, including the calculation of meaningful ratios (e.g., age of 20 is exactly twice as much as the age of 10).

Univariate Analysis

Univariate analysis is an analysis of one variable at a time. It includes numerical and graphical presentation of the variable.

Frequency Tables

Often, the first step in the analysis is to create *frequency tables* for the variables of interest. These tables show the number and percentage of people who answer a survey question in a particular way. For example, the survey of voting intentions for a sample of 1000 Canadians can show that 34 per cent of respondents, or 340 respondents, supported the Liberals in the last elections, 33 per cent, or 330 respondents, supported the Conservatives, and 33 per cent, or 330 people, supported the NDP. It can also show that 50 per cent of respondents, or 500 people, are females, and another 50 per cent are males.

Frequency tables and other statistical analysis are now created by using statistical software, for example SPSS, SAS, or STATA. The computer programs make data analysis easier, but they require certain competence to use them. Students usually are shown how these programs work in a statistics course.

What happens if an interval/ratio variable such as *age* is put into a frequency table? Because you might have respondents with a variety of ages, a simple frequency table might be too long and too detailed. In this case, it might be useful for the researcher to group the answers in a set of intervals. The values may be combined as long as they don't overlap and are *mutually exclusive (*e.g., age groups of 20–29, 30–39, . . .). The list of categories must also be *exhaustive*, that is, include all values. Combining categories makes the data more manageable and easier to comprehend.

Diagrams

Graphs: For one variable, the most common *graphical presentations* are bar charts, pie charts, and histograms:

- **Bar chart:** This chart graphs categories of a nominal or an ordinal variable. The height of each bar represents a number of people in each category.
- **Pie chart:** This chart is another way to graph categories of a nominal or an ordinal variable. Pie chart shows the size of different categories, but it more clearly brings out the size of each category relative to the total sample.
- **Histogram:** This graph displays interval/ratio variable. Each bar represents the number of people in a particular category. Histogram looks like a bar chart, but the bars touch each other to show that it is measuring an interval/ratio variable.

Measures of Central Tendency

The measures of central tendency describe the most common and average values of a variable, while the measures of dispersion show how spread a variable is.

- **Mode:** The most frequently occurring score, category, or value in variable. Can be used with all levels of measurement, although it is most applicable for nominal data.
- **Median:** The middle score when all scores have been arrayed in order. Can be used with ordinal or interval/ratio data, but not for nominal variables because nominal values cannot be ranked.
- **Mean:** The sum of all scores, divided by the number of scores. Can be used with interval/ratio data. Vulnerable to *outliers* (extreme scores of the distribution on either end), mean gets inflated or decreases when outliers are present in the data.

Measures of Dispersion

Measures of dispersion describe the amount of variation, or spread in a sample. Two groups of students might have the same average score in an exam, but in one group the scores will be more evenly spread out (=have higher dispersion), but in the other they would cluster closer to the mean (=have lower dispersion).

- **Range:** This is the most obvious way to measure dispersion. We simply take the difference between the highest and the lowest scores. Range is appropriate for interval/ratio variables, and is susceptible to the influence of outliers.
- **Standard deviation:** Standard deviation measures the amount of variation around the mean. It is calculated by taking the squared differences between each score and the mean, adding them up, dividing them by the number of cases in the sample, and taking the sure root of the result. Standard deviation is also affected by outliers.

Bivariate Analysis

Bivariate analysis looks at the relationship between the two variables. We can produce basic cross-tabulations of one variable by the other, graphs, or measures of association between two variables.

Contingency Tables

Contingency tables allow simultaneous analysis of two variables to identify patterns of association. They can be used for any variable type but are normally used for nominal or ordinal data because they get too large and inefficient with interval/ratio variables. For example, we can create a cross-tabulation of the Canadian region by voting preferences among three parties (which party the respondent voted). Because the region is the independent variable in the cross-tabulation, we would put them in columns. The party voted would be represented in rows, and the number in the cell of intersection between regions and party would indicate exactly the number of people from that region voting for that party.

Cells in contingency tables usually also contain percentages of the respondents falling in each cell. If the independent variable is in columns, we calculate *column percentages* in the table, and compare the percentages on dependent variable (party voted) by categories in independent variable (regions). We can see the pattern of association: which regions were more likely to vote each party.

Pearson's r

The correlation coefficient ***r*** is a measure of association to examine the relationship between two interval/ratio variables, for example age and income. Correlation coefficient is the most popular measure of relationship between two variables in social science, and has the following characteristics:

- Values of the correlation coefficient range from 0 (indicates no relationship)
  to +1 (indicates perfect positive relationship)
  or –1 (indicates perfect negative relationship)
- A correlation coefficient close to **1** or to **–1** (such as **0.8**) indicates a strong linear relationship between variables, a correlation coefficient close to **0** indicates weak correlation (e.g., **-0.2**)
- The relationship between variables can also be positive or negative. *Positive relationship* indicates that as one variable increases (e.g., age goes up), the other variable increases as well (income goes up). *Negative relationship* indicates a decrease in the second variables as the first variable goes up (as age goes up, income decreases).
- The relationship between two variables should be *linear*: is the correlation coefficient is strong, the data points on a plot should be clustering a long a straight line or a tube.

Scatter Diagrams

These are the graphic representations of bivariate relationship. The independent variable is plotted along the *X*-axis, while the dependent one on *Y*-axis.

If the value of the independent variable *X perfectly predicts* the dependent variable *Y*, the points on the scatter plot align exactly along the straight line. If the points on the scatterplot cluster in a tube-like fashion along the imaginary straight line, the graph shows *strong linear relationship*. If the points on the scatterplot have no particular shape, do not cluster along a tube, but rather look like a shapeless cloud, the correlation coefficient is predicted *as very low* or close to ***0***.

Other Measures of Association in Bivariate Analysis

Apart from the correlation coefficient Pearson's r, there are other measures of association that describe the relationship between two variables. **Measure of association** is a one-number summary which describes the *relationship* between the two variables. Measures of association describe whether the association exists, how strong it is, and what is its direction. There are different measures of association: the *correlation coefficient r*, *Kendall's tau-b*, *Cramer's V*, *Spearman's rho*, and others.

The use of the particular measure of association depends on the level of measurement of each variable involved. The measures of association for lower levels of measurements, such as for nominal variables, can be used for variables measured at ordinal or interval/ratio levels. However, the reverse situation is not correct: measures for interval/ratio variables, such as Pearson's *r*, can be used only with interval-ratio variables, and not with ordinal or nominal ones. Pearson's *r* is appropriate for the situations when both variables are measured at interval-ratio level of measurement.

Other measures of association include:

- **Kendall's tau-b:** This measure of association checks whether the pairs of cases on both variables change in the same direction, or whether the ordering of cases goes together. It is used for ordinal variables, or with one ordinal and one interval/ratio variable. Like Pearson's r, values range from 0 to ±1.
- **Spearman's rho:** This measure of association is used for ordinal variables, and it checks whether the ranks (orderings of cases) of two variables are correlated:
  - Like Pearson's *r*, Spearman's rho values range from 0 to ±1
  - Establishes a rank correlation coefficient (e.g., if this variable is in a particular rank position then we can predict the rank for the other variable)
  - Pearson's *r* cannot provide this prediction
- **Cramér's *V*:** – Used for two nominal variables
  - Values range from 0 to 1. It is always positive because nominal categories cannot be rank ordered.
  - Usually reported with a contingency table and a chi-square test
- **Eta**
  - Used with an interval/ratio variable and a nominal variable. For example, we can find out whether the salary of women in a company is different from the salary of men. We will calculate eta to see if there is a relationship between gender and salary. The nominal variable here, gender, is the independent variable. The interval/ratio variable, income is the dependent one. We can compare means of salaries for males and females.
  - Eta values range from 0 to 1, they are always positive.

Amount of Explained Variance

There is another possible interpretation for the above mentioned measures of association. By squaring eta, Kendall's tau-b, Spearman's rho, and Pearson's *r* the researcher can see how much variance in the dependent variable—in percentages—can be explained by variance of independent variable. This provides a quick measure of influence of one variable on another. For example, if the correlation coefficient between years of education and salary is **0.7**, squaring this value we can say that **0.49**, or **49 per cent** of variation in salary is explained by years of education.

Statistical Significance and Inferential Statistics

Can the results of research on a sample be used to estimate a characteristic of the whole population? If the sample is chosen using a probability method, this can be done. However, even then, there is a possibility that we will have a *sampling error*, when the sample results will be a bit off mark from the population values.

Tests of statistical significance provide the researcher with the estimation of the error, or the risk we are taking to estimate the population characteristic from the sample. Usually we want the results from the sample to reflect the population characteristics with a 95 per cent degree of confidence, or with the 5 per cent probability of error. The tests of significance allow us to estimate the level of error.

The estimation of the risk is done by testing the **null hypothesis** of no difference. For example, we can test the null hypothesis that there is no association between the two variables, or that the means of the two populations are not different. An acceptable level of significance (or level of error) is established and the null hypothesis tested. The level of significance must be **.05** or lower (≤ **.05**) so that the error does not exceed **5 per cent**, the usual maximum acceptable level in social research. If the null is supported, it is not rejected and the original statement remains: there is no relationship between the variables, there is no difference in means, etc. But if the null is false and it is rejected, the level of error for making this conclusion is stated in the statistical significance number, usually denoted by the letter **p** (the probability of error). We want the level of significance **p** to be as low as possible, but at least at **0.05** or lower, corresponding to the **5 per cent** level of error.

When we test the null hypothesis, there is always a chance that we accept the false null hypothesis or we reject the true one. Hence the null hypothesis is never 100 per cent supported or rejected; there is always an error of prediction. There are two types of errors in this process: Type I error of rejecting a true null hypothesis and Type II error of not rejecting a false null hypothesis. A Type I error means that the results of a sample are by chance and that the researcher was mistaken in concluding that there was an association in the population. The two types of errors act in converse relationship to each other and cannot be minimized at the same time: if one is low the other is high. Researchers usually choose to minimize the Type I error over the Type II. This makes it less likely that the null hypothesis will be rejected by error.

If the null hypothesis is rejected, we obtain the *p-value* which states how much error we made in rejecting the hypothesis when it was in fact true. In this case, the results are called statistically significant. However, if the hypothesis testing shows significant results, we cannot automatically assume that the results are important. Significant results do not mean that the results are substantively important. Statistical significance only speaks to the results not occurring by chance alone, it does not speak to the importance of the results.

Correlation and Statistical Significance

The researcher can test the statistical significance of the correlation, to check whether the correlation exists not only in the sample, but in the population as well. The significance of a *Pearson's r* is determined by (1) the size of the computed coefficient and (2) the sample size. If the calculated correlation coefficient is small, the results are less likely to be significant. By contrast, the larger the sample size, the more significant the results can be, regardless of whether there is a true relationship in the population or not.

Correlation and statistical significance must be considered together. A correlation of **.25** may not seem fairly small, but if it is significant (say, *p*-value is equal to **.01**), this means that the correlation in the population is present and it is unlikely that the correlation would occur by chance.

The Chi-Square Test

This test is applied to contingency tables and measures the likelihood that a relationship between the two variables exists in the population. It is calculated by comparing the observed frequency in each cell with frequency expected by chance (which would mean that there was no relationship between the two variables). The chi-square is test is good for nominal or ordinal variables. For example, we can check whether respondents in different regions in Canada (regions=nominal) vote for different political parties (parties=nominal variable). If the *p*-value for the chi-square test is small (***p*≤ .05**), we can conclude that there is a statistically significant relationship between variables, for example that different regions have different political preferences.

It is important to note that chi-square is always more likely to be more significant if the sample size is large.

Comparing Means and Statistical Significance

The analysis of variance test (F statistic) works to test whether there is a significant difference between the means of several groups. For example, we can test whether there is a statistically significant difference between the mean salary of males and females in a company. The F-test works by comparing the variance *within* each group (males and females) and *between* them. The F statistic explains the amount of *explained variance* (variation between groups) in relation to *error variance* (variation explained by the group one is in). If the F-statistics is significant, we conclude that the difference between groups is significant: for example, the finding about differences in salaries between gender groups.

Multivariate Analysis

Multivariate analysis examines the relationship between *three or more* variables.

Is the Relationship Spurious?

Although the prior logical position of the demographic variables (age, gender, education, etc.) makes it more likely to consider them to be causes, sometimes the causal relationship we suspect might be spurious. **Spuriousness** exists if two variables are correlated, but not because one is the cause of the other, but because there is the third variable *Z* that explains both variable *X* and variable *Y*. One example is the relationship between income and voting for a conservative candidate. We may observe that people who own more are more likely to vote conservative. Does this mean that being rich causes people to be conservative? Not necessarily so. It is more likely that there is a third variable, age, which is generally better at explaining people's ownership and their conservativism: older people tend to be both richer and more conservative.

Is There an Intervening Variable?

This variable suggests that the relationship between the two original variables is indirect. The test for intervening variables is quite straightforward. A control for the outside variable (that is thought to be intervening) is put in place and the relationship between the original variables is measured. If the relationship between the original variables disappears, then the outside variable is considered an intervening variable.

$$X \rightarrow Y \text{ (intervening variable?)} \rightarrow Z$$

e.g., Education → Income → Happiness

Control for income (e.g., look only at people with high income) and see if there is still a positive relationship between education and happiness. If the correlation between education and happiness disappears, then income is an intervening variable.

Is There an Interaction?

An interaction exists if the effect of one independent variable on the dependent one varies at different levels of a second independent variable. For example, is it only the lack of exercise that leads to weight gain? Is so, the weight gain effects would be similar at different levels of exercise. If the presence of exercise is not the only influence on weight gain, then the influence of exercise on weight would be different for different levels of this third variable, for example, stress. If so, people who exercise more but have a higher level of stress but achieve a lower gain loss than the people who exercise and have less stress. In this case, exercise affects the gain loss in interaction with stress.

Multiple Linear Regression

Multiple linear regression can determine (1) how much of the variation in the dependent variable is explained (predicted) by the independent variables, and (2) which, if any, of the independent variables is a significant predictor of the dependent variable. Regression quantifies the influence of one variable on the other. For example, it can tell you that having one extra year of education is likely to produce $XXX amount of increase on your income. Multiple regression quantifies the common influence of all independent variables (age, education, experience) on one's income. This common influence is reflected by R^2 statistic: the percent of variance explained by the model. The higher the per cent of variance explained (out of 100 per cent), the better the independent variables explain the dependent variable.

The other important parts of regression output in SPSS are the following:

- ANOVA table indicates whether the proposed model with all independent variables is a good model as a whole in explaining the dependent one. The significance level of the F-statistic indicates that if the p-value is low (≤ **.05**), the model is a good model overall.
- The "Coefficients" table gives us three important indications.
  1. It shows the *net effect* of one unit increase in the independent variable on change in the dependent variable, keeping the other variables constant. This influence is reflected in the column "unstandardized coefficients."
  2. The table shows the *relative importance* of each independent variable for predicting the dependent variable. This effect is reflected in the "standardized coefficients" column. The higher the standardized coefficient (in absolute terms), the more important that independent variable is for predicting the outcome.
  3. Finally, the coefficients table also indicates whether a particular independent variable is significant for the model overall. This is indicated in "Significance" column of the coefficient's table. If the *p*-value for the variable is small, it means that this variable is a significant predictor of the dependent variable.

Learning Objectives

In this chapter, you should learn to do the following:

- Understand the main goal of quantitative analysis, which is producing numerical and graphical data to check the theory put forward in a research project
- Differentiate among the various types of variables (levels of measurement): nominal, ordinal, interval/ratio
- Be able to produce (and/or read and comment on) the output for univariate analysis: frequency tables, graphs, measures of central tendency, and measures of dispersion
- Interpret the basic bivariate analysis results: cross tabulations, graphs, and correlation coefficients. Understand that the purpose of measures of association is to assess whether

there is a relationship between the two variables, evaluate its direction and strength, and appreciate that measures of association also depend on levels of measurement of variables

- Know that the basic purpose of hypothesis testing is to assess whether the results from the sample can be generalized to the population: whether the mean, the correlation or the chi-square results are due to a chance alone or due to some deeper patterns or relationship between the variables
- Understand the concept of statistical significance as a test which allows the researcher to estimate the level of error while generalizing the results from the sample to the population
- Appreciate that association does not mean causality, and that the described statistical analysis cannot establish causality by itself. It can establish only that there is a relationship, an association
- Be able to name the basic principle of multivariate analysis and understand the goals of multivariate regression, which are to assess the impact of several independent variables on the outcome

Media Resources

Create Pie Charts in Excel https://www.youtube.com/watch?v=8g_0W9Xh8U
How to . . . Plot a Simple Scattergram in Excel 2010
https://www.youtube.com/watch?v=g471UfimO3M

- What is the difference between one chart style and another?
- How does charting help the researcher to analyze research data before reporting findings?
- How does the selection of one charting style over another make a difference in reporting findings?

Frequency Distribution Tables. Statistics Canada.
http://www.statcan.gc.ca/edu/power-pouvoir/ch8/5214814-eng.htm

- What is the difference between a frequency distribution table and cumulative frequency table?
- When is each used?
- How is a frequency table used for standard deviation calculations?

Exercises. Statistics Canada.
http://www.statcan.gc.ca/edu/power-pouvoir/ch11/exer/5214866-eng.htm

- What are the difficulties with relying on the mean?
- How does the median help to overcome the difficulties that stem from relying on the mean?
- What is the value of using the mean, median and mode together?

Variance and Standard Deviation. Statistics Canada.
http://www.statcan.gc.ca/edu/power-pouvoir/ch12/5214891-eng.htm

- Why is standard deviation the most commonly used measure of spread?
- How can the researcher accommodate a single extreme score?
- What are limitations of using standard deviation?

Correlations
http://www.youtube.com/watch?v=1j6JNzqHHTI&feature=related

- What is the difference between a positive correlation and a negative correlation?
- When is correlation most useful in data analysis?
- How are correlations significant to data analysis?

Tutorial: Introduction to SPSS
https://www.youtube.com/watch?v=SL2bZXfWQls