Social Statistics in Action

Chapter Summaries

Chapter 1: Learning to Think Statistically

This chapter introduced you to the origins of **social statistics** as a tool initially conceived to generate information about the people governed by nation states. Today in Canada this responsibility falls largely to Statistics Canada who use statistical analysis to shape evidence-based policy creation and social scientists who use statistical analysis to describe and understand social inequalities in Canada. Statistical analysis relies on **quantification** or the transformation of information into numerical variables for the purposes of recognizing patterns in the data about the social world. You learned about the building blocks of quantification, **variables**, in this chapter as well.

Variables describe the changes and variety of characteristics of the social world. **Independent variables** represent outcomes and the **dependent variables** represent the causes of the outcomes. A variable's **attributes** represent the variable's range of change or variation in its value. When a variable's attributes describe a quantity or an amount they are known as **count or continuous variables** and the quantity or amount becomes the value for the variable. When the attributes describe groups or people the variable is a **categorical variable** and the value becomes whatever numerical quality the researcher assigns to the categories of the variable. The variable's attributes also determine its **level of measurement**, a concept used to determine what types of statistical analysis may be done with the variable.

Finally, you learned that variables are used in two branches of statistical analysis. The first branch, **descriptive statistics**, are statistical techniques for summarizing information whereas the other branch, inferential statistics, are statistical techniques that use samples collected from a population in order to make estimates or inferences about that population.

Chapter 2: Summarizing Data Using Numbers and Graphics

In this chapter you learned the necessary techniques needed to develop **frequency distributions**, which is a tally or count of the frequency of some attribute(s) of interest of categorical variables in a group of people. **Frequency distribution tables** are a common way of displaying this data and usually take the form of a table displaying a calculation representing the counts of cases with some

attribute of interest. You learned about the various statistical techniques for developing this data for display in these tables including the formulas for calculating **percentages** and **proportions**, calculations demonstrating the fraction of cases of a variable that possess some attribute(s) of interest. You also learned how to calculate **cumulative percentages**. This calculation represents the percentage of cases with a specific attribute or one ranked below it of a lower value, which means cumulative percentages are not reliable for variables at the nominal-level of measurement since variables at this level do not possess an intrinsic order. This chapter provided a review of the formulas for calculating how common or how rare some attribute of interest is, specifically the formulas for calculating **ratios** and **rates**. Calculating ratios allows researchers to compare the frequencies of two different attributes of the same variable while calculating rates allows researchers to compare the number of times some attribute occurs to the number of times it might occur. In this chapter you also learned how to work with two categorical variables at a time using **cross-tabulations**, which demonstrates how the distribution of one variable is conditional on membership in a group as defined by another variable. Finally, in this chapter you were introduced to several different forms of graphs which researchers typically use to visualize the data of frequency distributions. **Pie graphs** use the area of circles to display the percentage of cases with some attribute of interest whereas **bar graphs** use the length of horizontal bars or the height of vertical bars to display this percentage.

Chapter 3: Describing the Centre and Dispersion of a Distribution: Focus on Categorical Variables

In this chapter you learned how to calculate the centre of a variable and its dispersion or distribution around the centre with a focus on the affordability of housing in Canada. The chapter described the tools necessary to undertake these calculations starting with the mode, or the attribute of a variable that occurs the most in the data, and the median, or the middle point of a variable when every attribute of the variable is placed in order from the lowest to highest. This chapter also identified the tools necessary to understand how the cases of a variable are spread out around the centre of its distribution, a concept known as dispersion. Percentiles are one tool used to measure dispersion and are calculated by dividing the cases of a variable into an ordered list based on their attributes and then reporting the cut-off points. The cut-off points represent the percentiles, which are the percentage of cases on or below some attribute or value of the variable. Two other measures of dispersion you learned about in this chapter were the range and the interquartile range. The range is the difference between the lowest attribute of a variable and the highest whereas the interquartile range provides more detail about the dispersion between these minimums and maximums as it is a calculation of the distance between the twenty-fifth and seventy-fifth percentile. This chapter concludes by describing how to construct and interpret box plots, the visual depiction of a variable's median, range and interquartile range. The top and bottom of the plot represent the interquartile range, a line running through the box represents the median, and the whiskers of the plot represent the range of the variable's dataset. Finally, in this chapter you learned that box plots are useful for depicting outliers, or extreme values in the data, and how box plots are also useful for visually

depicting data that allows researchers to show how the dispersion of a variable is different in different groups.

Chapter 4: Describing the Centre, Dispersion, and Shape of a Distribution: Focus on Ratio-Level Variables

This chapter takes you through the steps necessary to calculate the centre and the dispersion for variables at the ratio-level of measurement and it did so by centering youth wage rates as its theme. These steps help you calculate the mean and the standard deviation. You were no doubt familiar with calculating the mean or the average from past educational experiences, but this chapter provided you with the more formal statistical notation describing the mean, or the total value of all the cases added up divided by the total number of cases. In this chapter you learned that the standard deviation requires the mean for its calculation, as the standard deviation represents a measure of a ratio-level variable's dispersion or spread around its mean. This statistic represents the standardized, average deviation from the mean. This chapter illustrates that by using both the mean and standard deviation it is possible to calculate the z-score for a case, which is a standardized score that shows how far an individual case is from the mean of a variable, a good tool for comparing the relative position of cases. This chapter presented material that also teaches you that those cases with unusually high or low values are called outliers and these outliers have a strong influence on the calculation of the mean, standard deviation and hence the variance. You were taught through this chapter that it is important to graph the distribution of the variable before performing any statistical analysis on it to identify any outliers. A common way of graphing ratio-level variables is with histograms, which is a graph that shows the number or percentage of cases that have values within equal-sized class intervals or bins. You learned that it does so through the relative height of a bar, though unlike the bar graph there are no gaps between the bars in a histogram. Finally, you learned that a common graph used as a point of reference for describing the shape and area of a variable's distribution is a graph depicting a normal distribution.

Chapter 5: Probability, Sampling, and Weighting

This chapter was the first one from this textbook to introduce you to inferential statistics, which are statistics that use randomly-selected samples from a population to make claims about that population. It explains some of the theoretical ideas necessary to generalize from samples to populations. One of those ideas you learned about was probability, which means the chance between 0 and 1 that an outcome or event will occur, and joint probably, or the chance that some outcome of interest will occur twice in a row. These probabilities are formally referred to as theoretical probabilities, which as you learned in the beginning of this chapter are calculated by dividing the number of outcomes by the total number of possible outcomes. You learned that observed probabilities are probabilities based on empirical trials that determine how often a specific outcome occurs. This calculation of probability based on empirical trials rests on the law of large

numbers, which means that the observed probability will look more and more like the theoretical probability as the number of empirical trials undertaken increases. This chapter explained to you how this is a crucial property to remember when thinking about the link between probabilities, frequency distributions and the normal distribution so crucial to sampling, the heart of inferential statistics. You learned in this chapter that sampling refers to the process whereby some population is surveyed by collecting either probability or non-probability samples of a specific number of people from the population to make inferences about that population through the data collected in the samples. Non-probability samples do not rely on the principle of random selection. Instead you learned that they rely on convenience samples, which are samples of easily-collected cases, and/or snowball samples, which are samples of other people recommended by a previous participant. Probability samples, of which there are a few different types, require the principle of random selection to ensure each case in the sample has a known, non-zero chance of being randomly selected.

Chapter 6: Making Population Estimates: Sampling Distributions, Standard Errors, and Confidence Intervals

With a focus on the gendered division of labour, this chapter introduced you to the tools and techniques necessary to make **estimations** about a population using a randomly-selected sample from the population. Whenever selecting a sample from a population, the question facing researchers is, how close does this sample's statistics match the population's parameters? One tool for analyzing this relationship was the **sampling distribution**, which is the distribution of a statistic for every equal-sized sample selected from a population. This chapter explained that researchers usually only draw one sample from a population to explore this relationship and that this exploration relies on the principles of the **sampling distribution** and the **central limit theorem**. The sampling distribution possesses certain properties, including that it is approximately normal in shape and is centred on the population mean of a statistic. In this chapter you learned that these properties are possible because of the **central limit theorem**, a law of large numbers that allows for these two characteristics to hold. Since the sampling distribution is approximately normal in shape it is possible to use the properties of the normal curve to analyze the distribution of sample means around the population mean. You were taught in this chapter that the **standard error** is the standard deviation for the sampling distribution and it is a measure of how much variation exists in a sample which, in turn, is a measure of variation in the population. This chapter also introduced you to the **95% confidence interval**, a tool that allows researchers to draw a sample and calculate a range of values the population mean is estimated to fall within based on the sample's statistics. These intervals are visually depicted using error-bar graphs with the middle dot representing the sample mean and the whiskers representing the confidence interval. With these depictions you can compare the population means for two or more different groups within the same population.

Chapter 7: Assessing Relationships by Comparing Group Means: T-Tests

This chapter introduced you to techniques necessary for using sample data to assess whether there is a relationship between two variables in a population, done so in the context of mental health. The primary way to assess this relationship is through **hypothesis testing**, which is a statement about an expected relationship between two or more variables in a population using a randomly-selected sample from that population. In this chapter you were taught that these types of relationships may be **directional or non-directional**, which determines whether a not a relationship between two or more variables moves in a certain direction. You were introduced to this idea by the question of whether income has a relationship to a person's mental health. To answer this question, you learned how researchers focus on the potential relationship's strength or its reliability. To assess its strength using **Cohen's d**, researchers compare the means of each group to see if membership in one group is related to a higher average score as measured with some ratio-level variable. The bigger the difference, the stronger the relationship. The reliability of the relationship refers to how confident researchers are that a relationship found in a sample exists in the population from which the sample was taken, and they measure reliability using **tests of statistical significance**. These tests estimate the chances of selecting a sample from the population with a relationship between two or more variables that does not exist in the population and they rely on a **research hypothesis** to make such estimates. Significance testing is premised on disproving the hypothesis that no relationship exists between the variables in the population. You learned in this chapter that researchers use a **t-test of independent means** along with a **t-distribution** to determine if the difference in means is large enough to say a sample comes from a population with two different means.

Chapter 8: Assessing Relationships by Comparing Group Means: ANOVA Tests

In this chapter, you learned about the **ANOVA, or analysis of variance test**, which is a test used to assess the reliability of a relationship between an independent, categorical variable with more than two attributes and a ratio-level dependent variable. It does so by using sample data to determine if two or more group means are different from each other in the population through the measurement of within group variation and between group variation. If there is more variation between the groups than within, then group membership is likely related to people's values on the dependent variable. In this chapter, you learned that researchers rely on calculating the variation within groups and the variation between groups using the sum of squares formula for each part of the ANOVA test. Using the data from running the numbers through these formulas, researchers use the results to calculate an F-statistic, which is a measure that shows whether the between group variation is larger or smaller than the within group variation after accounting for the sample size and number of groups. When paired with the **F-distributio**n, the F-statistic allows researchers to determine the probability of selecting a sample with the observed F-statistic ratio from a population in which the group means are equal. Researchers know that ANOVA tests only indicate that at least one of the group means is

different, not which one it is. As explained in this chapter, this is the reason why some researchers rely on **post-hoc tests**, which are tests researchers carry out, after a statistically significant relationship has been established, that provide more specific information about the relationship. Graphing the mean differences using error-bar graphs helps illustrate the results of ANOVA tests.

Chapter 9: Assessing Relationships Between Categorical Variables

This chapter introduced you to the knowledge and techniques necessary to assess relationships between categorical variables. The first way to assess such a relationship is through the **proportionate reduction in error measures (PRE)**. These are measures of the magnitude of a relationship between two categorical variables. The second way to assess such a relationship is through the **chi-square test of independence**, which as you learned in this chapter, assesses the reliability of a relationship between two categorical variables using the chi-square distribution. Two primary PRE-measures introduced in this chapter were **lambda (λ)** and **Gamma (γ)**. **Lambda** is used to measure the magnitude of an association between two nominal-level variables or between a nominal-level and an ordinal-level variable, while **Gamma** measures the size and the direction of the association between two ordinal-level variables. Despite their differences, this chapter clarified the importance of remembering how these two **PRE-measures** rely on measuring the distribution of the dependent variable as it changes depending on its group membership. Unlike **lambda**, **Gamma** accounts for the order or ranking of the attributes and provides information about the direction of the relationship. **Lambda**, instead, makes use of the mode or the most frequent attribute of the variable in its measurement of the magnitude of a relationship. In this chapter you also learned how researchers assess the reliability of a relationship between two categorical variables using a **chi-square test of independence**. Unlike **parametric tests**, which rely on means and variances, the chi-square test of independence is a **non-parametric test** as it relies on comparing the frequencies that are observed in the sample with the frequencies that are expected if the null hypothesis is true. The chi-square statistic you learned how to calculate in this chapter is the result of this comparison and is used to determine if a relationship between variables is statistically significant.

Chapter 10: Assessing Relationships Between Ratio-Level Variables

This chapter introduced you to the strategies researchers use to measure the magnitude and reliability of relationships between two ratio-level variables. You learned the three primary ways researchers characterize a relationship between two ratio-level variables. The most basic way is as either a **linear, non-linear,** or **curvilinear relationship**, which illustrates the pattern of the relationship between the variables. Researchers also describe such a relationship as either **homoscedastic** or **heteroscedastic**. A **homoscedastic relationship** is one in which the spread of the dependent variable is consistent across all values on the independent variable. With a

heteroscedastic relationship the spread of the dependent variable is different across the values on the independent variable. Finally, you learned in this chapter how relationships between two variables are described as either **monotonic** or **non-monotonic**. The former means that an increase or decrease in one variable is related to an increase or decrease in the other variable, whereas a **non-monotonic relationship** means the direction of the relationship between the two variables is not consistent across the values of a variable. The two statistical techniques introduced to you in this chapter, the **Pearson's correlation coefficient** and the **Spearman's rank-order correlation coefficient**, should only be used with relatively **homoscedastic** and **linear relationships**. As you learned in this chapter, the **Pearson's correlation coefficient** relies on the idea of **covariance**, as it is a measure of association designed to assess the magnitude and the direction of a relationship between two ratio-level variables by calculating the sum of products and the sum of squares for each variable. You also learned in this chapter how the **Spearman's rank-order correlation coefficient** is calculated using the rank of each case within a variable. The reliability of both correlation coefficients is assessed using the calculation and interpretation of the t-statistic and the t-distribution.

Chapter 11: Introduction to Linear Regression

This chapter introduced you to another means of assessing the relationship between two ratio-level variables: **linear regression**. **Linear regression** relies on identifying a general pattern of a relationship, which allows researchers to use one, ratio-level independent variable to predict the value on a ratio-level, dependent variable. Linear regression depicts the straight-line pattern that best fits the relationship observed in the data between two variables. The line considered the best fit to describe the relationship is the one with the lowest value after totaling the squared distances of the cases from a line. Finding this numerical best fit is accomplished through the **ordinary least squares method**, which is used to find the line with the smallest value after totalling the squared distances between every case and the line. This chapter describes how the line that best fits the pattern of a relationship between variables observed in the data is referred to as the **regression line**. This line relies on minimizing the distances between the cases and the regression line; any **influential cases**, which are those cases that substantially affect the location or direction of the regression line, must be addressed. This chapter introduced you to the idea that the regression line can be calculated using the **slope coefficient** and the **constant coefficient**, which when considered together, comprise the formula $\boldsymbol{y = a + bx}$. The **slope coefficient**, $\boldsymbol{b}$, shows that the size of the increase in the independent variable is associated with an increase or decrease in the dependent variable and shows how big that increase or decrease is. The **constant coefficient**, $\boldsymbol{a}$, sometimes called the intercept, shows the value of the dependent variable when the independent variable is equal to 0. You learned in this chapter how $\mathbf{R^2}$ is used to assess how well a regression line fits the data and how researchers test the significance of sample data, demonstrating a relationship between an independent and dependent variable, using statistical significance tests.

Chapter 12: Linear Regression with Multiple Independent Variables

This chapter introduced you to **multiple linear regression**—linear regression with more than one independent variable to make predictions about a dependent variable. You learned how multiple regression predicts the relationship between each independent variable and the dependent one by controlling for the other independent variables. This allows for much stronger predictions of the dependent variable, as researchers can account for multiple independent variables. Although there are multiple independent variables, you learned that the interpretation of the regression coefficients remains the same: there is a single constant coefficient that predicts the value of the dependent variable when all the independent variables are zero. However, each independent variable has a slope coefficient, capturing the change in the dependent variable associated with a one-unit increase in the independent variable, after controlling for all the other independent variables in the regression. In this chapter you learned how multiple regression is based on **model specification**, which refers to the decisions made by researchers about which independent variables to include in the regression analysis. Researchers strive to include in their analysis the independent variables that are the best predictors of the dependent variable and usually base their choices on prior research and theory. They seek to avoid including variables that lead to **spurious relationships**, which are relationships between two variables that disappear once additional variables are controlled for. On the other hand, you learned how researchers also look to avoid **omitted variable bias**, which means conducting a regression analysis while leaving out an important independent variable that might affect the dependent variable. Finally, this chapter introduced you to the way in which categorical variables with more than two attributes can be converted into **dummy** or **dichotomous variables**, which use a '1' or a '0' to denote whether a variable has a characteristic.

Chapter 13: Building Linear Regression Models

This chapter introduced you to **nested regression** and how to interpret it. It explains how to analyze the ways in which ascribed (characteristics an individual is born with) and achieved status (characteristics earned by an individual) as represented by independent variables, affect a dependent variable (an individual's weekly wages) using **nested regressions**. You learned how researchers build nested regressions by generating a series of linear regressions using the same dependent variable and adding more and more independent variables without removing any of the previous independent variables. Researchers evaluate the change in each partial slope coefficient for every independent variable as additional variables or groups of variables are added to the linear regressions. This chapter also introduced you to the strategies for selecting **parsimonious regression models**, or models that efficiently predict the values on the dependent variable while avoiding any unnecessary independent variables. It introduced you to the strategies researchers employ when deciding which independent variables to include in their regression models, including the **adjusted R^2** calculation. **Adjusted R^2** allows researchers to determine whether an added independent variable is a good or bad predictor of the dependent variable, increasing when it's good,

decreasing or staying the same when it's bad. The problem of **collinearity** exists when two independent variables are strongly related, making it difficult to assess how each of the variables is related to the dependent one. To remedy the problem of **collinearity**, researchers rely on **tolerance** and **variance inflation factors (VIF)** to ascertain, in the case of **tolerance**, how much of the variation within each independent variable is not predicted by the other independent variables in the regression. The **VIF** measures the inflation of a slope coefficient's variance due to a correlation with other independent variables in the regression.

Chapter 14: Manipulating Independent Variables in Linear Regression

In this chapter you learned how and why researchers manipulate independent variables in regression models. You learned how researchers utilize **interaction variables,** how **quadratic variables** allow researchers to model **curvilinear relationships**, and how to **transform** skewed variables. **Interaction variables** make it possible for researchers to model complex relationships between variables by demonstrating how two independent variables jointly affect a dependent variable. Regression may also be used to predict **curvilinear relationships** between an independent and dependent variable, or relationships which are predicted using curved regression lines, using a **quadratic variable**. A **quadratic variable** is one in which every value of a variable is squared to reflect the fact that the researcher believes the relationship between an independent variable and a dependent one is curvilinear. In this chapter you also learned how to **transform** skewed variables to create more accurate regression models. **Transforming** variables means replacing the variable's original values with values produced by running the original values through some sort of mathematical function. This chapter introduced you to three types of transformations: **linear**, **non-linear**, and **logarithmic**. With **linear transformations** the order of cases from smallest to largest and the relative distance between the cases remains the same in the new, transformed variable. **With non-linear transformations**, you learned how the order of the cases remains the same but the relative distance between the cases changes in the transformed variable. Finally, **logarithmic transformations** are necessary for right-skewed variables and they entail multiplying the original value of the variable by some base number raised to an exponent. This moves the cases with high values in the long tail on the right closer together and the cases with low values clustered in the tail on the left farther apart.

Chapter 15: Logistic Regression Basics

This chapter introduced you to the concepts which underlie **logistic regression**, a type of regression used to make predictions about a dichotomous dependent variable. You learned how researchers transform dichotomous dependent variables before using them in regression and how these transformations change the interpretation of the regression's slope coefficients. Transforming dichotomous variables in a logistic regression requires three steps. In the first step, researchers predict the probability that a case will have the value '1' on the dependent variable. The second step

requires a movement from probabilities to **odds**, or a calculation of the number of times something occurs relative to the number of times that it does not occur. The third and final step you learned about for transforming dichotomous dependent variables is a log transformation, a transformation often undertaken using the **natural log** with a common base number known as Euler's constant. When odds are transformed using the **natural log** of the odds they are called **log odds**, which represent the **natural log** of the odds of something occurring. As you learned in this chapter, the transformed variable which results from this three-step process is used as the dependent variable in **logistic regression**. This type of regression is nearly the same as the linear regression you've learned about so far, only **logistic regression** is non-linear, the slope and constant coefficients are calculated differently, and the interpretation of the regression coefficients changes with logistic regression. The interpretation of regression slope coefficients relies on the idea of **odds ratio**, which shows how the odds of something occurring in two different groups compare to each other. Finally, this chapter introduced you to **Nagelkerke's R^2**, which is a measure commonly used by researchers to assess how well a **logistic regression** fits. It measures how much the predictions made by a **logistic regression** using one or more independent variables are better than the predictions made using a null model.