Lecture Notes

# Chapter 18: Descriptive Statistics

## Learning Objectives

* 1. Explain the purpose of descriptive statistics.
	2. Distinguish between inferential and descriptive statistics.
	3. Explain the difference between a frequency distribution and a grouped frequency distribution.
	4. Read and interpret bar graphs, line graphs, and scatter plots.
	5. Calculate the mode, median, and mean.
	6. List the strengths and weaknesses of the mode, median, and mean.
	7. Explain positive skew and negative skew.
	8. Explain the impact of skewness on the measures of central tendency.
	9. Describe and interpret the different measures of variability.
	10. Calculate the range, variance, and standard deviation.
	11. Explain percentile ranks and *z* scores.
	12. Explain how to construct and interpret a contingency table.
	13. Explain the difference between simple and multiple regression.
	14. Explain the difference between the *y*-intercept and the regression coefficient.

## Chapter Summary

This chapter is about descriptive statistics. It presents the multiple ways that researchers can describe, summarize, or make sense of their data.

## Annotated Chapter Outline

1. Introduction
	1. This chapter focuses on **descriptive statistics**, which are the statistics that focus on describing, summarizing, or explaining data. Descriptive statistics are presented graphically or through quantitative methods. Descriptive statistics differ from inferential statistics. **Inferential statistics** are statistics that go beyond the immediate data and infer the characteristics of populations based on samples.
2. Descriptive Statistics: descriptive statistics begin with a **data set** (set of data) and then the researcher uses descriptive statistics to present the data in ways that make it more interpretable.
	1. Figure 18.1 illustrates the major divisions within the field of statistics.
		1. Chapter 18 focuses on descriptive statistics and Chapter 19 on inferential statistics.
	2. Table 18.1 contains a hypothetical data set that is used in Chapters 18 and 19 to illustrate points or as data students can work within class or on their own.
		1. Cases (individuals) are represented in rows and variables are represented as columns.
		2. This is a standard way of organizing data after data collection has ended.
		3. This practice data file is also available as a SPSS file in the student companion website.
	3. Discussion Question: Differentiate between descriptive and inferential statistics
3. **Frequency Distributions:** systematic arrangement of data values in which the data are rank ordered and the frequencies of each unique data value are shown.
	1. Steps in constructing a frequency distribution
		1. List each unique number in ascending or descending order in Column 1. If a particular number occurs more than once, list it only once.
		2. Count the number of times each number listed in Column 1 occurs and put that number in Column 2.
		3. (optional) Construct a third column by converting Column 2 into percentages by dividing each number in Column 2 by the total number of numbers.
		4. See Table 18.2.
	2. If the variable has a wide range of variables, collapse the values into intervals and create a **grouped frequency distribution** (the data values are clustered or grouped into intervals, and the frequencies of each interval are given).
		1. Intervals must be **mutually exclusive:** the property that intervals do not overlap.
		2. Intervals must also be **exhaustive**: the property that a set of intervals covers the complete range of data values.
		3. See Table 18.3.
	3. Discussion Question: How do you construct a frequency distribution and a grouped frequency distribution? Why do we have two types of frequency distributions?
4. Graphic Representations of Data: display data in two dimensions.
	1. Setting up graphs
		1. If graphing data of a single variable: values of variable are on the *x*-axis (abscissa, horizontal dimension) and the frequencies or percentages are represented on the *y*-axis (ordinate, vertical dimension)
		2. If graphing data from two variables, the independent variable is on the *x*-axis and the dependent variable is put on the *y*-axis.
		3. Discussion Question: why is important to always structure graphs as presented in the book?
	2. **Bar Graphs**: A graph that uses vertical bars to represent the data.
		1. Use with categorical variables
		2. See Figure 18.2.
		3. Discussion Question: What information is relayed by Figure 18.2.
	3. **Histograms**: a graphic that shows the frequencies and shape that characterize a quantitative variable.
		1. Used for quantitative variables
		2. Useful because it shows the shape of the distribution of values.
		3. Unlike bar graph, the bars in histograms are next to each other.
		4. See Figure 18.3, a histogram of salary for the sample data set.
		5. Discussion Question: When should you use a histogram instead of a bar graph?
	4. **Line Graphs:** a graph that relies on the drawing of one or more lines.
		1. Can be used for one variable or more
		2. Can be used to illustrate a factorial design’s findings, trends over time
		3. See Figure 18.4 line graph of data set participants’ GPAs.
		4. Discussion Question: Why might a researcher us a line graph rather than a histogram or bar graph to represent her data from her data?
	5. **Scatter Plots:** A graph used to depict the relationship between two quantitative variables.
		1. Used with correlations
		2. Can illustrate the relationship between variables and if the relationship is linear (a straight line) or curvilinear (a curved line)
		3. If the relationship that is present is positive (southwest to northeast direction) or negative (northwest to south east direction)
		4. How strong the relationship is: Straight line is stronger relationship and in a circular pattern is a weak correlations.
		5. See Figure 18.5, a scatterplot of starting salary and GPA.
		6. Discussion Question: Explain how information about a relationship and its quality can be learned from a scatterplot.
5. **Measures of Central Tendency:** the single numerical value considered most typical of the values of a quantitative variable will discuss the three most commonly used measures of central tendency.
	1. **Mode:** the most frequently occurring number
		1. Can have multiple modes with a set of data: bimodal (2 modes), multimodal (3 or more modes)
		2. If no number has a frequency of more than one you can have multiple modes or no modes.
	2. **Median:** or 50th percentile, the middle point in a set of numbers that has been arranged in order of magnitude (either ascending or descending order)
		1. If you have an odd number of numbers, the median is the center number.
		2. If an even number of numbers, of the median is the average of the two innermost numbers.
	3. **Mean:** the arithmetic average
		1. Formula for mean:
			1. Where *X* is the variable whose values we have*.*
			2. Σ is the Greek letter sigma that tells us to “sum what follows”
			3. *n* = the number of numbers
			4. Sum of the *X* values divided by the number of numbers.
		2. Discussion Question: Explain why when using the example data in the book, the mode and median were the same number, and the mean was a different number.
	4. A Comparison of the Mean, Median, and Mode
		1. **Normal distribution** or normal curve: a unimodal, symmetrical, bell-shaped distribution that is the theoretical model of many variables. See Figure 18.6b.
			1. In a normal distribution, the mean, median, and mode are the same number.
		2. **Skewed:** not symmetrical; one tail is stretched out longer than the other tail, numbers in the tail occur less frequently than the numbers in the “mound” of the distribution.
			1. Negatively skewed: skewed to the left, stretched in the negative direction, where numbers are decreasing in value.
			2. Positively skewed: skewed to the right, stretched in the positive direction, where numbers are increasing in value.
			3. In skewed distributions mean ≠ median ≠ mode
			4. In negatively skewed distribution: mean < median < model, see Figure 18.6a; if the mean is less than the median, the data are skewed to the left or negatively skewed.
			5. In a positively skewed distribution mean > median > mode, see Figure 18.6c; if the mean is greater than the median, the data are skewed to the right or positively skewed.
		3. Mean changes more than the mode or median because it takes into account the size of all of the scores whereas the median only looks at the number of scores, and the mode the score that occurs the most.
		4. Typically, the mean is the measure of central tendency that is used the most, it is the best and most precise.
			1. It is also the most stable from sample to sample.
			2. Mode is used when you want to report the most common data value.
			3. Median is usually preferred if data are highly skewed because it is less impacted by the skew than the mean.
				1. Often in skewed distributions, there is an **outlier** or number that is very atypical of the other numbers in the distribution.
		5. Discussion Question: Explain why the mean is more impacted by skew than the other measures of central tendency.
6. **Measures of Variability:** a numerical index that provides information about how to spread out the data values are or how much variation is present.
	1. How similar or different people are with respect to a variable.
		1. **Homogeneous:** a set of number with little variability.
			1. When numbers are more homogeneous, you can place more trust in the measure of central tendency.
		2. **Heterogeneous:** a set of numbers with a great deal of variability.
			1. When the distribution is heterogeneous, the measure of central tendency is less representative of the data values.
		3. Discussion Question: Explain the relationship between the variability in a data set and measures of central tendency.
		4. Three common indexes of variability
	2. Range: the difference between the highest and lowest numbers
		1. Range = *H –* L
			1. *H* is the highest number
			2. *L* is the lowest number
			3. Only takes into account the two extreme numbers, limited usefulness.
			4. Discussion Question: Explain how the range is calculated and why it is not a useful measure of variability.
	3. Variance and Standard Deviation: most stable measures of variability and the foundations of more advanced statistical analysis.
		1. **Variance:** the measure of the average deviation of points from the mean in squared units.
		2. **Standard Deviation:** the square root of the variance.
			1. An approximate indicator of how far the numbers vary from the mean.
		3. Calculating the variance and standard deviation (text below and example in Table 18.4)
			1. Find the mean of a set of numbers. As illustrated in Table 18.4, add the numbers in Column 1 and divide by the number of numbers. (Note that we use the symbol “X-bar” (i.e., ) to stand for the mean.)
			2. Subtract the mean from each number. As illustrated in Table 18.4, subtract the mean from each number in Column 1 and place the result in Column 2.
			3. Square each of the numbers you obtained in the last step. As illustrated in Table 18.4, square each number in Column 2 and place the result in Column 3. (To square a number, multiply the number by itself. For example, 22 is 2 × 2, which is equal to 4.)
			4. Put the appropriate numbers into the variance formula. As illustrated in Table 18.4, insert the sum of the numbers in Column 3 into the numerator (the top part) of the variance formula. The denominator (the bottom part) of the variance formula is the number of numbers in Column 1. Now divide the numerator by the denominator and you have the variance.
			5. You obtained the variance in the previous step. Now take the square root of the variance, and you have the standard deviation. (To get the square root, type the number into your calculator and press the square root [] key.)
		4. Discussion Question: Explain the relationship between the standard deviation and variance of a set of data.
	4. Standard Deviation and the Normal Distribution
		1. If there is a normal distribution, the following will always be true:
			1. 68.26% of the cases fall within 1 standard deviation.
			2. 95.00% fall within 1.96 standard deviations.
			3. 95.44% fall within 2 standard deviations.
			4. 99.74% fall within 3 standard deviations.
			5. The 68%, 95%, and 99.7% rule
			6. Figure 18.7



* + - 1. Some data are normally distributed: IQ, height, weight
			2. For data collected by researchers, the researcher needs to look at the data and determine if they are normally distributed.
1. Measures of Relative Standing: provide information about where a score falls in relation to the other scores in the distribution of data
	1. Figure 18.8



* + 1. Focus on percentile ranks and standard scores
		2. **Percentile Ranks:** scores that divide a distribution into 100 equal parts
		3. **Standard Scores:** scores that have been converted from one scale to another to have a particular mean and standard deviation.
		4. Also include IQ scores (mean of 100 and standard deviation of 15) and SAT scores (mean of 500 and standard deviation of 100).
	1. **Percentile Rank**: the percentage of scores in a reference group that fall below a particular raw score.
		1. Help to interpret people’s scores in comparisons to others’ scores.
		2. **Reference Group**: the norm group that is used to determine the percentile ranks.
		3. Should be used when the reference group is quite large and representative of a group of interest.
		4. Table 18.5 GRE General Test Interpretative Data: Have students name a student’s percentile rank on the Verbal Reasoning or Quantitative Reasoning test based on their scaled score.
		5. Discussion Question: How are percentile ranks interpreted? Are they the same thing as percentages?
	2. *z* Scores: a raw score that has been transformed into standard deviation units.
		1. *z* scores have a mean of 0 and a standard deviation of 1.
			1. *z* scores tell you how many standard deviations a raw score is from the mean: if it is positive, it is above the mean; if it is negative it is below the mean.
			2. Can be used to compare raw scores between different tests by converting the test raw scores into *z* scores.
			3. The *z* score transformation does not change the shape of the data distribution.
			4. Calculating a *z* score
			5. If distribution is normally distributed, we can use Figure 18.8 to compare *z* scores to percentile ranks.
			6. Discussion Question: Discuss how *z* scores can be used.
1. Examining Relationships Among Variables
	1. Correlation coefficient, contingency tables, and regression analysis are used in both descriptive and inferential statistics.
	2. Correlation Coefficient
		1. Measure of the relationship between two variables.
		2. Numerical index varies between -1.00 and +1.00, negative sign indicates that variables move in opposite directions, positive sign indicates that the variables move in the same direction.
		3. 0 indicates no relationship, the farther from 0 the stronger the relationship.
		4. Measure of a linear relationship.
		5. Discussion Question: Explain how a correlation coefficient is a descriptive statistic.
	3. **Contingency Tables**: a table displaying information in cells formed by the intersection **of** two or more categorical variables.
		1. Table 18.6 describes formation of contingency table
		2. **Rate:** the percentage of people in a group who have a specific characteristic.
		3. Determining whether variables in a contingency table are related.
			1. If the percentages are calculated down the columns, compare across the rows.
			2. If the percentages are calculated across the rows, compare down the columns.
			3. When you follow these rules you will be comparing the appropriate rates).
			4. More variables can be added but if you have three categorical variables, you should examine the original two-dimensional table separatedly for each level of the third categorical level.
	4. Regression Analysis: a set of statistical procedures that are used to explain or predicted the values of a dependent variable on the basis of the values of one or more independent variables.
		1. **Simple regression:** regression based on one dependent variable and one independent variable.
			1. You obtain a **regression equation**, the equation that defines the **regression line** (the line that best fits a pattern of observations).
			2. Important characteristics of any line (including a regression line) are slope (how steep the line is) and *y*- intercept (the point where the line or regression crosses the *y*-axis).
			3. Ŷ = *a* + *bX* where

Ŷ (called Y-hat) is the predicted value of the dependent variable,

*a* is the *y*-intercept,

*b* is the regression coefficient or slope, and

*X* is the single independent variable.

* + - 1. Can use regression equation to make predictions.
				1. Book example of predicting starting salary for people with a 3.00GPA.
				2. When predicting from the regression equation remember you should only use it for values or *X* that are in the range of *X* values in the data.
		1. Multiple Regression: used when there are two or more independent variables.
			1. **Partial regression coefficient:** a regression coefficient obtained in multiple regression.
			2. Partial regression coefficient show the predicted change in *Y* given a 1-unit change in the independent variable while controlling for the other independent variable(s) in the equation.
			3. Can also be used to make predications, book example for predicting starting salary based on GRE Verbal score and GPA.
		2. Discussion Question: Why is regression included as a descriptive statistic.