

Solution and Answer Guide

BUSINESS ANALYTICS: DATA ANALYSIS AND DECISION MAKING, 8E 2025,
9780357984581; CHAPTER 2: DESCRIBING THE DISTRIBUTION OF A VARIABLE

TABLE OF CONTENTS

Answers to Conceptual Questions	1
--	----------

ANSWERS TO CONCEPTUAL QUESTIONS

Note to Instructors: Student answers will vary. The responses here are intended to provide general guidance in terms of concepts that could be discussed.

- C.1.** An airline analyst wishes to estimate the proportion of all American adults who are afraid to fly because of potential terrorist attacks. To estimate this percentage, the analyst decides to survey 1500 Americans from across the nation. Identify the relevant sample and population in this situation.

Answer: The relevant population consists of all people (adults?) in the U.S. who would consider flying commercially. This might be discovered by a survey of a lot of people (several thousand?) that first asks whether they ever have flown or might fly in the future and then asks whether the fear of terrorist attacks would change their minds. Only the people responding “Yes” to the first question would remain in the sample for analyzing the second question.

- C.2.** The number of children living in each of a large number of randomly selected households is an example of which data type? Be specific.

Answer: The “number of children” is a count, so this is a discrete data type.

- C.3.** Does it make sense to construct a histogram for the state of residence of randomly selected individuals in a sample? Explain why or why not.

Answer: A histogram is typically relevant only for a continuous variable that is binned into discrete categories, so it isn’t relevant here. A column chart of counts in states *is* relevant, and it’s a close relative of a histogram.

- C.4.** Characterize the likely shape of a histogram of the distribution of scores on a midterm exam in a graduate statistics course.

Answer: Depending on the difficulty of the exam, the shape could easily be skewed to the left, with the students who didn’t study comprising the long left tail. But it could also be symmetric or even skewed to the right, with a few “brains” comprising the long right tail.

- C.5.** A researcher is interested in determining whether there is a relationship between the weekly number of room air-conditioning units sold and the time of year. What type of descriptive chart would be the most useful in performing this analysis? Explain your choice.

Answer: This should be a time series graph, which is really a scatterplot of the number of air conditioner sales versus time with the “dots” connected.

- C.6.** Suppose that the histogram of a given income distribution is positively skewed. What does this fact imply about the relationship between the mean and median of this distribution?

Answer: The mean will be larger than the median, maybe considerably larger. This is because the large incomes in the right tail pull up the average, but they have no effect on the median.

- C.7.** “The midpoint of the line segment joining the first quartile and third quartile of any distribution is the median.” Is this statement true or false? Explain your answer.

Answer: This is true only if the distribution, or at least the middle 50% of it, is reasonably symmetric. If it’s skewed to the right, say, the median will be closer to the 1st quartile than to the 3rd quartile. This is often clear in boxplots, where the median is *not* in the middle of the box.

- C.8.** Explain why the standard deviation would likely *not* be a reliable measure of variability for a distribution of data that includes at least one extreme outlier.

Answer: The standard deviation, like the mean, is highly sensitive to outliers. Remember that in the definition of standard deviation, deviations from the mean are *squared*, so outliers on either side can really increase the standard deviation. As stated in the book, an accepted procedure when outliers are clearly present is to report measures such as the standard deviation *with* the outliers and *without* them.

- C.9.** Explain how a box plot can be used to determine whether a distribution of values is essentially symmetric.

Answer: There are basically two indications: (1) the median will be about in the middle of the box, indicating that the middle 50% of the distribution is symmetric, and (2) the lines (whiskers) and outliers on either side of the box will be similar, indicating that the more extreme observations are distributed about the same on the low side as on the high side.

- C.10.** Suppose that you collect a random sample of 250 salaries for the salespersons employed by a large PC manufacturer. Furthermore, assume that you find that two of these salaries are considerably higher than the others in the sample. Before analyzing this dataset, should you delete the unusual observations? Explain why or why not.

Answer: It all depends. If these two salaries correspond to employees who are somehow outside the population of interest, such as top executives, it is fine to delete them. But if they aren’t, they shouldn’t be deleted. For example, they might reveal a “shady” company salary policy and hence could represent the most interesting finding of the study.

Solution and Answer Guide

BUSINESS ANALYTICS: DATA ANALYSIS AND DECISION MAKING, 8E 2025,
9780357984581; CHAPTER 3: FINDING RELATIONSHIPS AMONG VARIABLES

TABLE OF CONTENTS

Answers to Conceptual Questions1

ANSWERS TO CONCEPTUAL QUESTIONS

Note to Instructors: Student answers will vary. The responses here are intended to provide general guidance in terms of concepts that could be discussed.

- C.1.** When you are trying to discover if there is a relationship between two categorical variables, why is it useful to transform the counts in a crosstabs to percentages of row or column totals? Once you do this, how can you tell if the variables are related?

Answer: When analyzing a possible relationship between two categorical variables, the joint counts (i.e., the counts of joint categories of the two variables) are all you have to go on. Specifically, it is useful to show these joint counts as percentages of row totals or percentages of column totals. If they are shown as percentages of row totals, for example, where the percentages in each row add to 100%, then the rows should be nearly the same if there is no relationship. Otherwise, you can study the differences between the rows to understand the type of relationship.

- C.2.** Suppose you have a crosstabs of two “Yes/No” categorical variables with the counts shown as percentages of row totals. What will these percentages look like if there is absolutely no relationship between the variables? Besides this case, list all possible types of relationships that could occur. (There aren’t many.)

Answer: As in the previous problem, the two rows will be nearly the same, such as the following, if the variables aren’t related:

		Variable 1	
		No	Yes
Variable 2	No	65%	35%
	Yes	65%	35%

But they could be as follows, indicating that those in the Yes category for Variable 2 are more likely to be in the Yes category for Variable 1 than those in the No category for Variable 2:

		Variable 1	
		No	Yes
Variable 2	No	65%	35%
	Yes	30%	70%

Or they could be as follows, which is the opposite.

		Variable 1	
		No	Yes
Variable 2	No	60%	40%
	Yes	35%	65%

And with only two categories per variable (Yes and No), these are the only two possible forms of dependence, corresponding (roughly speaking) to positive correlation and negative correlation.

- C.3.** If you suspect that a company’s advertising expenditures in a given month affect its sales in *future* months, what correlations would you look at to confirm your suspicions? How would you find them?

Answer: You can imagine two columns, one for monthly sales and one for monthly advertising. If you found the correlation between the two columns, it would indicate the effect of advertising on sales *in that month*. If you think the effect is delayed, you should create one or more lagged columns of advertising (easy with StatTools) that “pushes down” the advertising column by one or more rows. Then, you can check the correlation between the sales column and any of the lagged advertising columns.

- C.4.** Suppose you have customer data on whether they have bought your product in a given time period along with various demographics on the customers. Explain how you could use pivot tables to see which demographics are the primary drivers of their “yes/no” buying behavior.

Answer: You can proceed exactly as in Example 3.5 (frozen lasagna triers). Put any demographic variable in the Rows area, put the “Have Bought” Yes/No variable in the Columns area, and summarize Have Bought by count in the Values area, displayed as percentages of row totals.

- C.5.** Suppose you have data on student achievement in high school for each of many school districts. In spreadsheet format, the school district is in column A, and various student achievement measures are in columns B, C, and so on. If you find fairly low correlations (magnitudes from 0 to 0.4, say) between the variables in these achievement columns, what exactly does this mean?

Answer: Because the correlations mentioned are across various achievement measures, these low correlations indicate that the measures themselves don’t appear to be measuring the same thing. For example, if one measure is of reading ability and another is of mathematical ability, the low correlation would discredit the possibility that the best readers tend to be the best at math. Such a finding might or might not be surprising.