SEAN G. CARVER, PH. D.

THE DATA PROFESSOR'S GUIDE TO BASIC STATISTICS

SELF PUBLISHED

PUBLISHED BY SELF PUBLISHED

http://www.seancarver.org/

Draft, July 2016

I THINK THAT A FAILURE OF STATISTICAL THINKING IS THE MAJOR INTELLECTUAL SHORTCOMING OF OUR UNIVERSITIES, JOURNALISM AND INTELLECTUAL CULTURE. COGNITIVE PSYCHOLOGY TELLS US THAT THE UNAIDED HUMAN MIND IS VULNERABLE TO MANY FALLACIES AND ILLUSIONS BECAUSE OF ITS RELIANCE ON ITS MEMORY FOR VIVID ANECDOTES RATHER THAN SYSTEMATIC STATISTICS.

AUTHOR STEVEN PINKER, *IN AN INTERVIEW WITH THE OBSERVER*

Contents

Dedicated to my students.

Introduction

You hold in your hand a draft of several chapters of a textbook that I have started writing to use in a Basic Statistics class that I periodically teach at American University. The book will be more than just a textbook. In addition, it will simultaneously serve as lecture notes, and a workbook. My lectures will follow this book very closely, and students will follow with their copies of the text during class. Observe the large margins for adding notes. Much of the material will be written in bullet points, and, together with figures, the text might resemble a printout of a PowerPoint presentation, only much better. In the text, I will pose questions to the reader that I will also pose to, and discuss with, the class. I will include materials for active learning exercises, planned for the class. I will provide blank space for answering questions and completing activities. The book will include homework problems, and a companion website will provide data.

Some homework problems in the text will have answers included in the text. My teaching philosophy inclines me to assign problems from this set, although other instructors using the text could make other decisions. However, I plan to make many problems without answers have corresponding similar problems with answers, and identified as such in the text. Finally, I will include problems not guided in this way, for teachers who want to assign them, and students who want to wrestle with a challenge.

My ideas for this book present a huge undertaking, but hopefully I will not work alone. I plan to release all source files for this book on GitHub, available for free download by students and teachers alike. GitHub is a cloud based service for hosting Git repositories. Git is an extremely sophisticated version control software, created to coordinate the activities of thousands of developers working on the Linux kernel. Git will facilitate the maintenance of a virtually unlimited number of versions of this textbook. Instructors can easily change the book to suit their needs, even changing it again for different sections of the of the same, or different classes. Contributors can share these changes back to the central repository, or not, either

as separate branches, or merged with other versions in a myriad of different ways.

I plan to make this book available to students, instructors, and contributors under an open-access, attribution, share-alike license, where contributors keep their copyrights to their contributions, but must provide access to their work under a compatible license. I hope to encourage many to contribute material to this endeavor and make this text truly outstanding.

[Aside: Until I have a chance read through and understand all the legal ramifications behind my choice of license, this printing is offered with "all rights reserved." Additionally, the source is still maintained under a private Git repository. I expect all of this to change in the coming weeks.]

I plan to also write a guidebook intended to be used and read by collaborators of this project. The software tools I plan to use to write this document have substantial learning curves, all of which merit the allocation of my time and energy to help my would-be collaborators surpass. Additionally, the guidebook will draw from, and cite, the GAISE report (Guidelines for Assessment and Instruction in Statistics Education), and maybe other sources, as valuable references. Of course, the guidebook, itself, will be a collaborative document, just like this one. Finally, for both of these projects, I plan to make use of the wiki and issue tracker that come standard with a repository hosted on GitHub.

One final issue has to do with the statistical software I will use to present the graphs and results of computations in this textbook. At American University many instructors use StatCrunch. StatCrunch is an effective pedagogical tool that proves easy, even trivial, for many students to learn. StatCrunch can be accessed and used with a browser (and used for free with American University credentials).

All that said, I do not plan to use StatCrunch at all for this textbook. I will use R, exclusively. The text will only show graphs and results generated with R. Nevertheless, for now, the lectures will still involve StatCrunch. In other words, while discussing the R figures and results in the text, I will, during class, teach students how to generate similar graphs and results with StatCrunch, if at all possible. (Much of what one can do in R remains impossible in StatCrunch.) During class, students will take notes and try StatCrunch out on their laptops. Students interested in R can read the *behind the scenes* addendum to each relevant chapter, which will explain how to generate, with R, all of the actual graphs and results shown. These optional addenda, not discussed in class, will give interested students a complete course on R. Class projects may motivate students to use R to exceed the capabilities of StatCrunch.

During exams, given in a computer lab, students will have both R and StatCrunch available for them to use to complete their exam problems. For these tests, I will provide a crib sheet, available ahead of time, and/or at the end of this book, listing the R commands presented in the R addenda. The commands on the crib sheet will include all the commands needed to solve the exam problems given to the class.

Here are the advantages of this R approach:

1. Although learning StatCrunch may provide skills transferable to learning other statistical software, by itself, StatCrunch is useless in the real world. No employer, to my knowledge, desires candidates with StatCrunch skills. Many want R skills. We should oblige Basic Statistics students who want to learn something more sophisticated, even if we do not require it at this level.

2. Although the main curriculum for Basic Statistics at American University does not require use of anything beyond StatCrunch, many students exceed its capabilities with projects assigned in the class.

3. Guiding students through the menu-based (GUI) StatCrunch can be difficult in a textbook. Command-line based R is much better suited for explaining what to do.

4. R is free for everyone, whereas StatCrunch may not be free outside of the American University community. I am looking far and wide for users of and contributors to this textbook. Open source software becomes an imperative for this endeavor.

5. The real kicker for me (which elevates R above both StatCrunch and *all* its other alternatives): R has sophisticated tools for authoring books. I can embed R code into the LaTeX files that comprise the source for this book. Compiling the document will run the R code that will create the results, tables and figures, and embed these results into the document. Additionally, the corresponding code (the actual code that was used to generate the results) can be automatically embedded into a different part of the document (e.g. the behind the scenes addendum). On top of all that, these tools facilitate version control, not just on the text, but also, on everything else involved including the results, tables, figures, and code. Collaborative authoring without these tools would be a *nightmare*.

Defining familiar terms

We are going to start with a game. I will give you a familiar word, and you will try to articulate a precise definition. Don't look ahead: I do give answers on subsequent pages, but the purpose of this exercise is to help you remember the definitions I use. The effort of trying to formulate your own definitions should help you remember mine. This game will be hard to play, so let yourself be pushed beyond your comfort zone.

In the box below, jot down your ideas or those discussed in class. When we are done playing the game, we will find out what the American Statistical Association has to say about this matter.

What does the word "statistics" mean?

So, what does *statistics* mean? On their website, the American Statistical Association (ASA) provides the following definition and a citation[1]: "*statistics is the science of learning from data...*" [1]

6. The ASA classifies statistics as a science, not as a type of mathematics, although everyone agrees that statistics draws heavily on mathematics—as do many other sciences.
7. Specifically, statistics is the science of *learning from data*.
8. What about data? Round two.

What does the word "data" mean?

So what does the word *data* mean?
Data are descriptions of objects, people, or events under study.

9. Let's unpack this definition. It has three parts. First, data are *descriptions*. And second, data are descriptions of *things under study*. And third, these things are always *objects, people or events*.

10. The word *data* is plural; its singular form is *datum*. Be careful with subject-verb agreement. *The data is compelling* (incorrect). *The data are compelling* (correct).

11. Examples of data: heights, weights and ages of varsity student athletes at a university. These are all *numbers* describing athletes.

12. More examples for the athletes: their gender (female, male) and the team (e.g. basketball, lacrosse, swimming, etc.) that they play for. These are both *categories* describing athletes.

13. Can you think of other examples of data that are numbers?

14. Can you think of other examples of data that are categories?

15. Most statistics involve data of one of two kinds: numbers and categories. Different statistical methods apply to different kinds of data. We will study methods for both of these kinds.

16. Another kind of data is raw data, described below.

17. *Raw data* are data that require processing before statistical analyses can be performed.

18. Raw data might be numbers or categories, but often they might best be described as something else. See examples, below.

19. Examples: Images, audio and video are raw data that are not best described as numbers or categories. Individual pixels can be described by numbers, but usually individual pixels are not, by themselves, useful to statisticians.

20. Example: your raw data consists of videos of the courtship rituals of songbirds. By watching the videos, you derive a number (length of song in seconds) and a category (success or failure to mate) to describe the courtship event.

21. We will only work with numbers and categories in this class.

Let's collect some data!

What is your favorite color? I'll count the number of people in the class for each color. Record the results below. We will use it later.

color (blanks for other)	number
red	
orange	
yellow	
green	
blue	
purple	
white	
gray	
black	
brown	

Concepts of structured data

22. Computers always store data (numbers, categories, or raw) as patterns of bits. Almost all statistics involve computers these days, but can you think of different ways to store data, not as a patterns of bits?

23. Data can be *structured* or *unstructured*. I explain the distinction below.

24. Structured data can be naturally stored in a spreadsheet, using one or more tables (also called sheets)

25. Recall: a table of a spreadsheet is a two-dimensional structure with rows and columns. Structured data has this format.

26. An example of structured data: the records of varsity student athletes, mentioned above. Traditionally, each player gets a row and each characteristic gets a column. Thus, there are columns for height, weight, age, gender, and team.

27. Can you think of more examples of structured data?

28. Unstructured data cannot be naturally stored in a spreadsheet.

29. An example of unstructured data: the archive of data from twitter including its author, text, hash tags, mentions and places.

30. Can you think of more examples of unstructured data?

31. In this class, we are only going to consider structured data.

32. A *case* is a single object, person or event described by the data.

33. Remember: we defined *data* as "descriptions of objects, people, or events under study," so the cases are those objects, people, or events.

34. Examples of cases: lots of a drug being manufactured, patients under care of a hospital, or stock trades made by a firm.

35. What were the cases in the student athlete data example?

36. A *variable* is a characteristic of a case, recorded in the data.

37. Variables could be dosage of the drug, age and diagnosis of patient, and company, price, and date of trade of stock.

38. What were the variables of the student athlete example?

39. Traditionally, rows hold cases, whereas columns hold variables.

40. The *values* make up the individual entries in the table.

41. We say that a variable has a value for a case.

42. If the cases are people, we often call them *subjects*.

43. If a data set contains more than one table, each table could refer to different cases.

44. For example: Amazon.com might have a table for all its *customers*, a table for all its *products*, and a table for all its customers' *orders*.

45. In the Amazon.com example, relationships exist among the data from different tables: certain customers place certain orders for certain products. You would use a *relational database* to manage these issues.

46. We will not deal with these complexities in this class. All our data will fit onto a single table.

47. What are the cases in our favorite color data set? What is (or are) the variable(s)?

What are the cases in the favorite color data set?

What are the variables in the favorite color data set?

48. So what are the cases of the favorite color data set? It is tricky to answer this question because what often comes to mind first, while not being wrong, is not really the best answer. Many people say the cases in the favorite color data set are the colors, and variable (only one) is the number of people in our class who have that color as their favorite. This answer is suggested by the structure of the data we collected.

49. But are we really studying colors? What are we studying? That's the best answer for what are the cases!

What objects, people, or events are under study in the favorite color data set?

50. So what are we studying in the favorite color data set? I think the best answer is that we are studying *the students in our class*. This answer suggests that the students in our class comprise the cases for this data set.

51. But what about the variables? And what about the structure of the data?

52. Could we rewrite the table so that each student has their own row?

53. Look below, the two tables hold the same data, although the second also has students' names, which were not recorded in the previous data set.

white	2
gray	3
black	1
brown	0

sally	white
john	white
zoe	gray
ivan	gray
charlotte	gray
jane	black

Technically, the first data set is a *summary* of the second. The concept will be important, later.

54. The second way makes clear that the cases are the *students* and the variables are the student's favorite *color* and the student's *name*.

55. Note that we did not need to add the names: our new data set could have been just one column of colors, with some colors repeating.

56. But if the new variable was not there, could you understand the data as easily?

57. The new name variable involves data that we did not record in our original (summary) data set.

58. The *student name* variable exemplifies the concept of a *label*, defined below.

59. A *label* is a variable that distinguishes or identifies the cases.

60. To be a label, it must hold a unique value for each case, otherwise it would not distinguish or identify the cases. But see complication, below.

61. Complication: sometimes data sets use more than one variable as a label. See example below.

62. For example, we might need both the *first name* and the *last name* to distinguish or identify the students (if names repeat).

63. What other options would we have?

Kinds of variables

64. *Quantitative* variables hold numbers whereas *categorical* variables hold categories—the only types of variables we will consider in this class.

65. What about labels? While not terribly useful, we can think of labels as categorical variables. If the data set uses just one label, then each case has its own unique category under this variable.

66. I have heard some people say *qualitative* as a synonym for categorical.

67. However, do not say *numerical* instead of quantitative: the two terms refer to different concepts. Consider the next bullet point.

68. The data records "1" for male and "2" for female. (Things like this happen all the time with real data).

69. The ones and twos are numbers, so the variable is numerical, but is it quantitative and not categorical?

70. Best way to tell the difference: if you have a variable holding numbers, ask yourself, are arithmetic operations (especially adding and averaging) meaningful for these numbers? If so, it is quantitative. If not, it is probably categorical or raw.

71. On the other hand, if the variable places each case into one of two or more categories, it is categorical.

72. It is usually very easy to tell the difference between a categorical variable and a quantitative variable, but in some cases the variable could be interpreted in either way. How?

73. Consider a different data set that records "0" for male and "1" for female (and let's say the cases are the people in our class). Obviously this is still a categorical variable, but is there a way to interpret the data as quantitative?

74. What if we interpret the variable (either 0 or 1) as *the number of females that the presence of the subject adds to the class*.

75. Interpreted as above, the variable is quantitative.

76. What is the sum of the values of this variable (0 for men and 1 for women), for all cases in the data set (i.e. all students in this class)?

What is the sum of the values of this variable?

77. What about average?

What is the average of the values of this variable?

78. What is the sum of the values of the variable across for all case? The sums of the zeros and ones in the above example is the *count* of the number of women in this class.

79. What is the average of the variable across this data set? The average of the zeros and ones in the above example is the *proportion* of women in the class. If 60% of students in this class are women, then this average is 0.6.

80. Both counts and proportions are very important in statistics. We will see them both again later.

81. The example above applies to any categorical variable with only two possible categories. Just assign 0 to one category, and 1 to the other.

82. A *binary categorical variable* is a categorical variable with only two possible variable categories.

83. The example, above, reveals a connection between statistics for binary categorical variables (involving counts and proportions) and statistics for quantitative variables (involving sums and averages). We will explore this connection, later.

84. There is another distinction between categorical variables: *ordinal* categorical variables versus *nominal* categorical variables. I explain the difference below.

85. *Ordinal categorical variables* are categorical variables with a natural *order*.

86. For example, some surveys pose a statement to the respondent then require a multiple choice answer: (1) Strongly Disagree (2) Disagree (3) Agree (4) Strongly Agree. Can you see the order in these categories?

87. Can you think of other ordinal categorical variables?

88. *Nominal categorical variable* are categorical variables that lack a natural order and are related by *name* only. An example of this kind of variable are the favorite colors that we collected above.

89. Can you think of other examples of nominal categorical variables?

90. In many universities in the U.S. grade students with a "letter" (one of A, A-, B+, B, B-, C+, C, C-, D, or F). What kind of variable is the grades variable?

What kind of variable is the "grades" variable?

91. So what kind of variable is the grades variable? I believe the best answer that it is a ordinal categorical variable.

92. Why the hesitation? The situation is a little complicated by the fact that there is a standard translation between grades and numbers: an A is a 4.0; an A- is a 3.7, etc.

93. If the categories were expressed as numbers 4.0, 3.7, etc., would the variable be quantitative?

94. Would summing or averaging the grade points make sense?

95. Consider this: one's grade point numbers are frequently averaged to form the much fretted over "grade point average" (GPA). That clearly suggests the variable is quantitative.

96. Still people argue that the answer is no, the variable is not really quantitative, because the numbering is arbitrary.

97. Consider the following question: is the difference between an A and an A- really the same as the difference between a B and a B-? (If you are unfamiliar with this scheme a B gets 3.0 points and B- gets a 2.7, suggesting the differences should be the same).

98. Many students (and employers) would think the difference between an A and and A- is much smaller than between a B and a B-.

99. Thus people have argued that, no, averaging the grade points to create the GPA fundamentally does not make sense. It follows that the grades variable is ordinal categorical.

100. But this position is open to interpretation. The question concerning whether grades are quantitative or categorical really depends on your perspective. Do you have an opinion? Do you agree or disagree that the translation between letters and numbers is arbitrary, in the sense explained above?

Distributions

101. The concept of a *distribution* is absolutely central in probability and statistics.
102. In an advanced book, you will get a mathematical definition of a distribution.
103. We have to settle for the following (which while imprecise, conveys the idea):
104. The *distribution* of a variable tells us (1) what values the variable takes and (2) how often the variable takes these values.
105. The best way to visualize a distribution is with a graph.
106. The kinds of graphs we draw for categorical variables is different from the kinds of graphs we draw for quantitative variables.
107. For categorical variables we draw pie charts and/or bar graphs.
108. For quantitative variables we draw stem plots and histograms.
109. Let's graph the favorite color variable of our favorite color data set.
110. Let's graph the categorical variables of the diamonds data set.
111. Homework 1.
112. Stem plots and homework 2.
113. Histograms and the call center data set.
114. Let's graph (some of the) quantitative variables of the diamonds data set.
115. Homework 3.

Exploratory data analysis

116. When you do *exploratory data analysis* you examine data to describe its main features.

117. The key word in the above definition is *describe*. With exploratory data analysis, our goal is simply a description of a data set's main features, not inference from the data.

118. Exploratory data analysis is generally the first thing you do with a new data set.

119. If there are only a few variables, you can start by graphing the distribution of each.

120. Single variables tell only a limited story. You also want to look at relationships between and among variables.

121. The next level of complication is to look at relationships between *pairs* of variables.

122. Of course you don't have to stop there. You can look at relationships among 3, 4, 5, or more variables. But with more than two variables, things can get very complicated.

123. In this class, we will look at single variables and pairs of variables, but no more.

124. For multiple variables, there is a generalization of the concept of distribution for more than one variable. It is called joint distribution of two or more variables. More about that later...

125. After creating graphs to understand the variables, alone or in pairs, the next step is to create numerical summaries of the data. We will soon talk a lot about that.

126. If there happens to be many variables in the data set (some data sets have thousands), graphing each one is impractical. And graphing each pair of two is even worse.

127. In that situation, look at the cases and variables. What cases do the data describe? What characteristics of the cases do the variables describe? You might graph the distribution of a few variables, but ultimately what you want to do is formulate a question about the data.

128. Formulating a question about the data is still a good thing to do with small data sets, as well.

129. Once you have a question, you try to answer it.

130. Once you answer your question, you try to formulate another question.

131. You repeat the process until you have gleaned some insight into the data.

132. That's all you can hope for. With a really big data set, with many variables, it may not be possible to completely understand the whole body of data.

133. The quality of your questions, and your success in answering them, will determine the value of your work.

134. What questions can we formulate about the diamonds data set?

Mean

135. The *mean* is a statistic used for a single quantitative variable.

136. Thus, we can take the mean of a set of quantitative observations like IQ, shoe size, height, weight, etc., but not a set of categorical observations like gender, party affiliation, etc.

137. Test scores are quantitative. Let's say we have the following test scores (and everyone did really well): 90, 92, 94, 96, 98, 100. We want to find the mean.

138. Most students are already familiar with the formula:

$$\bar{x} = \frac{90 + 92 + 94 + 96 + 98 + 100}{6} = 95.$$

139. Now we give each data point a number, called an index:

$$\begin{aligned} x_1 &= 90 \\ x_2 &= 92 \\ &\vdots \\ x_6 &= 100. \end{aligned}$$

140. If we want to refer to an *arbitrary* data point we use the letter i. In other words x_i is the i^{th} data point. Here i stands for a number, either 1, 2, 3, 4, 5, or 6. The subscript i is called the *index*.

141. Finally, if we want to refer to the *total* number of data points (in this case 6) we use the letter n. This use of n is common in statistics.

142. We use the sigma notation to write the formula for the mean:

$$\bar{x} = \frac{1}{n} \sum x_i.$$

143. The symbol Σ is the capital form of the Greek letter *sigma*. It stands for *sum*.

144. Other branches of mathematics require *limits* on the sum, such as

$$\sum_{i=1}^{6} x_i$$

This notation means to sum the data points x_i for values of the index i ranging from 1 to 6.

145. Statisticians often leave the limits off the sum. In this case, it is implied to sum over all of the data: sum from i ranging from 1 to n, which is the same thing as above.

146. Finally the coefficient $\frac{1}{n}$ in front of the Σ tells us to divide the sum by the total number of data points, n, in this case 6, as above.

147. The mean is a measure of the center of the distribution.

Sampling

148. What is sampling? Let's proceed with our example from the previous chapter: we have the test scores of 6 students: 90, 92, 94, 96, 98, 100. The mean of these test scores is 95.

149. Sampling would be indicated if, in addition to these 6 students, there were many more in the class and we wanted to use the 6 students to study the properties of the whole class.

150. Exactly what properties? Later, we will work with other statistics, but for now our focus rests squarely on the mean: we want to assess the *mean* test score of the whole class by just looking at the 6 students.

151. My example is not really best, because, in practice, the teacher would probably have all grades for the all students (even if there were 1000 students). Often the grades are in BlackBoard or in a spreadsheet and it takes one command to calculate the class mean.

152. Sampling only gives an approximation to the right answer. In the case of grades, the right answer is readily available, so sampling is not necessary or even advisable.

153. However, in many situations it is impractical to collect data on all imaginable cases.

154. Example: if you are doing a survey, you can't feasibly ask every person in the world. But it is still possible to study the world's population by sampling using a much smaller group.

155. Let's proceed with our grades example ignoring that it is often inappropriate for this application.

156. For our grades example, our sample size was 6. Six is an unusually small sample size. The larger the sample, the better.

157. Let's say the 6 students are among 1000 students in the whole class.

158. The 6 students comprise the *sample*.

159. The 1000 students (which must include the 6) comprise the so-called *population*.

160. We want to study the whole class mean: if we had access to all the grades we could find this number exactly (without sampling) by adding all 1000 grades and dividing by 1000: the usual mean.

161. But let's restrict ourselves to only the 6 students. What could we do?

162. A reasonable approach is to calculate the *mean* of the sample, or sample mean. As shown above, the sample mean is 95. That's our *estimate* of the population (i.e. whole class) mean.

163. The sample size is customarily written with the lower case letter n. In our example, $n = 6$.

164. If the variable in question is x, the sample mean is customarily written with the notation $\bar{x}$:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

165. The sample mean is an example of a *statistic*.

166. A statistic is a number that describes a sample.

167. The population size is customarily written with the upper case letter: N. In our example, $N = 1000$.

168. The population mean is written with the greek letter mu: μ. Alternatively, if the variable in question is x, this is sometimes indicated as μ_x:

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i$$

169. The population mean is an example of a *parameter*.

170. A parameter is a number that describes a population.

171. Mnemonic: *S*ample and *s*tatistic go together: they both start with *s*. *P*opulation and *p*arameter also go together: they both start with *p*.

172. You use $\bar{x}$ as an estimate of μ_x, but your answer depends on your sample.

173. Specifically, if your 6 students all happen to be above average, your estimate will clearly be too high. This scenario is clearly possible.

174. If your 6 students all happen to be below average, your estimate will clearly be too low. This scenario is also clearly possible.

175. If some of your students are above average and others are below, then your estimate might be too high or it might be too low, but it is unlikely (though possible) that it will be *exactly* right.

176. Therefore sampling rarely gives the right answer.

177. But in this case, it gives an *unbiased* estimate, a concept that will be made more precise below.

178. An *unbiased* estimate is one that is neither prone to being too high nor prone to being too low.

179. What does that mean? With 1000 students we can pick our 6-person sample in many ways. In fact, there are exactly 1,368,173,298,991,500 ways of picking a six member sample from a population of 1000. (Huh, you say? I'll show you how to count samples, in the next chapter.)

180. Each sample leads to a different sample mean (although some values may repeat). Some of these values are too high, and some of these values are too low, and maybe a few values are just right.

181. Thus, there are approximately 1.37 quadrillion possible sample means (including repeats).

182. Each sample mean is a mean of 6 values (but different values for each sample).

183. What would happen if I added all 1.37 quadrillion sample means then divided by 1.37 quadrillion?

184. I would get the mean of the sample means!

185. The sense in which the sample mean is an unbiased estimator of the population mean: the mean of all possible sample means equals the population mean!

186. In other words, the mean of all possible estimates is the quantity you are trying to estimate.

187. In this sense, an unbiased estimator is neither prone to being too high, nor prone to being too low.

188. An unbiased estimator is exactly correct on average. Individual estimates will likely be too high or too low, but those errors cancel out when the average is taken.

189. Caution: our result depends on the condition that the sample be chosen at *random*.

190. For example, if the high values are more likely to be chosen than low values, then clearly the estimator will be prone to estimates that are too high.

191. A *simple random sample* is one in which all of the possible samples have an equal chance of being chosen.

192. Our grades example employs a simple random sample only if each of the 1.37 quadrillion samples had an equal (1 out of 1.37 quadrillion) chance of being chosen as the sample that we used.

193. There are other strategies for sampling, which will be discussed in time.

194. However if the professor were to select her favorite 6 students as her sample, she should not expect an accurate assessment of the whole class mean.

195. Let's explore simple random samples:

196. Suppose we want to draw a sample of 2 people from the following population of 4 people: Amy, Betty, Carl, and Dennis, each denoted by his or her initial: A, B, C, and D.

197. There are 6 possible samples: AB, AC, AD, BC, BD, and CD.

198. Each person appears in exactly half the samples. Thus each person has an equal chance of being in the sample.

199. To draw a sample as a simple random sample, we could assign six-sided die face to each of the 6 possible samples, then roll the die to make the selection. Each person would have the same probability of landing in the sample: $1/2$.

200. Some people mistakenly believe that a simple random sample means that each person has a equal probability of being in the sample.

201. Let's explore this scenario. Let's suppose we don't have a die—we only have a coin and we get lazy. We assign AB to heads and CD to tails. Then again each person in the population has the same probability, $1/2$, of being in the sample, but not every sample can be chosen: there are no coed samples possible!

202. Characteristics (such height differences among members of the sample) for which single-sex samples do not fully represent the population would not be well-studied with this sampling scheme.

203. For a simple random sample, in this example, we must make our selection among 6 samples, not 2.

Counting samples

204. How many ways are there to choose a sample of n individuals out of a population of N individuals.

205. This number has a name. It is called, appropriately enough, "N choose n".

206. The following is a mathematical notation for this number:

$$\binom{N}{n}.$$

207. What is the number $\binom{N}{n}$?

208. Calculating this number is based on counting the number of ways of arranging the N individuals in the population into an order.

209. First, how many ways are there of arranging the letters $ABCD$? There are 4 choices for the first letter, 3 for the second, 2 for the third, and 1 for the fourth: $4 \times 3 \times 2 \times 1$.

210. This number is better denoted 4!, read "four factorial." Basic arithmetic will tell you that $4! = 24$. Likewise there are $N!$ (N factorial) ways of ordering the N individuals in the population.

211. Having enumerated all the ways of ordering the N individuals in the population, how do we pick a sample from the ordering?

212. It doesn't really matter how we pick the sample, so let's just pick one way and be consistent: from each ordering, pick the first n individuals from the ordering as the sample.

213. We have found a way of counting orderings of the population, and picking samples from ordering. Now we try to count the ways of sampling n individuals from the population of N.

214. Unfortunately, counting orderings of the population will over count the number of samples, because we can change the ordering of the population without changing the sample. Indeed, if we just reorder the first n individuals, the sample, as we have picked it, doesn't change.

215. In our example, with sample size 2, every sample of the correct size has two reorderings: for example, we can reorder AB as AB or BA. Note that we count the original ordering AB as one its possible reorderings.

216. So we should divide the number of orderings by at least 2 to get the number of samples—but we are not quite done, yet, because there is a second way of reordering the population without changing the population. In general, every sample of size n will have $n!$ possible orderings, so we should divide $N!$ by at least $n!$ to get the number of samples—but we are not quite done yet.

217. We are not quite done, yet, because, as mentioned, there are actually two ways of reordering of the N individuals in the population without changing the sample: we can reorder the first n chosen as the sample, as done above, *or* we can reorder the last $(N - n)$ left out of the sample.

218. The following 4 orderings all give the same sample AB: $ABCD$, $BACD$, $ABDC$, and $BADC$.

219. Indeed, there are 4 orderings for each of the 6 of the possible samples; so we need to divide 24 by 4 (or divide 24 by 2 twice), which gives 6, as expected. In general, we need to divide $N!$ by $n!$ and then divide again by $(N - n)!$, which gives the following result:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

220. Another way of calculating $\binom{N}{n}$ is with Pascal's Triangle. Here is Pascal's Triangle:

$N = 0$:											1										
$N = 1$:										1		1									
$N = 2$:									1		2		1								
$N = 3$:								1		3		3		1							
$N = 4$:							1		4		6		4		1						
$N = 5$:						1		5		10		10		5		1					
$N = 6$:					1		6		15		20		15		6		1				
$N = 7$:				1		7		21		35		35		21		7		1			
$N = 8$:			1		8		28		56		70		56		28		8		1		
$N = 9$:		1		9		36		84		126		126		84		36		9		1	
$N = 10$:	1		10		45		120		210		252		210		120		45		10		1
											⋮										

221. The rows are labeled by the population size, N, top to bottom, $N = 0, 1, 2, 3, \ldots$,

222. There is no limit to the number of possible rows in Pascal's triangle, but the first row, the top row, the apex of the triangle, corresponds to $N = 0$, the smallest population size.

223. It is not much of a population, if it has no individuals, but the row for $N = 0$ is there for completeness.

224. The Nth row has $(N + 1)$ entries.

225. These entries correspond to the sample size, n; left-to-right as $n = 0, 1, 2, 3, \ldots N$.

226. The smallest possible sample size is zero: $n = 0$; and the largest possible sample size is the size of the whole population: $n = N$.

227. For both smallest and largest samples, there is only one possible way to draw the sample (respectively, no one in the sample, or everyone in the sample).

228. The entry corresponding to N and n equals $\binom{N}{n}$, the number we want to calculate. Note the entry for $N = 4$ and $n = 2$ is 6 as we expected. (Remember to count n from 0, not 1.)

229. The first and last entries in a row are always 1, for the smallest and largest samples. After that, do you see the pattern? To get any other entry, add the two entries above it: the one to the left, and one to the right.

Problem 1: A deli gives patrons the option of 3 different breads (rye, pumpernickel, and white), 2 different meats (chicken and roast beef) and 8 different toppings (lettuce, tomato, banana peppers, avocado, grated cheese, relish, black olives and garlic). How many ways can you make a sandwich with exactly 1 bread, exactly 1 meat, and exactly 4 different toppings? (One possible sandwich that meets the criteria is roast beef on rye with lettuce, tomato, avocado, and black olives.)

Problem 2: What is the row of Pascal's triangle corresponding to $m = 11$?

Solution 1:

$$3 \text{ breads } \times 2 \text{ meats } \times \binom{8}{4} \text{ toppings}$$

This simplifies to 420 sandwiches.

Solution 2: The $m = 11$ row of Pascal's Triangle is

$m = 11$: 1 11 55 165 330 462 462 330 165 55 11 1

Standard deviation

230. The standard deviation is a measure of the spread of the distribution—in other words, how close, or how far, do the data tend to fall from the mean?

231. I'll start with a formula, explained below. Confusing: there are actually two formulas for standard deviation, and many calculators give you a choice.

$$\begin{aligned} s_n &= \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2} \\ s_{n-1} &= \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2} \end{aligned}$$

232. In the context of *sampling* (discussed below), the second formula is correct. Indeed, if there is only one choice given in a book or a calculator, it is usually the second one. I will call the second formula the *more common formula*, and the first formula the *less common formula*.

233. Which formula should you use? Short answer: always use the more common formula. Use of the less common formula should be noted and justified, and I would say just don't bother!

234. Why are we discussing the less common formula? Because most students have a hard time understanding the more common formula, and think the less common formula makes more sense. It *does* make more sense in certain contexts. And I think it is important to discuss these contexts to better understand the more common formula.

235. Let's unpack the formulas.

236. The quantity $(x_i - \bar{x})$ is called the deviations, or deviations from the mean.

237. Deviations are important because we are trying to estimate how far the data points fall from the mean.

238. Each data point has a corresponding deviation from the mean.

239. Consider the same test score example. The first observation, 90, is 5 points *below* the mean (which was 95). So its deviation from the mean is -5.

240. Can you guess what the other deviations from the mean are?

241. The other deviations from the mean are: -5, -3, -1, 1, 3, 5.

242. Because the data points are equally spaced, the deviations have a nice pattern to them. This pattern will not be there in most data sets.

243. However, it will always be the case that the deviations add to zero.

244. Unless all deviations are actually zero, some will be positive and others will be negative, in such a way that they will balance out, adding to zero—this property results from the fact that the deviations are from the mean and the mean is the center of the distribution.

245. Because it is useless to average the deviations (the average will always be zero), we first square the deviations: $(x_i - \bar{x})^2$. For our data set the squared deviations are: 25, 9, 1, 1, 9, 25.

246. Unless a deviation is zero, its square is positive.

247. The next step is to "average" the squares of the deviations. This number will be positive unless all of the deviations are zero.

248. The less common formula for s_n uses the mean of the deviation as the average: $\frac{70}{6} = 11.6667$.

249. The more common formula for s_{n-1} uses an adjusted mean—adjusted for the so-called number of degrees of freedom or $n - 1$: $\frac{70}{5} = 14$. The adjusted mean is what is used as the average.

250. The last step, in both formulas, is to take the square root of the result: either $\sqrt{\frac{70}{6}} = 3.4157$ or $\sqrt{\frac{70}{5}} = 3.7416$. Because we square the deviations in a previous step, we take the square root, so that the result can more easily be compared with the mean (without taking the square-root the units change).

251. The more common formula for s_{n-1} always gives a larger value than the less common formula for s_n.

252. The larger the value of n, the less difference there is between the results given by the two formulas. The difference between dividing by 6 or dividing by 5 is much greater than the difference between dividing by 1000 or dividing by 999.

253. If you skip the square root step you are left with a quantity that is also important in statistics. It is called the *variance*. The variance has different units than the data.

254. As mentioned above, in the context of sampling, we should use the more common formula for s_{n-1}. In this context, $\bar{x}$ is called the sample mean, and s_{n-1}, also written as s, is called the sample standard deviation.

255. The more common formula arises in the context of sampling.

256. Now, in addition to estimating the population mean to assess center of the distribution, you may want to estimate the population standard deviation to assess the spread in the distribution—things get complicated.

257. The unequivocal right answer to the population standard deviation uses the *less common formula!*, summing over all 1000 students and using the unadjusted average and substituting the population mean for $\bar{x}$.

258. The question is: what is an appropriate estimate of the population standard deviation using our sample of 6 students, rather than all 1000?

259. Which formula you should use depends on what you use for $\bar{x}$.

260. If you use the population mean for $\bar{x}$, as above, you would use the less common formula, employing the usual unadjusted average. This is almost never done for lack of access to the population mean.

261. If you use the sample mean for $\bar{x}$ (after all, you want to avoid dealing with all 1000 students) you will get an answer with is prone to being too small, unless you correct it by changing the notion of average.

262. To fix this problem, you use the adjusted mean, which appropriately increases your estimate, so that on average, you get a result which is neither prone to being too high, nor too low.

263. The question is: why does the less common formula lead to an estimate which is prone to being too low?

264. Consider this fact: the correct result involves an average of squared-deviations from the population mean.

265. But now consider this fact: We want to take an average of squared deviations from the *sample mean*, not population mean.

266. The problem is, deviations from the sample mean are prone to being smaller than deviations from the population mean.

267. Why? Consider this example: What if the population mean test score, instead of being 95, was in fact 70. Our sample wouldn't be representative of the population, but that can happen, some times.

268. The deviations from the population mean would be between 20 and 30 whereas the deviations from the sample mean would be between 1 and 5, as shown above. The sample mean deviations would be too small.

269. The example above is extreme, but any time the sample mean is different from the population mean, we have a problem, because the sample mean is a better estimate for just the sample (it was derived from the sample) than for the whole population involving all the data.

270. The sample is closer to the sample mean than the population mean, but the population mean gives the right answer. The sample mean's result is too small, but we can correct this by adjusting our notion of average.

271. So why divide by $n - 1$ instead of something else, like $n - 2$. I am not sure if anyone has a satisfying non-mathematical answer to this, although it is clear from a mathematical calculation.

272. It should be pointed out that $n - 1$ is the number of "degrees of freedom" in the deviations.

273. There is one less degree of freedom in the deviations than the total number of deviations because they are constrained to add to zero as mentioned above, so you are really averaging $n - 1$ independent quantities instead of n.

274. Some books justify the more common formula with this argument concerning the degrees of freedom in the deviations, but for me the explanation falls flat and doesn't tell the whole story.