SEAN G. CARVER, PH. D.

THE DATA PROFESSOR'S GUIDE TO BASIC STATISTICS

SELF PUBLISHED

PUBLISHED BY SELF PUBLISHED

http://www.seancarver.org/

Draft, June 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..."*

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.

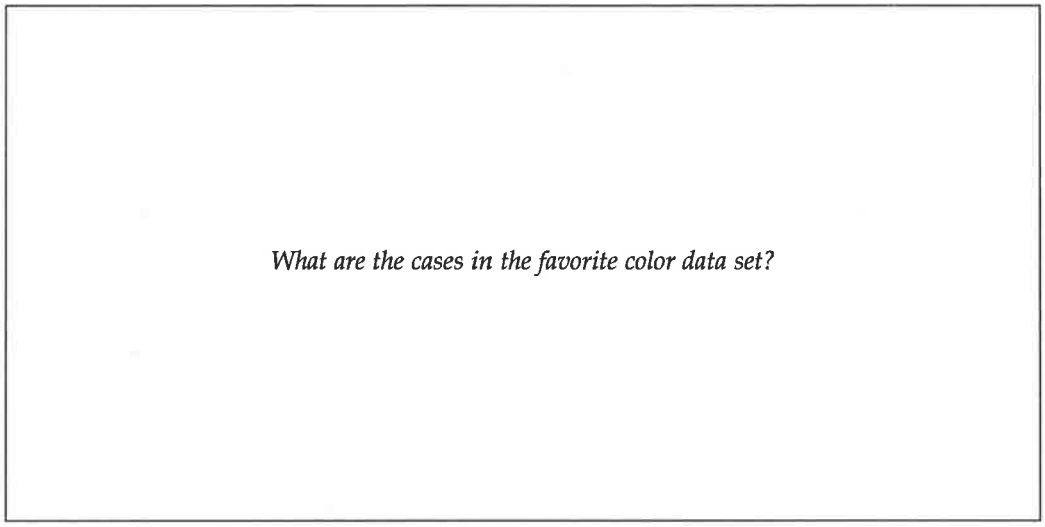
33. Remember: we defined *data* as "descriptions of objects, people, or events under study," so the cases are those objects, people, or events.

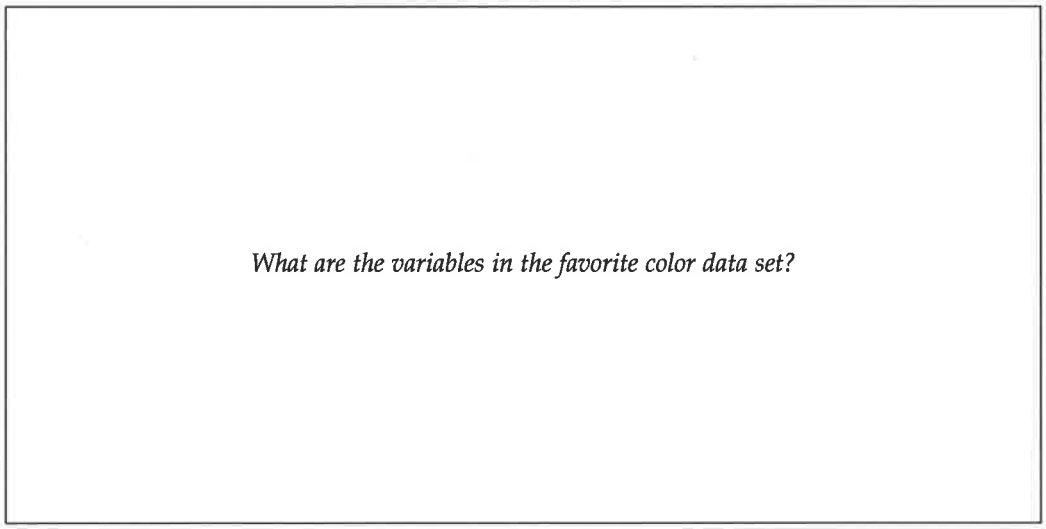
34. Examples of cases: lots of a drug being manufactured, patients under care of a hospital, or stock trades made by a firm.

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)?





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?