SEAN G. CARVER, PH. D.

THE DATA PROFESSOR'S GUIDE TO
BASIC STATISTICS

Draft, July 2017

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 text-book. 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 statistics

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

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

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

25. 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.

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

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

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

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

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

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

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

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

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

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

36. Variables could be dosage of the drug, age and diagnosis of patient, and company, price, and date of trade of stock.
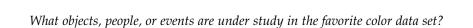
37. What were the variables of the student athlete example?

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

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

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

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

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

43. 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*.

44. 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.

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

46. 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?

47. 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.

48. 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?

49. 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.

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

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

52. 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.

53. 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*.

54. 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.

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

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

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

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

59. 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.

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

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

62. What other options would we have?

Kinds of variables

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

64. 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.

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

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

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

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

69. 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.

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

71. 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?

72. 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?

73. 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*.

74. Interpreted as above, the variable is quantitative.

75. 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?

76. What about average?

What is the average of the values of this variable?

77. 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.

78. 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.

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

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

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

82. 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.

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

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

85. 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?

86. Can you think of other ordinal categorical variables?

87. *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.

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

89. 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?

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

91. 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.

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

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

94. 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.

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

96. 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).

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

98. 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.

99. 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?

Installing Software

100. To follow this text, you need three software packages: R, R-Studio, and StatCrunch.

101. R is undoubtably the most powerful statistical software package available, and its available for free.

102. R-Studio, also free, is a graphical interface to R. Without R-Studio, R is entirely text based.

103. StatCrunch is an alternative to R which is easy to learn.

104. StatCrunch provides a pedagogical tool helpful for learning statistics.

105. Unfortunately, StatCrunch is not particularly useful outside of the classroom.

106. Please install R and R-Studio for your system now. Install R first, then R-Studio, because R-Studio requires R. When you install R, you may be asked to "choose a mirror." A mirror is an alternative download site. Pick one close to you for fastest download. When you install R-Studio, you have the choice of version. Choose the free desktop verision of R-Studio. Follow directions online for installing both R and R-Studio; the directions depend on your operating system, and may change from time to time.

107. You access StatCrunch through your browser. Just about any browser is suitable: Firefox, Chrome, Safari, Internet Explorer, etc.

108. StatCrunch costs money, but if you are enrolled at a university, your institution may have already paid for your subscription. For example, at American University, StatCrunch is available via http://statcrunch.american.edu/ with American Univerisity credentials. If you are not at a university that provides free access, you can purchase a subsription through http://statcrunch.com/. On the other hand, there is no need buy StatCrunch if you just want to use R. R can certainly do everything StatCrunch can, perhaps with just a little more effort. I'll show you how.

Storing Data on a Computer

109. A big part of being able to use R for statistics, is being able to understand R variables.

110. R variables remain similar to the variables defined above for statistics, but they must be extended in the context of computer memory storage.

111. All statistical software packages store data similarly, behind the scenes.

112. They differ in how much detail you need to understand to use them.

113. Unlike R, StatCrunch is similar to Excel in how much detail you need to understand to use it. Maybe even less detail is needed for StatCrunch.

114. When you want to put data into StatCrunch or Excel, you just type it into the spreadsheet on the computer. You don't have to think about what kind of data you are entering, or how it is represented in memory.

115. Behind the scenes, StatCrunch figures out how to store your data in computer memory. The details of how it accomplishes this feat are completely hidden to you.

116. R also tries to hide details from you, but at the same time, it gives you the power to control your data at a low level.

117. Much of the workings of R remains mysterious to new users unless they understand R variables.

118. Even a basic knowledge of R is informed by a basic knowlege of R variables.

119. Unlike with StatCrunch, students using R need to learn about how variables are stored in a digital computer, and what constraints this implementation leads to.

120. We will discuss these contraints now.

121. Above, we defined a *binary categorical variable* as a categorical variable allowing only two possible categories.

122. Binary variables restrict (perhaps unrealistically) their values to two possibilities—such "Democrat" and "Republican", or "Male" and "Female."

123. Binary categorial variables are also just called *binary variables*, especially in the context of computer science.

124. Turns out, all memory in a digital computer is composed of (collections of) binary variables.

125. Why just binary variables?

126. Because computer memory is made of billions of transistors, each of which can be in one of two states: OFF and ON.

127. OFF and ON are the two categories of the binary variables in computer memory. Sometimes these are written as 0 and 1.

128. Current passing through the transistor can either read the state, or change the state, depending on how the current is passed.

129. Each transistor stores one *bit* of data. Bit stands for *b*inary dig*it*. Digits in a digital computer are binary.

130. You may be more familiar with the term *byte*, as in kilobyte, megabyte, gigabyte, terabyte.

131. A single byte equals eight bits. Storing a single byte of data requires 8 transistors. Storing a megabyte requires 8 million transistors, etc.

132. When you want to store data, you ultimately represent the information in bits. This includes numbers and categories, pictures, videos, your favorite songs on iTunes, and any program that will be run on your computer, etc.

133. How do you store data on a computer?

134. Let's start with categories: male and female.

135. Male: 0 and Female: 1. One bit (one transistor) for each patient in your health care data set. If you had 8 thousand patients, you would need 8000 bits, or 1000 bytes, or one kilobyte to store all their genders.

136. In practice, a computer might use more bits than the minimum theoretically needed to store data; but that remains beside the point.

137. As many of you know binary categories do not capture the full range of human gender identity. To account for more diversity, you might include a third option on your survey concerning gender identity: "Male," "Female," and "Other."

138. To store all the data from this more inclusive survey, you would need a second bit for each case. The value of the *pair* of bits would now determine the value stored for the gender identity variable of each patient in your health care data set. For example, Male: 00, and Female: 01, and Other: 10. In this example, the bit pair 11 would never be used. For 8000 patients, using two bits per patient, you would need 16,000 bits for these data, or 2000 bytes, or 2 kilobytes—twice what you needed, before.

139. One bit stores up to two categories. Two bits stores up to 4 categories. Three bits stores up to 8. Four bits stores up to 16. With n bits, you can store 2^n categories. But unless your number of categories is a power of 2, some bit patterns will not be used.

140. In R, such categorical data are stored in a variable called a factor. A factor is actually considered a complex data type, derived from one of 6 simpler types, called atomic classes. (And of course all 6 atomic classes are ultimately patterns of bits.)

141. The six atomic classes in R are logical, integer, numeric, complex, character, and raw.

142. Logical variables are the easiest to understand: they are binary variables, where the categories are interpreted as TRUE or FALSE.

143. What about numbers?

144. In R, there are three atomic classes devoted to numbers: (1) integers, and (2) numeric (called floats in other computer languages) and (3) complex.

145. Numeric is perhaps the most mysterious. They are real numbers— think decimals in scientific notation.

146. Integers are easy to understand. With two bits you can store 4 integers: zero: 00, one: 01, two: 10, three: 11.

147. Two bits are, of course, not enough for most purposes. R uses 32 bits (4 bytes) to store integers.

148. This 4 byte constraint on integers is considered a significant limita-
     tion of the R language. Most other computer languages use at least
     8 bytes (64 bits) for each integer.

149. Because of this limitation, in R, don't try to store an integer bigger
     than 2 billion (in absolute value). However, storing such a big
     number as a numeric/float is OK, though.

150. Floats are stored as decimals in scientific notation, for example
     -23.52 is stored as $-.2352 \times 10^2$. R stores the sign (in this case,
     "negative", using 1 bit), the exponent (in this case, "2", using 11
     bits) and the fraction (in this case, ".2352", using 52 bits). In total,
     64 bits or 8 bytes are used by R for each numeric/float.

151. Originally in computer science, floats spanned only 32 bits. Such
     32-bit floats are pretty rare today, although many graphics cards
     still use 32-bit floats (as graphics doesn't require that much preci-
     sion).

152. Today, 64-bit floats, used in R, are sometimes called "doubles," or
     "double precision."

153. There is such a thing as "quadruple precision", a.k.a."real-16" (for
     16 bytes), a.k.a. 128-bit floats, but that choice is not available in R
     and it is still pretty rare.

154. Using the scheme described above, R can store at least 15 signifi-
     cant digits of precision (that's base-10 digits) in the fraction.

155. In total, R can (only!) store 2^{64} different numeric values/floats.

156. As a fourth atomic class, complex numbers can also be stored in R,
     but we won't consider those.

157. As for atomic classes, so far we have considered logical, integer,
     numeric, and complex.

158. A fifth atomic class is *character*, appropriate for storing the pa-
     tient's name.

159. How would you store the patient's name?

160. You could store each letter as a separate byte: for example A: 0000,
     B: 0001, C: 0010, etc. This scheme is not exactly the encoding that
     is used, but you get the idea.

161. The last atomic class, raw, just stores the data as bit patterns and
     does not interpret them, or allow any specialized operations like
     addition or multiplication. We won't deal with data of that class.

Using R-Studio and R

162. To use R, you will always launch R-Studio, the graphical front-end to R.

163. Launch R-Studio now.

164. What you see when you launch R-Studio differs depending on how it is set up. But if you run it for the first time you should have one window broken into three panes (later it will be four).

165. The left-most pane is the Console, this is where you will type commands into R.

166. The upper-right pane has several tabs. The most important tab is the Environment. The environment will be a list of variables that you have defined. Presumably the Environment will be empty, because you haven't yet defined any variables.

167. One of the other tabs in the upper-right pane is the History, a list of the commands you have typed into R. The History should also be empty, until you type your first command.

168. The lower-right pane also has a number of tabs. Some of the most important of these are: Files, Plots, Packages, and Help.

169. The Packages tab requires explanation. Not every command in R is immediately available from R: for many, you need to install a packages. Installing packages can be done from the Packages tab. There are many thousands of packages available in R.

170. Want to create a new R command that can be used by others? Any programmer can write and contribute packages to the R archive.

171. The archive is called the Comprehensive R Archive Network, or CRAN.

Using R as a calculator

172. Let's say we have three students: Sally, Joe, and Zoe. Each takes an exam. Sally scores a 90, Joe scores and 86, and Zoe scores a 97.

173. Let's say you want to calculate the average of all the scores. Later we will do this with R variables, but for now, lets do it without variables, as you would with a calculator.

174. You probably know the formula already: $(90 + 86 + 97)/3$.

175. This formula gives something called the *mean* of the scores. The mean is one kind of average. We will discuss others, later.

176. Now we type it into R:

```
(90+86+97)/3

## [1] 91
```

177. Considering that I have just included an R formula in this document, a comment is in order about how this textbook is written.

178. This textbook is written in R with the R-package knitr. This package allows me to type commands into my document and have R execute the commands, draw figures, make tables, etc when I create my document. In other words, I just type the commands and it takes care of the rest, putting R's output into the document. And everything is professionally typeset.

179. Why is this feature useful?

180. What if you create professional looking report with statistics and figures, but then your data changes: let's say you add 10 more cases to the data set (e.g. 10 more patients).

181. In the old days, you would have to redo all the calculations, recreate all the figures and tables, and paste them all into your document—hoping you didn't miss something. And you would

have to go through this tedious process every time the data changed, which might be every day!

182. With R-Studio, you can put in the commands into your document to generate the statistics, and create the figures and tables. After you've done that, each time the data changes, you would simply press one button, labeled "Knit," which would re-typeset your document after recomputing the statistics and regenerating the tables and statistics with R.

183. The benefits of such dynamic documents are tremendous. In my opinion, knitr is the most compelling selling point for R.

Exercise: Compute in the R-Console, using operators +, -, *, /, ^ :

$$5^3 + \frac{7 \times 50}{8} - 9$$

Variables in R

184. We have seen how to use R as a calculator without using variables. Now let's repeat this calculation with variables.

```
sally <- 90
joe <- 86
zoe <- 97
(sally+joe+zoe)/3

## [1] 91
```

185. We have defined 3 variables here: sally, joe, and zoe. Each variable stores a single number, the exam score for the named student.

186. Variables always have names, like sally, joe and zoe. Names can be any sequence of letters and numbers, but they can't have special characters or spaces, and the first character must be a letter.

187. The variable name is a handle which we use to access the value they contain later.

188. Why use variables?

189. There are probably dozens of compelling reasons why use variable in R. I'll give just a few.

190. Concerning our discussion of knitr above, suppose the instructor was writing a report to convey the performance of the class.

191. The instructor computes several statistics from the class. Actually there are many statistics she might consider, many of them we will taking about, but for now lets just deal with the average, the minimum, and the maximum.

192. The instructor could direct R to compute these three statistics for the scores $90, 86, 97$, or she could define variables for sally, joe, and zoe, and compute the statistics for the variables.

193. Does it matter? Which do you prefer?

     Option 1:

```
(90+86+97)/3
min(90,86,97)
max(90,86,97)
```

     Option 2:

```
sally <- 90
joe <- 86
zoe <- 97
(sally+joe+zoe)/3
min(sally,joe,zoe)
max(sally,joe,zoe)
```

194. In my opinion, option 2 is a lot easier to read.

195. But an even more compelling reason is that if one of the scores changes—let's say the instructor realizes that she made a mistake in grading or transcribing the scores—you only need to change the code in one obvious place, not in several (and a possibly unknown number of) obscure places. That advantage is huge!

196. Remember there are 6 atomic classes of variables, but we only consider 4: logical, integer, numeric, and character. (We skip complex and raw.)

197. How would we store variables as different classes? Consider

```
# sally is stored as numeric/float/double
sally <- 90
# Joe is stored as an integer
joe <- 86L
# ZoePass is stored as a logical
ZoePass <- TRUE
# ZoeGrade is stored as a character
ZoeGrade <- "A"
```

198. Some explanation is needed. If the number sign (#) appears in the code, the rest of the line is ignored by R. This is useful for adding comments to code.

199. The assignment operator "<-" is used to define variables. In other languages "=" is used as the assignment operator. This works in R, too, however there is an unwritten understanding among R programmers to only use the "<-" operator for assignment, and, even though it is allowed by the language, to never use "=" for assignment. The equal sign is used in other contexts for other things, and R programmers think that this restriction avoids confusion.

200. If a number is assigned to a variable, it will be stored as a numeric/float/double. If you want to store a number as an integer, do as above, with L following the number as in "86L."

201. For logicals TRUE and FALSE have special meaning in R. T and F also work, but only sometimes. If you define a variable with the name T, the letter T will no longer mean TRUE. On the other hand, TRUE and FALSE cannot be reassigned.

202. Enclosing text into quotation marks indicates that you want the text to indicate data stored as character. Two or more characters, such as "A+" would also work here—these are called strings instead of just characters. There is no restrictions on the text that can be saved in strings, although adding quotation marks to the text obviously requires special tricks.

203. There are at least 3 different ways of checking how R stores a particular variable, typeof, class, and mode, with subtle differences.

```
typeof(sally)

## [1] "double"

class(sally)

## [1] "numeric"

mode(sally)

## [1] "numeric"

typeof(joe)

## [1] "integer"

class(joe)

## [1] "integer"

mode(joe)
```

```
## [1] "numeric"

typeof(ZoePass)

## [1] "logical"

class(ZoePass)

## [1] "logical"

mode(ZoePass)

## [1] "logical"

typeof(ZoeGrade)

## [1] "character"

class(ZoeGrade)

## [1] "character"

mode(ZoeGrade)

## [1] "character"
```

Beyond the Datum

204. If we have 10 exam scores, we don't need 10 variables. We can store them in what is called a *vector*.

205. Observe:

```
scores <- c(33,77,50,89,94,78,88,100,98,77)
scores

##  [1]  33  77  50  89  94  78  88 100  98  77
```

206. The 'c' above stands for *combine*. It is the function that creates vectors. You can also create vectors with other atomic types. Here we do it for the logical type:

```
Pass <- c(FALSE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE)
Pass

##  [1] FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

207. There is a better way of figuring out who passed. Let the computer do it for you:

```
Pass2 <- (scores >= 60)
Pass2

##  [1] FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

208. We can create vectors of integers:

```
classof <- c(2017L, 2018L, 2020L, 2017L, 2021L, 2021L, 2019L, 2020L, 2021L, 2020L)
classof

##  [1] 2017 2018 2020 2017 2021 2021 2019 2020 2021 2020
```

209. Or strings:

```
firstname <- c("Sal", "Sy", "Bob", "Jo", "Jane", "Ivan", "Zoe", "Terry", "Bill", "Jill")
firstname

## [1] "Sal"   "Sy"    "Bob"   "Jo"    "Jane" "Ivan" "Zoe"   "Terry"
## [9] "Bill"  "Jill"
```

210. What happens when you create a vector with different types?

```
student1 <- c("Sal", 2017L, FALSE, 33)
student1

## [1] "Sal"   "2017"  "FALSE" "33"
```

211. Notice what happened? A clue is that all components are in quotation marks. What does that mean?

212. When vectors are created with different types of input, all input is coerced into one type. The result is not only strings. It can be other things depending on the types involve. If you are curious, investigate.

213. You can coerce things yourself:

```
as.numeric(FALSE)
## [1] 0
as.numeric(TRUE)
## [1] 1
as.character(TRUE)
## [1] "TRUE"
as.logical(33)
## [1] TRUE
as.integer(33)
## [1] 33
as.numeric("33")  # coercing a string into a number
## [1] 33
as.numeric("Sal") # can't do this coercion
## Warning:  NAs introduced by coercion
## [1] NA
```

214. If you want to put things together of different types, you need to create a list.

```
student1list <- list("Sal", 2017L, FALSE, 33)
student1list

## [[1]]
## [1] "Sal"
##
## [[2]]
## [1] 2017
##
## [[3]]
## [1] FALSE
##
## [[4]]
## [1] 33
```

215. Lists can also include other lists.

216. As an alternative to a spreadsheet, R has something called a dataframe:

```
gradebook <- data.frame(firstname, classof, Pass2, scores)
gradebook

##     firstname classof Pass2 scores
## 1         Sal    2017 FALSE     33
## 2          Sy    2018  TRUE     77
## 3         Bob    2020 FALSE     50
## 4          Jo    2017  TRUE     89
## 5        Jane    2021  TRUE     94
## 6        Ivan    2021  TRUE     78
## 7         Zoe    2019  TRUE     88
## 8       Terry    2020  TRUE    100
## 9        Bill    2021  TRUE     98
## 10       Jill    2020  TRUE     77
```

217. In a dataframe, each column contains data of the same atomic type. In this case, this statement is obvious because we created it with vectors. That said, there are other ways of creating dataframes, and no matter how you create them, if the types don't match, coercion takes place.

218. The last complex data type we will discuss the *factor*. We men-
     tioned earlier that factors are appropriate for categorical data.

219. Let's add genders to our student list.

```
genders <- c("Female", "Male", "Male", "Female", "Female", "Male", "Female", "Male", "Male", "Female")
genders.factor <- factor(genders)
genders.factor

##  [1] Female Male   Male   Female Female Male   Female Male   Male   Female
## Levels: Female Male
```

220. Both the period and the underscore are characters that can be used
     in variable names, just like any other letter.

221. Now the 10 students genders are stored as binary variables (be-
     cause there are only two categories) instead of strings of characters.
     This saves memory, and makes computation more efficient.

222. The "Levels" are also stored as "Female" and "Male." But these
     character strings are only stored once.

223. There is also a way to store ordinal categorical variables. Consider

     [resume]

```
birthmonth <- c("Feb", "May", "Mar", "Sep", "Dec", "Apr", "Mar", "Jun", "Jul", "May")
months=c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")
birthmonth.factor <- factor(birthmonth, levels=months, ordered=TRUE)
birthmonth.factor

##  [1] Feb May Mar Sep Dec Apr Mar Jun Jul May
## 12 Levels: Jan < Feb < Mar < Apr < May < Jun < Jul < Aug < Sep < ... < Dec
```

224. Now we add the two new columns to our dataframe.

```
gradebook <- cbind(gradebook, genders.factor, birthmonth.factor)
gradebook

##    firstname classof Pass2 scores genders.factor birthmonth.factor
## 1        Sal    2017 FALSE     33         Female               Feb
## 2         Sy    2018  TRUE     77           Male               May
## 3        Bob    2020 FALSE     50           Male               Mar
## 4         Jo    2017  TRUE     89         Female               Sep
## 5       Jane    2021  TRUE     94         Female               Dec
```

```
## 6       Ivan   2021  TRUE    78          Male              Apr
## 7        Zoe   2019  TRUE    88        Female              Mar
## 8      Terry   2020  TRUE   100          Male              Jun
## 9       Bill   2021  TRUE    98          Male              Jul
## 10      Jill   2020  TRUE    77        Female              May
```

225. Actually, the command (actually it's called a function) data.frame(),
     will by default, convert a vector of strings into a factor, before its
     put into the dataframe. This behavior can be overwritten in several
     ways. Ask Google.

Indexing Data

226. What if you want the scores of just the third student?

```
scores[3]

## [1] 50
```

Distributions

227. The concept of a *distribution* is absolutely central in probability and statistics.

228. In an advanced book, you will get a mathematical definition of a distribution.

229. We have to settle for the following (which while imprecise, conveys the idea):

230. The *distribution* of a variable tells us (1) what values the variable takes and (2) how often the variable takes these values.

231. The best way to visualize a distribution is with a graph.

232. The kinds of graphs we draw for categorical variables is different from the kinds of graphs we draw for quantitative variables.

233. For categorical variables we draw pie charts and/or bar graphs.

234. For quantitative variables we draw stem plots and histograms.

235. Let's graph the favorite color variable of our favorite color data set.

236. Let's graph the categorical variables of the diamonds data set.

237. Homework 1.

238. Stem plots and homework 2.

239. Histograms and the call center data set.

240. Let's graph (some of the) quantitative variables of the diamonds data set.

241. Homework 3.

Exploratory data analysis

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

243. 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.

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

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

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

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

248. 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.

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

250. 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...

251. 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.

252. 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.

253. 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.

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

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

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

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

258. 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.

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

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

Mean

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

262. 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.

263. 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.

264. Most students are already familiar with the formula:

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

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

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

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

267. 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.

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

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

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

270. 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.

271. 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.

272. 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.

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

Sampling

274. 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.

275. 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.

276. 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.

277. 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.

278. 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.

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

280. 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.

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

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

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

284. The 6 students comprise the *sample*.

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

286. 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.

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

288. 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.

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

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

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

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

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

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

294. 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:

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

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

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

297. 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*.

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

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

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

301. 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.

302. Therefore sampling rarely gives the right answer.

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

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

305. 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.)

306. 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.

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

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

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

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

311. 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!

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

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

314. 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.

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

316. 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.

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

318. 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.

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

320. 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.

321. Let's explore simple random samples:

322. 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.

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

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

325. 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.

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

327. 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!

328. 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.

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

Counting samples

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

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

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

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

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

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

335. 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: $4 \times 3 \times 2 \times 1$.

336. 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.

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

338. 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.

339. 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.

340. 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.

341. 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.

342. 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 chang-
ing 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.

343. We are not quite done, yet, because, as mentioned, there are actu-
ally 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.

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

345. 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:

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

346. 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

$$\vdots$$

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

348. 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.

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

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

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

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

353. 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).

354. 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.)

355. 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

356. 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?

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

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

358. 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*.

359. 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!

360. 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.

361. Let's unpack the formulas.

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

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

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

365. 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.

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

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

368. 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.

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

370. 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.

371. 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.

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

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

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

375. 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.

376. 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).

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

378. 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.

379. 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.

380. 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.

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

382. 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.

383. 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 $\bar{x}$.

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

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

386. 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.

387. 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.

388. 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.

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

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

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

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

393. 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.

394. 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.

395. 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.

396. 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.

397. 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.

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

399. 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.

400. 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.

Random phenomena and probability

401. If someone were to tell you that a phenomenon was random, what would you know about the phenomenon?

402. First of all, you would know that the outcome of the phenomenon was uncertain.

403. For example, you if you flip a coin, you might get heads and you might get tails.

404. There is no way to know in advance if each outcome is going to be heads or tails.

405. But notice that if you flip a coin many times, and the coin is fair (one side not weighted more heavily than the other), then about half the time you get heads and about half the time you get tails.

406. A coin flip is a random phenomenon.

407. A *random phenomenon* is one for which individual outcomes are uncertain but there is, nonetheless, a regular distribution of outcomes in a large number of repetitions of the phenomenon.

408. Is it possible for there not to be a regular distribution of outcomes in a large number of repetitions?

409. This situation is never studied, and I think it is fair to say that if you lack a regular distribution of outcomes in a large number of "repetitions," then you are not really repeating the same phenomenon. Something is changing.

410. Now that we know that there is a regular distribution of outcomes in a large number of repetitions, we can define probability.

411. The *probability* of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very large series of repetitions of the phenomenon.

412. The concept of probability can be made more precise with the concept of a limit. The probability is the limit of the proportion

of times the outcome would occur as the number of repetitions of the phenomenon goes to infinity. Limits are something studied in calculus, which you may or may not be familiar with.

Probability models

413. A *probability model* describes a random phenomenon in the language of mathematics, specifically set theory.

414. When we describe a random phenomenon, there are two things we care about: (1) what various outcomes are possible, and (2) how likely are the various outcomes?

415. For a fair coin flip, there are two possible outcomes: heads and tails. Each happens with probability 1/2.

416. For a fair 6-sided die toss, there are 6 possible outcomes, usually labeled with dots: ⚀, ⚁, ⚂, ⚃, ⚄, and ⚅. Each happens with probability 1/6.

417. Note how probabilities arise.

418. How do we formalize these statements into probability models?

419. We use set theory.

420. A set is a collection of objects.

421. The statement "a set is a collection of objects" is not a mathematical definition of the concept of a "set." It is a description of the concept in terms of a synonym: "a collection."

422. So, what is the precise mathematical definition of a "set?"

423. Funny you should ask, because "set" happens to be only one of two concepts in all of mathematics that do not have a definition. (The other is "element of a set.")

424. You may remember from Geometry in high school that most concepts studied had definitions in terms of simpler concepts.

425. But there had to be some concepts that were the simplest possible.

426. In high school geometry there were three simplest concepts. They were the so-called undefined terms.

427. Those simplest concepts were (1) point, (2) line, and (3) plane.

428. In the rest of mathematics the simplest concepts are "set" and "element of a set".

429. These concepts have no definitions in mathematics.

430. Numbers, functions, and other concepts are defined in terms of simpler notions that eventually lead to "set" and "element of set."

431. In early 20th century, the mathematician Hilbert defined "point," "line," and "plane," in terms of simpler notions of "set" and "element of a set."

432. But "set" and "element of a set" is as far as you can go. You have to start somewhere.

433. Remember that even though there are no definitions for point, line and plane in high school geometry, these concepts are pinned down by five postulates (also called axioms).

434. Set theory has its own set of axioms, (called the Zermelo-Fraenkel-Choice, or ZFC, Axioms), which precisely pin down the notion of "set" and "element of a set."

435. The ZFC axioms are too technical to discuss here, but feel free to Google.

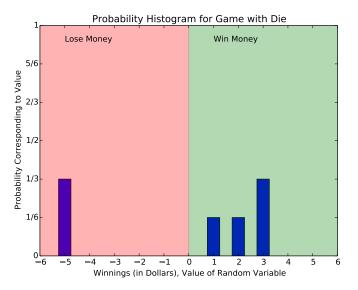436. At this level it is better to explain set and element of a set with examples.

Probability mass versus probability density

437. Consider the following game of chance: we roll a fair six-sided die. If the die shows ⚀, we win 1 dollar. If the die shows ⚁, we win 2 dollars. If the die shows ⚂ or ⚃, we win 3 dollars. And if the die shows ⚄ or ⚅, we lose 5 dollars.

438. Let the random variable X be our winnings for one game, which will be negative if we lose money.

439. One question we will answer in a future lesson: should we play the game?

440. For now, we are going to discuss the differences between discrete and continuous random variables: specifically a discrete random variable has probability mass at its possible values, whereas a continuous random variable has probability density.

441. Remember, a discrete random variable is one having a finite number of possible values.

442. And a continuous random variable is one having a continuous range (hence infinite number) of possible values.

443. For our die game, there are 4 possible values for the random variable, X, our winnings for one game: -$5, $1, $2, and $3. Thus X is discrete.

444. Let's create a probability table for this game. Remember, a probability table for a discrete random variable lists the finite number of possible values for the random variable, together with their respective probabilities.

Values	-$5	$1	$2	$3
Probabilities	1/3	1/6	1/6	1/3

445. A probability of 1/3 (or 2/6) corresponds to the value -$5, and $3, because, in each case, two out of the six die faces lead to that outcome. (E.g. for -$5, it's ⚄ and ⚅.) Likewise, a probability of 1/6 corresponds to winning $1 or $2.

446. We can graph the information in the probability table as a proba-
bility histogram. See below.

447. A probability histogram is a "theoretical histogram" for a dis-
crete random variable, just like a density curve is a "theoretical
histogram" for a continuous random variable.

448. What is a "theoretical histogram?"

449. If you (1) repeat the random phenomenon under consideration
many times, and (2) each time compute the random variable
from the outcome of the random phenomenon, then (3) put those
numbers into a column of a spreadsheet and finally (4) use those
numbers to generate a histogram, *then* that data histogram *should
look like* the theoretical histogram.

450. What does "should look like" mean?

451. Random variation will lead to deviations between the data his-
togram and the theoretical histogram—the match won't be perfect.
But the more data you generate, and the smaller your bin width
(for the continuous case), the smaller the deviations are likely to
be, and, with increasing certainty, your data histogram will will
look more and more like the theoretical histogram.

452. For any histogram, the values of the random variable (e.g. win-
nings, in our running example), lie on the horizontal axis, but the
vertical scale can differ.

453. The vertical scale for a discrete probability histogram is relative
frequency, the theoretical proportion of data points with that have
that value.

454. Theoretical histograms for continuous random variables (density curves) use a different vertical scale: density.

455. Remember, the density scale used for continuous random variables is the relative frequency of a very small bin around the given value, divided by the width of that very small bin.

456. The calculus notions of limit and derivative make the definition of density precise.

457. Why don't we use relative frequency for continuous random variable histograms?

458. And why don't we use density for discrete random variables?

459. Answer: first, because, as will be explained below, the relative frequency of all individual points for a continuous random variable is zero—useless for describing the random variable!

460. Second, because, as will also be explained below, the density of a discrete random variable is infinite at all possible values of the random variable—also useless!

461. Let's develop these ideas a little.

462. Theoretically, what is the relative frequency for each bin (proportion of observations falling inside) for a discrete random variable's histogram?

463. Answer: it is the sum of the probabilities of the possible values of X falling within the bin.

464. In our die game example, consider a bin that contains both $2 and $3, and no other possible values for X (that is, it doesn't contain $1 nor, (obviously), -$5). The proportion of observations falling within this bin (relative frequency) is 1/2 (or, equivalently, 3/6) because there are 3 possible die faces out of 6 (specifically, ⚃, ⚄, and ⚅) which lead to values of the random variable inside this rather wide bin, spanning both $2 and $3).

465. On the other hand, if a bin contains just one possible value for the random variable, its relative frequency is just the probability of obtaining that one possible value.

466. In the die game example, the relative frequency of a small bin containing $3, but no other possible values of X, is 1/3, *no matter how small the bin.*

467. This result means that the relative frequency of the value $3 is 1/3.

468. The situation is different for continuous random variables, where the proportion of observations in a bin *does* depend on bin width (unless the density is zero).

469. Consider a uniform random variable. As we discussed previously, the uniform distribution is a continuous distribution in which all values are possible within a continuous range between two extremes (e.g. 0 and 1), and moreover all of these values remain equally likely.

470. "Equally likely" ends up meaning equal density.

471. Remember, the density curve for the uniform distribution is a rectangular "box" between its extremes (e.g. 0 and 1) and zero beyond these extremes.

472. Remember, the area under any given density curve is 1, so the closer together the extremes fall, the higher the box.

473. Recall, the proportion of observations that land within an interval is the area under the density curve above that interval.

474. In calculus, this quantity has a name: we call it the definite integral of the density curve over that interval.

475. So the relative frequency of a bin of a histogram for data coming from the uniform distribution, contained within the extremes is

    relative frequency, uniform, bin within extremes = width of bin × height of box.

476. Notice, if we shrink the width of a bin by 1/2, the relative frequency of the bin goes down by 1/2.

477. And if we shrink the box to a single point, the relative frequency goes to zero.

478. Put another way, and said more generally, the relative frequency of any *individual* value within a continuous distribution is zero.

479. If we tried to use a relative frequency curve to convey the same information that a density curve conveys we would fail because the relative frequency curve would be everywhere zero—completely uninformative!

480. Of course a real histogram for continuous data can use relative frequency as its vertical scale and it won't be everywhere zero!

481. Why? Because the bin width won't ever be zero. With a non-zero bin width, you get non-zero relative frequencies.

482. So why don't we use relative frequency as our scale?

483. We can, but we have the following problem: For a continuous random variable, the height of the relative frequency histogram will depend on the bin width!

484. Look again at the equation for relative frequency, uniform.

485. Only if we *divide* the relative frequency by the bin width do we get a value that is independent of the bin width.

486. Remember: dividing the relative frequency by the bin width gives us density.

487. A relative frequency histogram is a poor description of a continuous distribution, because, for continuous random variables, the relative frequency depends on an arbitrary choice of bin width.

488. Unfortunately, for non-uniform, but still continuous distributions, (where the density curve is not a constant height), the relative frequency will depend not just on the bin width, but also on the endpoints of the bin—and so will density.

489. What you have to do instead is divide by the relative frequency by the bin width, then shrink the bin to a point, which is made precise with the calculus notion of a limit.

490. For people who know calculus, I'll fill you in. (Don't worry if you don't know calculus, just skip this bullet point.) Consider the function

$$f(x) = \text{Probability}(\text{Random Variable} \leq x)$$

Let h be the width of the bin, then the proportion of observations within a bin that starts at x is $f(x + h) - f(x)$. Divide by the bin width h and take the limit as h goes to ∞ and you get the definition of the derivative for f. In an advanced book you, will see that the definition of a continuous random variable is one for which this function f has a derivative. Its derivative is the density of the random variable. But for discrete random variables, this derivative does not exist (the limit is infinite).

491. Density, defined this way, does not depend on such arbitrary choices, making it a better description of a distribution.

492. Calculus aside, why don't we use density for discrete random variables?

493. Remember, for our example, a bin containing the value $3 but no other possible value of winnings had a relative frequency of 1/3:

$$\text{relative frequency, die game, \$3} = \frac{1}{3}$$

494. In this case, we get the same value, independent of bin width, *without dividing by bin width*.

495. If we *do divide by bin width* we get a density estimate that *grows as the bin width shrinks*.

496. Why grow? Because the numerator stays constant and we divide by smaller and smaller numbers. Dividing by 0.1 is the same as multiplying by 10. Dividing by 0.01 is the same as multiplying by 100, etc.

497. We might say the density is infinite.

498. The analogy to mass and density is helpful.

499. As an analogy, elements like lead and gold have different densities, even if they have the same mass.

500. An ounce of pure lead has more volume than an ounce of pure gold.

501. We define density as mass divided by volume; gold has higher density than lead, regardless of volume.

502. The analog of mass in probability is called probability mass.

503. All probability distributions have total mass 1.

504. Probability mass is spread out along the number line, either along a continuum for continuous distributions, or among isolated points for a discrete distribution.

505. Probability density is probability mass divided by length (no volume, because there is only one dimension here)

506. Think: Density for bin equals the proportion of observations in bin (mass) divided by bin width (length).

507. For discrete distributions, probability mass is concentrated on individual points, not spread out across a continuum,

508. When concentrated on individual points the density at those points is infinite, whereas, when mass is spread out along a continuum, the mass at any given point is zero.

509. A physicist might call a mass with infinite density a "black hole."

Sampling distributions

510. Remember the example we used in the *Sampling* chapter?

511. We had a sample of 6 students out of a population of 1000 and the 6 students had the following scores on an exam: 90, 92, 94, 96, 98, and 100.

512. In that example, the sample mean was $\bar{x} = (90 + 92 + 94 + 96 + 98 + 100)/6 = 95$.

513. In that chapter, we were interested in an estimate, based on these six students, of the population mean.

514. Remember, if we happened to have the scores for all 1000 students in the class, we would have been able to exactly compute the population mean adding all 1000 scores, then dividing that sum by 1000.

515. In the situation of the *Sampling* lesson, we only had the scores of the six students in our sample, not all 1000 students in the class, so we used the sample mean, 95, as an estimate of the population mean.

516. In other words, we used a *statistic* to estimate a *parameter*.

517. Remember, a *statistic* is a characteristic of a *sample*, like the sample mean.

518. Remember, a *parameter* is a characteristic of a *population*, like the population mean.

519. There is only one population mean: the sum of the 1000 scores divided by 1000. Parameters *always* have just one possible value.

520. Statistics generally have lots of possible values, because samples can generally be drawn in many different ways.

521. Indeed, there is a different sample mean for each different sample that can be drawn. (However sometimes values repeat.)

522. In our example, we counted almost 1.37 quadrillion different possible samples.

523. Remember what the distribution of a variable tells us? It tells us: what values the variable takes and how often it takes those values.

524. The distribution of a statistic, also called its *sampling distribution* tells us what values the statistic takes and how often it takes those values—across all its possible values, corresponding to all the different possible samples.

525. For our population, what values does the sample mean take, and how often does it take them?

526. The answer to the question should not yet be obvious (and indeed, we need more information to answer it), but it should already be clear that the answer interests us.

Spread in sampling distributions

527. For samples chosen randomly (specifically for samples chosen as what we called *simple random samples*), the sample mean is an *unbiased estimate* of the population mean.

528. Remember the definition of an unbiased estimate? It is was an estimate where the mean of all possible estimates equals the correct value of the quantity being estimated.

529. In this case, the mean of all possible sample means (for all possible samples that can be drawn) equals the population mean.

530. We desire unbiased estimates, but it also matters how much variation there is in the estimate across the samples.

531. Depending on both the sample size and the variability in the population, it could happen that if we repeat the phenomenon of drawing samples then using those samples to calculate sample means, the computed sample means could vary considerably from each other.

532. Sampling distributions (the distribution of a statistic computed from a sample) can be described as any other distribution: center, and spread, modes and skewness, mean and standard deviation, median, quartiles, percentiles, etc.

533. We know that the mean of the sample means is the population mean.

534. Therefore, the population mean is the mean of our sampling distribution. (This value measures the *center* of the sampling distribution for sample mean).

535. We could have also used the median or mode to measure the center of the sampling distribution, however we typically we use the mean to measure the center of sampling distributions.

536. What about the *spread* in the distribution?

537. We will use the standard deviation to measure the spread of the sampling distribution. We could also use quartiles or percentiles, but these alternative are less common.

538. If the standard deviation is relatively large, then we can expect widely different estimates of the population mean for different samples—even if they are simple random samples.

539. Thus, if the standard deviation of the sampling distribution is large, we might not want to trust any one particular estimate, even if the estimates are unbiased. (I.e. center is in the correct place but the spread is all over the place).

The sample mean as a discrete random variable

540. In a previous lesson, we defined a *statistic* as a characteristic of a sample.

541. A statistic can take on different values depending on how the sample is drawn.

542. What values the statistic takes and how often it takes them is the sampling distribution of the statistic.

543. The sampling distribution of a statistic has center (e.g. mean) and spread (e.g. standard deviation), as well as, perhaps, any other characteristics of a distribution, such as peaks, gaps, symmetry and/or skewness.

544. Now we go a little farther. But maybe this statement won't come as a surprise: *a statistic is a random variable.*

545. Let's develop this idea a little.

546. In the *Sampling* chapter, we said that there were exactly 1,368,173,298,991,500 different ways of selecting a sample of 6 students from a population of 1000.

547. We also identified this monstrous number, exactly, as "1000 choose 6," which could also be written as:

$$\binom{1000}{6}.$$

548. A simple random sample gives an equal chance to each possible way of selecting the sample (in this case, each sample has a 1 in approximately 1.37 quadrillion chance of being selected).

549. This sampling process is a random phenomenon, just like flipping a coin, or rolling a die.

550. Remember, random phenomena had *outcomes*. The set of all outcomes was called the *sample space* and subsets of the sample space were called *events*.

551. The *sample space* for sampling is the set of all possible samples.

552. I imagine the name *sample space* comes from sampling, but statisticians now apply it more generally to the set of outcomes of any random phenomenon.

553. In our case, the sample space is the set of approximately 1.37 quadrillion possible six-person samples drawn from the population of 1000 people.

554. Remember, we defined a *random variable* as a function that assigned a number to each outcome in a sample space.

555. Just like we assigned a number to each possible outcome of tossing a coin three times (we used the number of heads), let's assign a number to each possible outcome of the sampling phenomenon.

556. In other words, let's define a random variable mapping each of the 1.37 quadrillion possible samples to a number.

557. What number? A number that characterizes the sample: i.e. a statistic.

558. How about this idea? To each of the 1.37 quadrillion possible samples we assign the sum of the scores of the individuals in that same sample, divided by the sample size.

559. In other words, to each sample, we assign its sample mean.

560. Because there are just under 1.37 quadrillion outcomes in our running example, there are at most 1.37 quadrillion possible values for this random variable.

561. Actually, there will probably be somewhat less than 1.37 quadrillion values for the random variable because more than one sample can share the same sample mean.

562. Certainly this coincidence occurs if more than one student shares the same exam score—because there will be different samples with the same scores.

563. But sample means can coincide in other ways, as well—specifically, if different sets of scores average to the same value.

564. On the other hand, in theory it is possible that all 1.37 quadrillion sample means have different values.

565. But in our case, if the 1000 scores are all integers between 0 and 100, samples means will certainly repeat.

566. Either way, the sample mean has a finite number of possible values.

567. Remember, a discrete random variable is one with a finite number of possible values.

568. For sampling from a population of N individuals, as long as N is finite, the sample mean is a discrete random variable, with a finite number of values.

569. Though finite, the number of possible values for the sample mean may be astronomically large, even if N isn't.

The sample mean as a continuous random variable

570. Remember the exam score example: we know the exam scores of a sample of 6 students from a population of 1000 students in a large multisection basic statistics class.

571. If we had access to all 1000 exam scores, we could put them into a histogram, just like we could for any other quantitative variable.

572. Likewise, if we had access to all 1,368,173,298,991,500 six-person sample means, we could put them into histogram, again, just like we could for any other quantitative variable.

573. To be able to access 1.37 quadrillion non-integer numbers on a computer (typically, stored as double precision floats, requiring 8 bytes each) you would need about 11,000 one-terabyte hard drives. Google probably has such resources, but hardly anyone else does.

574. Let's stick with just the 1000 exam scores, for now.

575. As hinted at earlier, exam scores often follow a Normal distribution (but not always).

576. Following a Normal distribution implies that the QQ-plot for the 1000 exam scores approximates a line and the histogram for the exam scores approximates a bell curve.

577. If the exam score distribution is truly Normal, then any discrepancies between these approximations and the predictions of the Normal distribution stem from random, well-characterized, fluctuations.

578. However, if a random variable truly has a Normal distribution, it must be a continuous random variable with a continuous range of possible values—not a discrete random variable, with only 1000 possible values.

579. Indeed, choosing a small enough bin width for the histogram of the exam scores will inevitably reveal the gaps that must fall between the exam scores actually achieved by students.

580. If the exam scores are all integers, then gaps will appear at all non-integer values, as well as any integer value not attained as one of the 1000 students' exam scores.

581. Being discrete, the population distribution for the exam scores is not Normal.

582. Nonetheless, statisticians sometimes find it convenient to use continuous distributions to approximate discrete ones.

583. How?

584. Approximating a discrete random variable with a continuous random variable involves filling in the gaps of the histogram (in a reasonable way).

585. Going back to the material of a previous lesson, filling in the gaps amounts to smearing the probability mass across the continuum, so that mass within larger bins remains roughly the same, but the mass no longer remains concentrated at individual points.

586. From the discrete probability histogram, we get a continuous density curve for the exam scores.

587. The histograms of the discrete and continuous distributions should look the same *except at very small scales (i.e. except at very small bin widths)*.

588. Different software packages might accomplish this smearing (sometimes called smoothing) differently. We won't discuss the algorithms.

589. But if the QQ-plot for 1000 exam scores reasonably approximates a line (or equivalently, if the histogram for the exam scores reasonably approximates a bell curve) then a bell curve might be a reasonable way to fill in the holes.

590. In this case, we would approximate the discrete exam score distribution with a Normal distribution. But which Normal distribution should we use?

591. Normal distributions differ only in their mean and standard deviation.

592. If we had access to all 1000 scores, we might compute their population mean and population standard deviation to use as parameters for the continuous Normal distribution approximation.

593. If we didn't have access to all 1000 scores, we would have to settle for estimates based on samples. But hopefully, in this case, our sample size would be substantially greater than 6.

594. If the QQ plot didn't follow a line, we might use another distribution, with a density curve that was not a bell curve.

595. In theory, we could use any density curve to approximate the distribution of a discrete random variable, however we would want to use one that looked like the histogram, plotted with density on the vertical scale and using an appropriate bin width.

596. Remember, all continuous distributions have density curves that describe them.

597. Remember, any curve above or on the horizontal axis, enclosing an area of 1, is a valid density curve. The most commonly used distributions have names (such as Normal or uniform), but every different valid density curve gives a different distribution, so most distributions do not have names.

598. Let's get back to sampling.

599. We approximate drawing a six-person sample from a finite (1000-person) population as drawing six samples from an infinite population whose exam scores are a continuous random variable.

600. Instead of 1000 exam scores attainable within a sample of our population (possibly with repeating values), with the continuous approximation there are now infinitely many possible values for each exam score—each score can take on any value within a whole continuous range.

601. So, instead of a number (possibly as large as 1.37 quadrillion, but definitely finite) of different attainable six-student sample means, we now have infinitely many.

602. Have we gained anything?

603. Yes, because it is easier to deal with a density curve, that you can approximate with an equation, than it is to deal with up to 1.37 quadrillion different possible values for the sample mean.

604. Most calculations with density curves involve calculus.

605. Even with its density curve completely and correctly specified, saying that we draw six numbers from a distribution does not completely specify the random phenomenon of sampling. We need to say something more.

606. We need to also say how the different numbers *relate* to one another.

607. For instance, if you choose one student with a low exam score, are you more or less likely to pick the next student with another low exam score?

608. For sampling from continuous distributions, the usual answer is "previous selections have no effect on subsequent selections, and vice-versa."

609. This statement means the selections are *independent*.

610. We usually describe continuous sampling as drawing *n* people *independently* from the same continuous distribution.

611. But surprisingly, for discrete sampling from finite populations, this perhaps innocuous sounding statement does not hold!

612. And the smaller the sample size, the more egregious the difference.

613. Why? For independent selections, we said that previous selections did not affect the distribution of subsequent selections.

614. Consider the following example: suppose there was just one poor soul who did not study at all for the exam scored a zero while everyone else scored a perfect 100.

615. At first there is a 1 in 1000 chance of picking the zero, but once you have picked the slacker for your sample, you know you can never pick him again for the same sample, so in subsequent selections the probability of picking the slacker goes down to zero.

616. In other words, for finite sampling, previous selections *do* affect the probability distribution of subsequent selections—meaning the selections are *not independent*.

617. What about smaller sample sizes? Consider the following fact: for a similar three person population the probability of picking the slacker would go down from $1/3$, at first, to 0, after you select him, a considerably greater difference in probability than going from $1/1000$, at first, to zero after picking the slacker.

618. On the other hand, if you *don't* pick the slacker at first, the probability of picking him goes up in subsequent selections, and substantially so, with small population sizes. This result also demonstrates non-independence.

Sampling distribution versus population distribution

619. We highlight a distinction between the *sampling distribution* and the *population distribution*.

620. The *sampling distribution* tells you what the values a statistic takes (e.g. sample mean of test scores) and how often it takes them—across all possible *samples* from a population.

621. The *population distribution* tells you what values a variable takes (e.g. test scores) and how often it takes them—across all possible *members of a population, not samples*.

622. Actually, for the same population, there are different sampling distributions, depending on the sample size.

623. In our case, the sample size was 6, but in other scenarios the sample size could be 1, 2, 3, etc.—any integer, n, between 1 and and the population size, N, inclusive.

624. Question: How many ways are there to draw a one-person sample—i.e. how many samples have $n = 1$?

625. Answer: There are exactly as many ways to select one person from a population as there are people in the population. In other words, there are N ways, (where N is the population size), to draw a one-person sample.

626. In our exam score example, because the class size is $N = 1000$, there are 1000 one-person samples.

627. For any sample size n, to get the sample mean, you add up n numbers (e.g. n exam scores) and divide by n.

628. If $n = 1$, you add up just one number (e.g. one exam score) and divide by one.

629. So in our example, when $n = 1$, the sample mean is just the score of the one selected student.

630. So, what is the sampling distribution for the sample mean of one-person samples?

631. What values does the sample-size-one sample mean take, and how often does it take those values?

632. Can you see that sample-size-one sampling distribution *for the sample mean of a variable* is the same as population distribution for the same variable?

633. Specifically, the mean of "the sample-size-one sample means" is the population mean, the sum of 1000 test scores, divided by 1000.

634. While this coincidence holds for the sample mean, it does not necessarily hold for other statistics.

635. For example, mean of "the sample-size one sample standard deviations" is *not* the same as the standard deviation across the whole population.

636. Similarly, the standard deviation of "the sample-size one sample standard deviations" is *not* the same as the standard deviation across the whole population.

637. Indeed, with Bessel's correction, the sample standard deviation for a sample-size-one sample doesn't even exist, because with Bessel's correction you divide by $n - 1$ instead of n in the formula for standard deviation.

$$s_{n-1} = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2} \text{ with Bessel's correction.}$$

638. To reiterate, if $n = 1$, you can't calculate the sample standard deviation with Bessel's correction, because you can't divide by zero.

639. But the problem lies beyond just Bessel's correction.

640. Let's remove Bessel's correction and see what we get—we get another statistic that characterizes the sample.

641. Remember, with or without Bessel's correction, the sample standard deviation is a biased estimate of the population standard deviation, but the bias is somewhat worse (especially for small sample sizes) without Bessel's correction. Both statistics are used in practice, but the one with Bessel's correction is used far more commonly.

642. Remember, without Bessel's correction, the standard deviation is the square root of the mean of the squares of the deviations from the mean.

$$s_n = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} \text{ without Bessel's correction.}$$

643. With sample size one, there is just one data point, so there is just one deviation from the mean.

644. But then the data point must coincide with the mean, so the lone deviation from the mean must be zero.

645. So to find the sample standard deviation without Bessel's correction, we take the square root of the mean of the squares of...of just this one zero—which always equals zero.

646. The mean of "the sample-size-one sample standard deviations (without Bessel's correction)" is the mean of 1000 zeros—which equals zero.

647. At the same time, the standard deviation of "the sample-size-one sample standard deviations (without Bessel's correction)" is the standard deviation of 1000 zeros—which also equals zero.

648. At the same time, the population standard deviation is almost never zero. (It can only be zero if all N scores are the same.)

Probability tables for discrete sampling

649. In theory, we can describe finite sampling with a probability table, which, in our 6 chosen from 1000 example, would be way to large to write down.

650. Remember that probability tables have two rows: the first for possible values of the random variable, and the second for their respective probabilities.

651. Actually, sometimes it is useful to write a *redundant probability table* (my terminology), with three rows: one for outcomes, one for values, and one for probabilities.

652. In a redundant probability table, instead columns for each possible *value* of the random variable, there are columns for each *outcome* in the sample space. Thus, there could be more than one column with the same value for the random variable.

653. In the "flip three coins and count heads" example, there will be separate columns in the redundant probability table for TTH, THT, and HTT, even though the value of the random variable (number of heads) is, in each case, 1.

654. In the first row of the redundant probability table for "flip three coins and count heads" go the eight outcomes: from TTT to HHH.

655. In the second row of the redundant probability table go the values of the random variable, with repeats: in this case, number of heads from 0 to 3.

656. In the third row of the redundant probability table go the probabilities of each outcome, 8 repetitions of 1/8; not 1/8, 3/8, 3/8, and 1/8.

657. For the exam score sampling example, in the first row of the redundant probability table go all 1.37 quadrillion possible 6-student samples.

658. In the second row, go the sample means, possibly with repeats.

659. In the third row, go the probability of each sample. With a simple random sample, each of these probabilities is the exact number $1/1,368,173,298,991,500$.

660. It is impossible to know the probabilities in the non-redundant probability table without knowing how many samples get each score. For this knowledge, you would need to know the exam scores of all 1000 students.

661. Why? Consider the case of the sample mean of 95. At least one sample, the one we originally drew has a sample mean of 95.

662. Now consider: we know that the scores of the 6 students in our sample range from 90 to 100.

663. What if every other student in the class got a 0 on the exam? (That's 994 zeros: ouch!)

664. Then every other sample replaces at least one, and maybe more, maybe all 6, scores with a zero, so every other sample mean is less that the 95 we originally obtained.

665. In this case, there would be only 1 sample with a sample mean of 95 and the probability of getting 95 for a sample mean would be 1 in 1.37 quadrillion.

666. On the other hand, what if every other student (994 of them) scored a 95 on the exam? Then, all the many samples that do not contain one of our six students has a sample mean of 95, and that scenario is not the only way to get a 95 for the sample mean.

667. In this case, the probability of getting a 95 for a sample mean is much closer to 1 than it is to 0.

668. If you know the probabilities, and especially if they are equal, it can sometimes be easier to work with a redundant probability table.

Tests of significance

Introduction

669. In statistics, we are often faced with making decisions based on partial information—often samples, sometimes small samples, from a larger population.

670. For instance, a drug company must decide to put a new drug on the market. Should they?

671. When considering the benefits of a new drug to treat a specific disease, the population of interest is every person in the world with that disease.

672. Consider the following response variables: how much does the patient improve, and for how long do they survive?

673. The only way to definitively answer the drug company's question is to measure the response variables for all possible patients (all people with the disease), for all treatments (including placebo and control) while holding all other factors constant.

674. Obviously, no one can conduct such a study.

675. So we are left with random samples.

Running example

676. We are going to work with our running example, already familiar to you.

677. In this example, this year, a hypothetical university has implemented a novel teaching method, and the university wants to compare students' performance this year with their performance last year.

678. Using their comparison, the university will make a decision whether or not to continue using the new method and to push its implementation at other universities.

679. The university plans to use the scores on one exam—the same exam given this year and last—to compare performance.

680. They have framed the following question: have students scores improved, on average, this year, over last year? They have decided that any improvement, on average, no matter how small, would justify the change in teaching methods.

681. Of course, it is very easy to answer this question if you have the grade books for both classes: just compute the population mean for this year, and compare it to the population mean for last year.

682. But it gets difficult when you only have samples of the students grades to work with.

683. Of course, a university would probably have access to all the grades—but in other domains, working with samples is imperative.

684. We are going to restrict ourselves to samples, but remember: you are trying to answer a question about the population means. Specifically is the population mean higher this year?

685. Comparing sample means is not the same as comparing population means. Conclusions based on samples can differ, depending on the luck of the draw (i.e. the luck of how the sample is drawn).

686. The population means give the correct answer to the question, so conclusions based on samples may, or may not, be in error.

687. Therefore, with sample data and without population data, we can't even give a definitive answer to the question we posed.

688. So we revise our question: "Based on sampling, do we have sufficient evidence to conclude that the population mean has increased this year?"

689. We will show that we can answer this question, but the "sufficient evidence," our conclusion may still be wrong—we can still make an error.

690. How much evidence is sufficient? You design a test of significance to make the probability of error acceptable.

One sample or two?

691. Three situations are possible: (1) you have samples for both years, (2) you have a sample for one year and population data for the other, (3) or you have population data for both years.

692. First, you might be stuck with samples for both years. In this case, you would do a *two sample test*.

693. Second, you might have all the scores for one year, but only a sample for the other. In this case, you could do a *one sample test*.

694. With a one sample test you have more information than with a two-sample test, so assuming your information is correct, you will be more likely to make the right decision, all other things being equal.

695. Of course, if you had all the information—all grades for both classes—you wouldn't even have to do a test of significance.

696. That said, *even if the university had all the grades, they might decide to use a two sample test!*

697. Why? They might decide to think of each class as a sample from a larger population.

698. What population? The population of all students in the world, or maybe the population of all students who might conceivably enroll in their classes.

699. Unfortunately, a university's classes are not random samples from any larger population, so any conclusions drawn from such a two sample test are hard to interpret based on rigorous statistics.

The data

700. Let's continue with our example of the sample of 6 students (with exam scores 90, 92, 94, 96, 98, 100), from a class of 1000 students total.

701. It is important to realize that this sample was not, in fact, drawn randomly from any population. If it had been, we should be very surprised at the regular spacing. Still, this sample is familiar to you from earlier lessons, and for our purposes it will suit us just fine.

702. So assume that these scores come from a simple random sample from this year's class, and they are the only scores we know of for this year.

703. But, we have access to the entire grade book for last year: all 1032 exam scores in a spreadsheet (a few more people took the class last year). We will call this class the *regular* class because, later, we identify a different group of students as *honors*.

704. To simulate these data, I used a program I wrote in the software package R, that implemented modified distribution that is not available in many simpler software packages, like StatCrunch.

705. You won't be able to simulate these data without R and the program I wrote, but you can download the grade book (see below).

706. I am not going to tell you, until later, about the program that generated the data. These details are irrelevant for solving the problem and would just distract you. Clearly, this information would not even exist if you had real scores from a real class with real students.

707. There are too many scores to print, but here are the first 10 scores out of 1032 (the rest can be downloaded, see below):

```
##  [1] 80 85 81 86 80 70 80 80 85 82
```

708. I use the software package R to generate the data, then put the first 10 generated numbers into this document, then save all 1032 generated numbers to an Excel file. All of this is done automatically, each time I create the PDF you are reading. You might be confused by the "## [1]" before the list starts. That is R's way of saying that the whole list of 1032 begins with the first number shown. This embellishment can be helpful when the list spans multiple lines—each line will start with its own marker noting the position of the line in the list. I have decided to leave these markers in the document, throughout.

709. Let's look at a different group of students taking a similar class last year—all honors students. There were only 210 students in this group but that's still too many to print. Here are the first 10 scores (the rest can be downloaded, see below):

```
##  [1] 95 89 92 92 95 91 86 89 93 99
```

710. You should download the data sets now (regular one with 1032 scores, the honors one with 210 scores). You can find these data with the rest of the data for the course.

711. After you download the data you should load the data into the statistical software package that you use for class.

712. Now, answer the questions in the textboxes below. My answers follow each question.

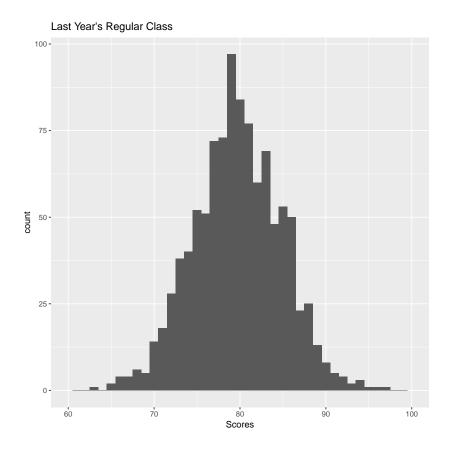> *What is the first thing you should do with new data?*
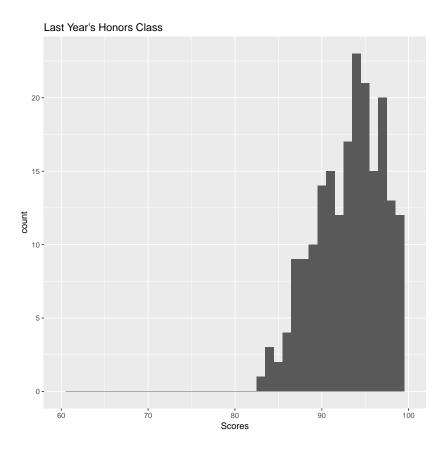
713. Answer: You should graph your data! Do it now.

> *How did you create your graphs?*

714. There is only one column here, "Scores," so we should ask: is the variable quantitative or categorical?

715. If you try to think about this question deeply you might be confused if you notice that all exam scores are integers. You should have noticed this when you loaded the data. So you might tempted to try to justify calling the scores variable ordinal categorical.

716. The problem is, we have already decided to base our decision on population means—the mean is meaningless for categorical variables—so we have actually already made our decision.

717. So we are left with two choices: stemplots and histograms. There is too much data for a stemplot. So we are left with a histogram.

718. Actually there are other graphs we have learned about for a single quantitative variable: box plots and QQ plots. These may be helpful later, especially the QQ plot.

719. But the histogram will usually convey the most information for a variety of distributions, so it is a good choice for the first plot. (But, of course, doing several graphs is always good, too.)

720. What bin width did you choose? Did you pick the default that your software package provided? The default is seldom the best choice. Start with it, then tweak it.

721. Because the scores are integers, we should consider a bin width of 1. That way each score gets its own bin. Try it and you will see

that are enough students, and the scores cluster together in the
right way, to make these choice helpful.

722. Let's take a look at the histogram for both classes:

Last Year's Regular Class

Last Year's Honors Class

723. After looking at these histograms, you might ask yourself, what would have happened if we choose to consider the scores as an ordinal categorical variable?

724. With an ordinal categorical variable, we would be stuck with pie charts and bar graphs. After working with the possibilities, we would find that a bar graph, with the natural ordering, conveys the most information. (Pie charts and Pareto orderings would be useless for this variable.)

725. Interesting: a bar graph with the natural ordering would look almost the same as the above histogram, except that it would have spacing between the bars—spacing that would also appear in the histogram if we chose a smaller bin width. What choices best convey the information in the distribution? What do you prefer?

The main question

726. Moving to the main question for today's lesson: *Based on the sample of 6 students from this year's class* (with exam scores 90, 92, 94, 96, 98, 100), *do we have enough evidence to conclude that students in this year's class have scored higher, on average, than students in last year's classes?*

727. Actually, we can identify two separate questions: the first compares this year scores to last year's *regular* student scores, and the second compares this year's scores to last year's *honors* student scores.

728. Note that, while you have all scores, for all students in both of last year's classes, I never gave you the scores for the other 994 students in this year's class. I don't have these scores. I never ran a simulation to create them. They are irrelevant for the question we are addressing, which was framed as "Based on [just] the sample of 6 students from this year's class, do we have sufficient evidence to conclude that students are doing better, on average, this year?"

729. With a *random* sample, even though we have *only 6 out of 1000 scores*, we will see that it is possible that we can have *very strong* evidence that students are doing better, on average, this year.

730. The histogram for last year's regular class already suggests that we will find such strong evidence for this class.

731. Notice that very few students scored in the 90's, and no student (out of 1032) scored higher than a 97.

732. Now look at the sample of students from this year: all 6 students scored above 90, and two of the 6 students scored above last year's maximum with scores of 98 and 100.

733. So clearly, we don't have the same scores this year as last year, but the question remains, could all the scores be drawn independently from the same distribution, this year and last?

734. We can't answer this question definitively with our data: we might see the same data either way, depending on the luck of the draw.

735. But the following might be clear already: *if all the scores, this year and last, were drawn from the same distribution, then our sample would be very unusual and unexpected,* provided, of course, it was drawn as a simple random sample.

736. Please note the importance of the *simple random sample* stipulation: if the professor her favorite 6 students as the sample of 6, we might not find our data unexpected—even if the scores of all students in both classes were drawn from the same distribution.

737. We interpret unusual and unexpected data as evidence in support of the statement the change we are looking for has occurred. But the data must unusual and unexpected if no change has occurred, and they are less unusual and less unexpected the change we are looking for has occurred.

Test statistic

738. We need to quantify "unusual and unexpected" for a sample.

739. We are going to use one number, called a *test statistic*.

740. The test statistic, as its name suggests, is a *statistic*—a number that characterizes a *sample*—so it must be a function of the six scores in our sample.

741. We are going to use the sample mean as our test statistic.

742. We interpret greater sample means, as stronger our evidence that scores have improved, on average, this year.

743. Our sample mean was 95. Is that high enough?

744. We are going to choose a *critical value for the test statistic.*

745. In our case, if our sample mean is greater than our critical value, we are going to deem the evidence sufficient to conclude that student's are doing better this year, on average. Otherwise, we will deem the evidence insufficient for this purpose.

746. Our sample mean was 95. But how should we pick our critical value? Is it greater than or less than our sample mean?

Hypotheses

747. We want to pick the critical value for the test statistic to control the probability of error.

748. But to control the probability of error, we must be precise about the statements we want to say are true or false.

749. This is done with what is called *hypotheses*. Tests of significance are sometimes called *hypothesis tests*.

750. The first hypothesis is called the *null hypothesis*. It is usually the statement that no effect or no change has occurred.

751. The second hypothesis is called the *alternative hypothesis*. It is usually the statement that the change you are looking for has occurred.

752. You should write these hypotheses in terms of the parameters of the population, but I'll use English.

753. Our null hypothesis: "the six exam scores from this year's sample came from the same distribution as the one last year's scores came from."

754. Our alternative hypothesis: "the six exam scores from this year's sample came from a distribution whose population mean is greater than the one that last year's scores came from."

755. Always write hypotheses in terms of parameters. Never write hypotheses in terms of statistics.

756. The truth or falsity of hypotheses should never depend on how the sample is drawn.

757. Notice that the null hypothesis is specific: it says that this year's sample's scores came from the *one* distribution that is the same as last year's.

758. Notice that the alternative hypothesis is not specific: it says that this year's sample's scores came from *any* distribution that has a greater population mean than last year's population mean.

759. Statisticians sometimes prefer the term *simple* over *specific*. The null hypothesis is simple whereas the alternative hypothesis is not.

760. The null hypothesis is often abbreviated H_0, whereas the alternative hypothesis is often abbreviated as H_a.
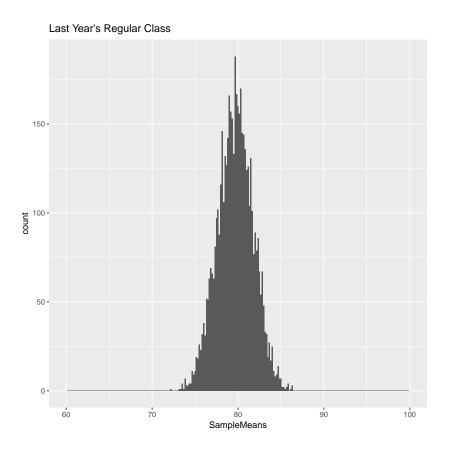
The distribution of the test statistic

761. Do we first need to know the distributions mentioned in our hypotheses to test our hypotheses?

762. Remember, we are going to use a test statistic (in this case, sample mean) to weigh the evidence for or against our hypotheses.

763. In our example, if the sample mean is unusually and unexpectedly high, we will deem the evidence sufficient to conclude that the scores have improved, on average, this year.

764. How will we know how high is high enough?

765. To answer this question, we need to know: What values does the test statistic take and how often does it take its values? Only then will we know what values are unusual.

766. In other words, we need to know the sampling distribution for the test statistic.

767. We derived our test statistic from our sample of this year's data. We only have one value for the test statistic, so how can we find its distribution?

768. We can simplify things with an assumption: we will assume one of
our hypotheses is true.

769. The alternative hypothesis is not good choice here, because there
are lots of possibilities for the sampling distribution under alterna-
tive hypothesis—the alternative hypothesis is not simple.

770. On the other hand, there is only one possibility for the sampling
distribution of the test statistic under the null hypothesis. In our
case, it is the same as the distribution of the sample means for last
year's class—and we know all the scores from last year's class.

771. As it turns out, finding sampling distribution of the test statistic
under the null hypothesis is good enough to be able to perform
tests of significance—and as you will soon see, it is often much
easier than finding the distribution of the scores.

772. But how do we find the distribution of last year's sample means?

773. We can simulate the distribution with software (you should al-
ready know how): Draw 5000 six-scores samples from last year's
regular class scores, calculate the sample mean for each.

774. Remember, there more than a quadrillion possible samples of size
6 that can be drawn from a population of last year's class size—we
are generating just 5000 of them.

Follow the instructions above for drawing sample means and put the results into one column.

Now plot a histogram of the sample means computed above.

Last Year's Regular Class

775. Did you tweak the bin width? I used a bin width of 1/6 for the above plot because every sample mean (being a sum of 6 integers divided by 6) is a multiple of 1/6. With this choice, every possible sample mean gets its own bin. I would not recommend making such a choice with a somewhat larger sample size.

776. Compare your histogram of sample means plotted above with the histogram for the last year's scores—and remember we are working with the regular class.

777. Some things to notice: First, the center of the sample means is almost the same as the center of the scores.

778. Second, the spread of the sample means is somewhat less than the spread of the scores.

779. Both the distribution for the sample means and the distribution for the scores seem close to Normal. However, we can say that the distribution for the sample means more closely approximates a Normal distribution, because the gaps in the histogram are smaller: width 1/6 versus 1.
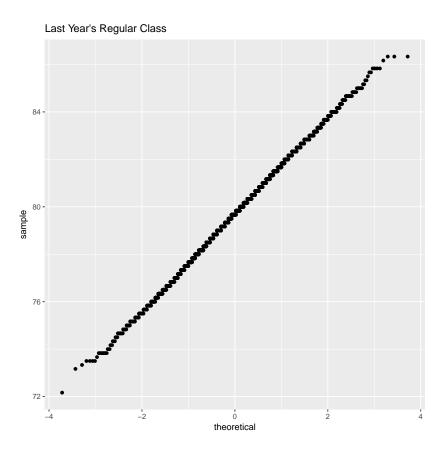
Compute summary statistics for the regular sample means: mean and standard deviation.

Compute summary statistics for the regular scores: mean and standard deviation.

Find formulas relating the parameters that correspond to the summary statistics, above.

To check your work, compare your answers for the previous three problems.

Check the Normality of the regular sample means. Use a graph different from a histogram.

Last Year's Regular Class

780. The QQ-plot looks pretty straight so we can say that our test statistic (sample mean) closely follows a Normal distribution.

781. Indeed, whenever our test statistic is the sample mean, we should *expect* the distribution of the test statistic to follow a Normal distribution.

782. Why? Remember, the Central Limit Theorem tells us that (aside from some technicalities that we don't usually need to worry about) sample means closely follow a Normal distribution whenever the sample size is large enough.

783. If you are interested in the technicalities, Google "Central Limit Theorem."

784. Our sample size, 6, happens to be small enough that we might indeed need to be concerned if there were strong outliers or substantial skewness.

785. But the QQ-plot of sample means confirms that, in our case, these concerns are unfounded.

786. Consider the importance of the Central Limit Theorem: no matter what distribution our data come from, we can use the sample

mean as a test statistic and expect its distribution to be Normal—at least as long as the sample mean is large enough.

787. Can you see the huge importance of the Central Limit Theorem?

788. And it turns out in many cases, like in ours, that the sample size doesn't even need to be especially large.

789. We've made a QQ-plot of *sample means*, now make a QQ-plot of *scores*.

Make a QQ-Plot for last year's regular scores.

Make a QQ-Plot for last year's honors scores.

790. Observe that the honors scores are less normal than the regular scores—there is some skewness to the honors scores.

Make a QQ-Plot for last year's honors sample means.

791. Observe that despite the increased skewness of the honors scores, the honors sample means still closely follow a Normal distribution—even though the sample size was only 6.

792. Had there been stronger skewness or outliers in the scores distri-bution, we might have needed a larger sample size to accurately use a Normal distribution to approximate the distribution of sample means.

793. There are methods for dealing with non-standard distributions of the test statistic, however, if the QQ-Plot of sample means reveals

a problem with using a Normal distribution as an approximation, the easiest solution is usually to up your sample size.

794. Now repeat the calculations done above for the honors class.

Compute summary statistics for the honors sample means: mean and standard deviation.

Compute summary statistics for the honors scores: mean and standard deviation.

Find formulas relating the parameters that correspond to the summary statistics, above.

To check your work, compare your answers for the previous three problems.

---

795. The work you performed above answers the question: what distribution does the test statistic follow?

796. We established that the test statistic follows a Normal distribution, but remember the Normal distribution requires parameters.

797. The parameters of the Normal distribution are *mean* and *standard deviation*.

798. You have already found these parameters—in two ways.

799. First, you sampled 5000 sample means (out of over a quadrillion) and computed the sample mean and sample standard deviation (summary statistics) of these 5000 sample means.

800. Second, you computed the mean and standard deviation of the entire population of last year's scores. You probably used the same *summary statistics* to do make this calculation, but you should interpret the results returned as the population parameters for last year's scores.

801. The formulas you found above relate these two results. Your answers should have been:

$$\mu_{\text{sample means}} = \mu_{\text{scores}}$$
$$\sigma_{\text{sample means}} = \frac{\sigma_{\text{scores}}}{\sqrt{n}}$$

802. Using these equations you can find the parameters for the distribution of the test statistic without sampling the sample means. You just need to know the population mean and the population standard deviation for the scores from last year—which we do know, or can calculate, because we know all the scores.

803. The equations above are exact whereas the sampling method is prone to sampling error.

804. You should definitely know the equations above, but also remember the sampling error can be made as small as you want with a large enough sample size.

805. What was our sample size for sampling the sample means?

806. The relevant sample size for sampling sample means above is 5000, even though the sample size for each individual sample mean is only 6. Yes, this gets confusing.

807. Note that the sample size for sampling the sample means is constrained only by your computational power, not by your ability to collect data. If you know all of last year's scores, which we do, you can essentially make the sampling error (for sampling the sample means) as small as you want.

808. We have found the parameters of the test statistic's distribution two different ways.

809. Now that we know the distribution of the test statistic and its parameters—for both of last year's classes—we can decide upon a critical value for the test statistic that controls the probability of error.

Type I and type II errors

810. We can make two types of errors: a false positive, and false negative.

811. In our example, a false positive means we say students *are* doing better, on average, this year, when in fact they are *not*.

812. Likewise, a false negative means we say students are *not* doing better, on average, this year, when in fact they *are*.

813. False positives are called type I errors.

814. False negatives are called type II errors.

815. In terms of hypotheses, a type I error means we *reject* the null hypothesis, when in fact it is true.

816. A type II error means we *fail to reject* the null hypothesis when in fact it is false.

817. Statisticians traditionally speak of errors in terms of the null hypothesis, although it would be possible to talk about errors in terms of the alternative hypothesis.

Controlling the probability of errors

818. Let's compare the one sample mean from this year's sample with the population means from last year's classes. Remember the sample of scores for this year was (90, 92, 94, 96, 98, 100), so its sample mean was 95.

Compare 95 with the population mean for last year's regular class.

Compare 95 with the population mean for last year's honors class.

819. In each case, the sample mean for this year's class is *greater* than the population mean for last year's classes.

820. Remember we used the sample mean as an estimate of the population mean. If we make this estimate for this year's class, our estimate will be greater than the known population mean for both of last year's class.

821. So should we say students are doing better this year? Better than both the regular students and the honors students?

822. The statement that the sample mean from this year's class is greater than the population mean of last year's scores is, of course, evidence that students are doing better this year.

823. But from a statistical point of view, that statement alone, in and of itself, is *very weak evidence* that students are improving.

824. Consider this: what if the distribution of test scores for this year was, in fact, *the same* as the distribution of one of last year's classes? Put another way, what if our null hypothesis was correct? What is the probability that we would make an error based on one sample mean being greater than the population mean?

825. If the null hypothesis is correct, we can only make a type I error. Our error cannot be a False Negative because in this case, Negative is not False.

826. So what is the probability of a type I error?

827. We can answer this question with our sampling distribution.

828. Using your sampling distributions for last year, *and assuming the distribution hasn't changed this year*, compute the probability that a sample mean drawn randomly from this year's class will be greater than the population mean for last year's honors class? This is the probability of a type I error.
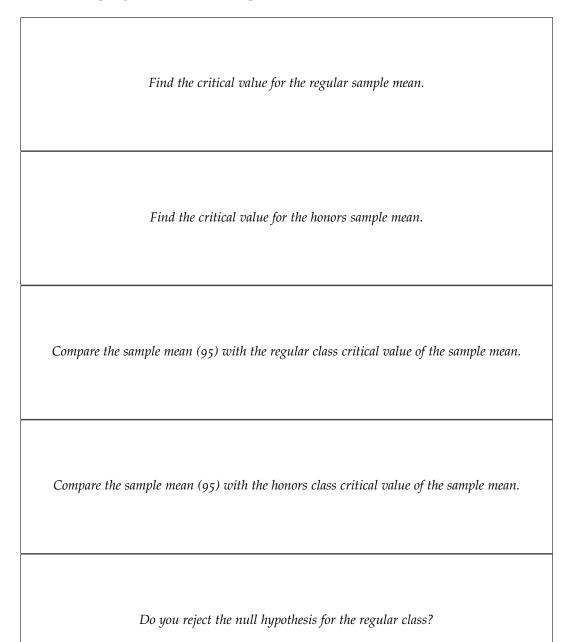
Compute the probability mentioned above.

829. Having trouble? Remember our sampling distribution is approximately Normal and its mean is the same as the population mean last year. But the population mean last year is the same as this

year's sampling mean. So we need to know, what's the probability of observing a statistic (the sample mean) greater than the mean of its sampling distribution.

830. With any symmetric distribution, such as the Normal distribution, the mean equals the median, and with all continuous distributions, the probability of landing above the median is 0.5.

831. So the probability of a type I error, if we reject the null hypothesis whenever the sample mean for this year is greater than the population mean for last year is 0.5—for most applications this probability is unacceptably high.

832. For most application we want better control on type I errors.

833. Tests of significance are designed to control for the probability of a type I error.

834. Type II errors are also important, but they are harder to pin down because when the alternative hypothesis is true there can be many distributions for the test statistic (the alternative hypothesis is not simple).

835. The probability of a type I error is denoted α and is called the *level of significance of the test*.

836. Let's give the name *weak test* to the test that deems student's performance better than last year if the sample mean for this year is greater than the population mean for last year. The name is appropriate because with very weak evidence it deems an improvement.

837. You found the level of significance for the weak test to be 0.5.

838. For comparison, traditionally statisticians choose the level of significance as 0.05.

839. Statisticians are starting to move away from the $\alpha = 0.05$ tradition as they recognize that the best α is a trade off between the costs of type I and type II errors. These costs are different for different problems.

840. I should point out that type II errors are usually controlled by adjusting the sample size. The higher the sample size, the smaller the probability of a type II error, all other things being equal, including the probability of a type I error. But without changing the sample size, as type II errors become less likely, type I errors become more likely, and vice-versa.

841. Let's design different tests to determine if this year's students are doing better—tests that have the traditional level of significance.

842. We look for evidence *against* the null hypothesis. If we succeed, we reject the null hypothesis in favor of the alternative hypothesis. If we reject the null hypothesis, we have found an effect, and our results are *significant*.

843. Remember, we say that if the test statistic is greater that a the critical value then we deem our evidence sufficient to conclude that scores have improved, on average.

844. A result that the test statistic exceeds critical value is called *significant*. This is where the name *test of significance* comes from.

845. How do we choose the critical value? If our test statistic is $\bar{x}$, then this critical value will usually be denoted $\bar{x}^*$.

846. Answer: we want to choose $\bar{x}^*$ so the probability of the type I error is our desired level of significance, in this case 0.05.

847. How do we do that?

848. Remember a type I error can only occur when the null hypothesis is true.

849. When the null hypothesis is true, a type I error occurs when the sample mean exceeds the critical value.

850. Why? Because then we erroneously reject the null hypothesis, by deeming the evidence sufficient to conclude (incorrectly) that the students are doing better this year, on average.

851. It is very important to realize that *even if the null hypothesis is true* it is possible to draw a random sample for which the sample mean exceeds the critical value of the test statistic. When this happens you incorrectly reject the null hypothesis (a type I error).

852. To reiterate, the probability of type I error is α, the level of significance of the test. You pick α. Your choice determines $\bar{x}^*$.

853. How are we going to pick $\bar{x}^*$?

854. We want to pick $\bar{x}^*$ such that the probability that the sample mean is bigger than $\bar{x}^*$ is equal to the level of significance.

855. You have already found the distribution for the sample mean under the assumption that the distribution has not changed from last year.

856. Find $\bar{x}^*$ for each test (regular and honors). Hint: use the functionality of your software that allows you to compute percentiles corresponding to values and values corresponding to percentiles of the relevant distribution.

857. Hint use the Normal calculator in StatCrunch, put the parameters of the sampling distribution in and put in the α.

Find the critical value for the regular sample mean.

Find the critical value for the honors sample mean.

Compare the sample mean (95) with the regular class critical value of the sample mean.

Compare the sample mean (95) with the honors class critical value of the sample mean.

Do you reject the null hypothesis for the regular class?

Do you reject the null hypothesis for the honors class?

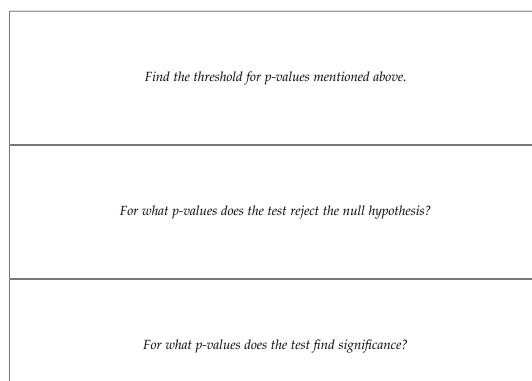*For the regular class, do you have a significant result?*

For the honors class, do you have a significant result?

P-values

858. You have now done tests of significance by comparing the test statistic with the critical value for the test statistic.

859. However, tests of significance are usually performed *without* computing the critical value of the test statistic explicitly, unlike what you did above.

860. Instead of the critical value for the test statistic, something called the *p-value* is computed. Tests with p-values are completely equivalent to tests with critical values, however p-values are usually easier to interpret than critical values.

861. What is a p-value?

862. At least for the tests we have done here, the p-value is the probability *assuming the null hypothesis is true* of drawing a six-person sample whose sample mean is *greater* than the sample mean actually observed in the data (which, in this case, greater than 95).

863. The greater the sample mean (the one actually observed in the data, but not necessarily 95), the stronger the evidence that the test

scores for this year are better, on average, than the test scores from last year.

864. The greater the sample mean, less the probability *assuming the null hypothesis is true* of drawing another sample whose mean is even *greater* than our original sample mean.

865. The probability of drawing a sample whose sample mean is greater than 95 is the p-value for our sample.

866. The p-value quantifies the strength of the evidence against the null hypothesis.

867. What threshold for the p-value determines significance? To answer this question, consider what is the p-value for a sample mean that coincides with critical value of the sample mean. The answers to the next three questions immediately below should be the same for both tests under consideration.

Find the threshold for p-values mentioned above.

For what p-values does the test reject the null hypothesis?

For what p-values does the test find significance?

868. What is the p-value assessing the evidence provided by our sample of from this year's exam scores (90, 92, 94, 96, 98, 100) that students are scoring the same on average (null hypothesis) or better on average (alternative hypothesis) than the students in last year's *regular* class?
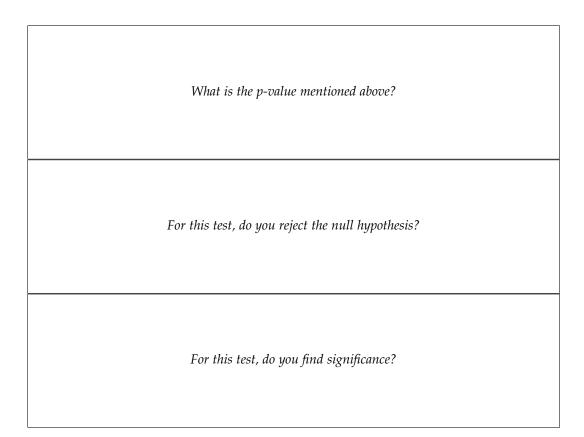
869. You can calculate p-values with the Normal calculator in a manner
complementary with the manner of calculating critical values.

What is the p-value mentioned above?

For this test, do you reject the null hypothesis?

For this test, do you find significance?

870. What is the p-value assessing the evidence provided by our sam-
ple of from this year's exam scores (90, 92, 94, 96, 98, 100) that
students are scoring the same on average (null hypothesis) or bet-
ter on average (alternative hypothesis) than the students in last
year's *honors* class?

What is the p-value mentioned above?

For this test, do you reject the null hypothesis?

For this test, do you find significance?

871. What is the p-value if the sample mean for this year's exam scores is actually equal to the population mean for last year's exam scores (instead of being 95, as assumed above). The answer should be the same whether or not you consider the regular class or honors class, although the population mean is different in each case.

What is the p-value mentioned above?

For this test, do you reject the null hypothesis?

For this test, do you find significance?

872. I have heard that most non-statisticians who nevertheless use tests of significance in their research do not understand what a p-value is.

873. I think it's a disgrace, and it is really a failure of most basic statistics courses, like this one. Many of you will use and study statistics in other courses but won't take another class from a statistician. If you don't understand p-values and tests of significance now, you may never.

874. The reason I bother to bring this up is that the misconception of many people may actually help you understand p-values.

875. If you ask non-statisticians to explain p-values, many will tell you, if they think they know, that "the p-value is the probability that the null hypothesis is true." (This is incorrect.)

876. This misconception is compelling for the following reasons:

877. First, the p-value actually *is* a probability, just not the probability that the null hypothesis is correct. In our example, it is the probability, *assuming the null hypothesis is true*, of drawing a six-person sample whose sample mean is greater than 95 (the sample mean seen in our data).

878. Second, being a probability it acts like a probability: specifically, it is a number between 0 and 1, inclusive.

879. Finally, we can interpret the p-value in a way that does suggest that the misconception is, to the contrary, accurate.

880. Specifically, if the p-value is low, the null hypothesis "probably isn't correct" and if the p-value is high the null hypothesis "may indeed be correct." This interpretation, as stated here, is, indeed valid, and is the way you should interpret it.

881. The problem is: how would we compute the p-value, if we think, incorrectly, that it is the probability that the null hypothesis is true? How would we even make sense of this statement?

882. Consider this: although we only have scores for the six students in our sample, the test was given to all 1000 students, and although we don't know all 1000 scores, the exams have been graded and all students have their scores.

883. If we had access to all 1000 scores we could definitively answer the question, without a shadow of a doubt: are student doing better, on average, this year than last.

884. If the distributions are truely the same, the null hypothesis is correct, regardless of what data we collect

885. On the other hand, if the population mean this year is greater than the population mean last year, even by just a little bit, then the correct answer is that you should reject the null hypothesis in favor of the alternative hypothesis. The null hypothesis is false, again, regardless of what data we collect.

886. It should be pointed out that if the population mean this year is actually *lower* than population mean last year (i.e. if the students are doing *worse* this year not better), then technically the null hypothesis is not correct, but because of the way we framed the alternative hypothesis (students are doing better) our p-values will likely be higher than if the null hypothesis is true and we will be even less likely than the level of significance to find a significant result.

887. Note that the truth or falseness of the null hypothesis has everything to do with the populations and nothing to do with the sample.

888. But the probability of the null hypothesis is true can only be interpreted as the proportion of samples (in the discrete distribution) for which the null hypothesis is true.

889. The null hypothesis either is or isn't true, independent of what sample we draw.

890. Therefore the null hypothesis either is true for *all* 1.37 quadrillion samples or for *none* of these samples.

891. In other words the probability that the null hypothesis is true either is 1 or it is 0.

892. What's worse we have no way of knowing from just the data in the sample whether the probability is 1 or 0. If we had this information we would not have to use a test.

893. All that said, go ahead and remember the incorrect statement that the p-value is the probability that the null hypothesis is true.

894. The incorrect statement will at least give you the right intuition. Specifically, you will remember that when the p-value is low, the null hypothesis probably isn't correct and when the p-value is high, the null hypothesis may indeed be correct.

895. There is a reason I say "when the p-value is high, the null hypothesis may indeed be correct," rather than "when the p-value is high the null hypothesis *probably* is correct."

896. The reason for the linguistic gymnastics is that even if the p-value is high, the population mean for this year may indeed be greater than the population mean for last year: the null hypothesis is false.

897. This situation is likely to arise if both the difference between the population means is small, and the sample size is small. In this case, you tend need to gather more data to have compelling evidence that the population means are different.

898. When a p-value is high there is not that the probability of the null hypothesis is high—its that you lack evidence to refute the null hypothesis.

899. This is the main problem with the incorrect statement "the p-value is the probability that the null hypothesis is correct."

900. If you remember the incorrect statement as a mnemonic for intuition, remember the statement is incorrect and why, and also remember the following statement which gives better intuition: "The p-value measures *the strength of the evidence against the null hypothesis*. The lower the p-value, the stronger the evidence."

Big Picture Highlights

What follows below lists some of the main themes of the semester. The list should not be considered a complete list of topics that need to be studied. Rather the list should be seen as an attempt to convey the scope of what we covered throughout the semester.

901. The first half of the semester concerned ways to **describe** data; the second concerned **inference**.

902. Describing data was called **exploratory data analysis**.

903. Exploratory data analysis involved (1) making **graphs**, and (2) deriving **summaries**.

904. Graphs were used to show the **distribution** of variables.

905. Pie charts and bar graphs were used to show the distributions of single **categorical variables**.

906. Stemplots and histograms were used to show the distributions of single **quantitative variables**. Later we introduced box plots for this purpose.

907. To display the **relationship between two** quantitative variables, scatterplots were used.

908. For summaries, we used **mean, median, and modes** to describe the center of the distributions of single quantitative variables.

909. We used **standard deviation, quartiles, and percentiles** to show the spread of single quantitative variables.

910. We used **counts and proportions** for categorical variables

911. And we used **correlation and regression** to describe relationships between pairs of quantitative variables.

912. In the second half of the semester we considered the alternative to describing data: **inference**.

913. For inference, we draw conclusions about **parameters** of a **population** based on **statistics** from a **sample**.

914. Our first foray into inference was to derive point **estimates**.

915. The **sample mean** was an estimate of the population mean for quantitative variables.

916. The **sample standard** deviation was an estimate of the population standard deviation for quantitative variables.

917. The **sample proportion** was used an estimate of the population proportion for categorical variables.

918. For means and proportions, the estimates were **unbiased**.

919. For standard deviation, we could not get around the fact that the estimate was **biased**, though **Bessel's correction** improved matters some.

920. The next thing we did was to look at **sampling distributions** for our statistics.

921. Studying sampling distributions required us to develop **probability theory**.

922. We derived formulas for the mean and standard deviation of the sampling distribution of the sample mean, sample counts and sample proportions (for different kinds of variables).

923. The shape of sampling distribution for the sample mean was given by the **Central Limit Theorem**—approximately Normal if the sample size is large enough (and often it didn't need to be that large).

924. And we tested hypotheses about parameters with **tests of signficance**. There were a number of choices that needed to be made here.

925. We extended point estimates to **confidence intervals**.