Distributions

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

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

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

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

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

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

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

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

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

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

111. Homework 1.

112. Stem plots and homework 2.

113. Histograms and the call center data set.

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

115. Homework 3.

Exploratory data analysis

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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