Student's Solution Manual

William B. Craine III
Kimberly F. Smith
Mahinda Samarakoon

Stats: Data and Models
Canadian Edition

Richard D. De Veaux
Williams College

Paul F. Velleman
Cornell University

David E. Bock
Cornell University

Augustin M. Vukov
University of Toronto

Augustine C.M. Wong
York University

Pearson Canada
Toronto
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2</td>
<td>Data</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Displaying and Describing Categorical Data</td>
<td>4</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Displaying and Summarizing Quantitative Data</td>
<td>15</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Understanding and Comparing Distributions</td>
<td>28</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>The Standard Deviation as a Ruler and the Normal Model</td>
<td>43</td>
</tr>
<tr>
<td>Review of Part I</td>
<td>Exploring and Understanding Data</td>
<td>57</td>
</tr>
<tr>
<td>Chapter 7</td>
<td>Scatterplots, Association, and Correlation</td>
<td>68</td>
</tr>
<tr>
<td>Chapter 8</td>
<td>Linear Regression</td>
<td>77</td>
</tr>
<tr>
<td>Chapter 9</td>
<td>Regression Wisdom</td>
<td>91</td>
</tr>
<tr>
<td>Chapter 10</td>
<td>Re-Expressing Data: Get It Straight!</td>
<td>107</td>
</tr>
<tr>
<td>Review of Part II</td>
<td>Exploring Relationships Between Variables</td>
<td>123</td>
</tr>
<tr>
<td>Chapter 11</td>
<td>Understanding Randomness</td>
<td>134</td>
</tr>
<tr>
<td>Chapter 12</td>
<td>Sample Surveys</td>
<td>139</td>
</tr>
<tr>
<td>Chapter 13</td>
<td>Experiments and Observational Studies</td>
<td>143</td>
</tr>
<tr>
<td>Review of Part III</td>
<td>Gathering Data</td>
<td>152</td>
</tr>
<tr>
<td>Chapter 14</td>
<td>From Randomness to Probability</td>
<td>158</td>
</tr>
<tr>
<td>Chapter 15</td>
<td>Probability Rules!</td>
<td>161</td>
</tr>
<tr>
<td>Chapter 16</td>
<td>Random Variables</td>
<td>169</td>
</tr>
<tr>
<td>Chapter 17</td>
<td>Probability Models</td>
<td>177</td>
</tr>
<tr>
<td>Review of Part IV</td>
<td>Randomness and Probability</td>
<td>189</td>
</tr>
<tr>
<td>Chapter 18</td>
<td>Sampling Distribution Models</td>
<td>200</td>
</tr>
<tr>
<td>Chapter 19</td>
<td>Confidence Intervals for Proportions</td>
<td>212</td>
</tr>
<tr>
<td>Chapter 20</td>
<td>Testing Hypotheses About Proportions</td>
<td>223</td>
</tr>
<tr>
<td>Chapter 21</td>
<td>More About Tests</td>
<td>232</td>
</tr>
<tr>
<td>Chapter 22</td>
<td>Comparing Two Proportions</td>
<td>246</td>
</tr>
<tr>
<td>Review of Part V</td>
<td>From the Data at Hand to the World at Large</td>
<td>258</td>
</tr>
<tr>
<td>Chapter 23</td>
<td>Inferences About Means</td>
<td>271</td>
</tr>
<tr>
<td>Chapter 24</td>
<td>Comparing Means</td>
<td>293</td>
</tr>
<tr>
<td>Chapter 25</td>
<td>Paired Samples and Blocks</td>
<td>306</td>
</tr>
<tr>
<td>Chapter 26</td>
<td>Comparing Counts</td>
<td>316</td>
</tr>
<tr>
<td>Review of Part VI</td>
<td>Learning About the World</td>
<td>327</td>
</tr>
<tr>
<td>Chapter 27</td>
<td>Inferences for Regression</td>
<td>342</td>
</tr>
<tr>
<td>Chapter 28</td>
<td>Analysis of Variance</td>
<td>355</td>
</tr>
<tr>
<td>Chapter 29</td>
<td>Multifactor Analysis of Variance</td>
<td>360</td>
</tr>
<tr>
<td>Chapter 30</td>
<td>Multiple Regression</td>
<td>365</td>
</tr>
<tr>
<td>Chapter 31</td>
<td>Multiple Regression Wisdom</td>
<td>368</td>
</tr>
<tr>
<td>Review of Part VII</td>
<td>Inference When Variables Are Related</td>
<td>370</td>
</tr>
<tr>
<td>Chapter 32</td>
<td>Rank-Based Nonparametric Tests</td>
<td>386</td>
</tr>
<tr>
<td>Chapter 33</td>
<td>The Bootstrap</td>
<td>394</td>
</tr>
</tbody>
</table>
Chapter 2 – Data

1. **The news.** Answers will vary.

3. **Walking in Circles.** *Who* – 32 volunteers. *What* – Sex, height, handedness, the number of yards walked before going out of bounds, and the side of the field on which the person walked out of bounds. *When* – Not specified. *Where* – Not specified. *Why* – The researcher was interested in whether people walk in circles when lost. *How* – Data were collected by observing the people on the field, as well as by measuring and asking the participants. *Variables* – There are five variables. Sex, handedness, and side of the field are categorical variables. Height and number of yards walked are quantitative variables.

5. **Investments.** *Who* – 48 China/India/Chindia funds listed at globeinvestor.com. *What* – 1 month, 1 year, and 5 year returns for each fund. *When* – The most recent periods of time. *Where* – globeinvestor.com website. *Why* – To compare investment returns for future investment decisions. *How* – globeinvestor.com uses reports from the fund companies. *Variables* – There are three variables, all of which are quantitative. 1 month return; 1 year return; 5 year return, annualized; all variables are measured as percentages.

7. **Tim Horton’s doughnuts.** *Who* – Donut types for sale at Tim Horton’s. *What* – Various nutritional characteristics (see variables, below). *When* – Not stated, but presumably the measurements were taken recently. *Where* – Tim Horton’s website. *Why* – To help customers make good nutritional choices. *How* – Further research would be needed to learn how they made these measurements, but presumably at some specialized food analysis lab. *Variables* – There are eight variables, all quantitative. Number of calories (kca/s), amounts of trans fat (g), total fat (g), sodium (mg), sugar (g), protein (g), % daily value of iron (percentage), % daily value of calcium (percentage). Units found by going to the website.

9. **Air travel.** *Who* – All airline flights in Canada. *What* – Type of aircraft, number of passengers, whether departures and arrivals were on schedule, and mechanical problems. *When* – This information is currently reported. *Where* – Canada. *Why* – This information is required by Transport Canada and the Canadian Transportation Agency. *How* – Data is collected from airline flight information. *Variables* – There are four variables. Type of aircraft, departure and arrival timeliness, and mechanical problems are categorical variables, and number of passengers is a quantitative variable.

13. **Weighing bears.** *Who* – 54 bears. *What* – Weight, neck size, length (no specified units), and sex. *When* – Not specified. *Where* – Not specified. *Why* - Since bears are difficult to weigh, the researchers hope to use the relationships between weight, neck size, length, and sex of bears to estimate the weight of bears, given the other more observable features of the bear. *How* – Researchers collected data on 54 bears they were able to catch. *Variables* – There are four variables; weight, neck size, and length are quantitative variables, and sex is a categorical variable. No units are specified for the quantitative variables. *Concerns* – The researchers are (obviously!) only able to collect data from bears they were able to catch. This method is a good one, as long as the researchers believe the bears caught are representative of all bears, in regard to the relationships between weight, neck size, length, and sex.

15. **Cars.** *Who* – Automobiles. *What* – Make, country of origin, type of vehicle and age of vehicle (probably in years). *When* – Not specified. *Where* – A large university. *Why* – Not specified. *How* – A survey was taken in campus parking lots. *Variables* – There are three categorical variables and one quantitative variable. Make, country of origin, and type of vehicle are categorical variables, and age of vehicle is a quantitative variable.

17. **Babies.** *Who* – 882 births. *What* – Mother’s age (in years), length of pregnancy (in weeks), type of birth (Cesarean, induced, or natural), level of prenatal care (none, minimal, or adequate), birth weight of baby (unit of measurement not specified, but probably grams), gender of baby (male or female), and baby’s health problems (none, minor, major). *When* – 1998–2000. *Where* – Large city hospital. *Why* – Researchers were investigating the impact of prenatal care on newborn health. *How* – It appears that they kept track of all births in the form of hospital records, although it is not specifically stated. *Variables* – There are three quantitative variables: mother’s age, length of pregnancy, and birth weight of baby. There are four categorical variables: type of birth, level of prenatal care, gender of baby, and baby’s health problems.

(herbal or sugar solution) is categorical, and severity rating (on a scale from 0 to 5) is quantitative. **Concerns** – The severity of a cold seems subjective and difficult to quantify. Also, the scientists may feel pressure to report negative findings about the herbal product.

21. **Streams.** **Who** – Streams. **What** – Name of stream, substrate of the stream (limestone, shale, or mixed), acidity of the water (measured in pH), temperature (in degrees Celsius), and BCI (unknown units). **When** – Not specified. **Where** – Northern Ontario. **Why** – Research is conducted for an ecology class. **How** – Not specified. **Variables** – There are five variables. Name and substrate of the stream are categorical variables, and acidity (pH), temperature (in degrees Celsius), and BCI are quantitative variables.

23. **Refrigerators.** **Who** – 41 refrigerators. **What** – Brand, cost (probably in dollars), size (in cu. ft.), type, estimated annual energy cost (probably in dollars), overall rating, and repair history (in percent requiring repair over the past five years). **When** – 2002. **Where** – United States. **Why** – The information was compiled to provide information to the readers of *Consumer Reports*. **How** – Not specified. **Variables** – There are seven variables. Brand, type, and overall rating are categorical variables. Cost, size (cu. ft.), estimated energy cost, and repair history (percentage) are quantitative variables.

25. **Horse race 2008.** **Who** – Kentucky Derby races. **What** – Date, winner, margin (in lengths), jockey, winner’s payoff (in dollars), duration of the race (in minutes and seconds), and track condition. **When** – 1875–2008. **Where** – Churchill Downs, Louisville, Kentucky. **Why** – It is interesting to examine the trends in the Kentucky Derby. **How** – Official statistics are kept for the race each year. **Variables** – There are seven variables. Winner, jockey, and track condition are categorical variables. Date, margin, winner’s payoff, and duration are quantitative variables.
1. **Graphs in the news.** Answers will vary.

3. **Tables in the news.** Answers will vary.

5. **Forest fires.** The relative frequency histogram is shown below:

<table>
<thead>
<tr>
<th>Cause of fire</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lightning</td>
<td>47.29</td>
</tr>
<tr>
<td>Human activities</td>
<td>46.04</td>
</tr>
<tr>
<td>Unknown</td>
<td>6.68</td>
</tr>
</tbody>
</table>

Causes for forest fires are about equally split between human activities and lightening. Only 6.68% of forest fires are due to unknown causes.
(Example: 47.29% = 0.4729 = 3719/7865)

7. **Ghosts.**
   a) It is NOT reasonable to assume that 66% of those polled expressed a belief in either ghosts or astrology. The percentages in the table add up to 185%! This tells us that we are not dealing with parts of a whole and that some respondents believe in more than one of the psychic phenomena listed. In other words, belief in ghosts and belief in astrology are not mutually exclusive. There is no way to know what percent of respondents believe in ghosts or astrology.
   b) As in part a, since the percentages are not meant to add up to 100%, there is no way to determine what percentage of respondents did not believe in any of these phenomena. With data collected in this fashion, the only way to determine the percentage of people who did not believe in any of the psychic phenomena would be to add another category to the survey when collecting the data.
   c) Since the percentages were not intended to add up to 100%, a pie chart is not appropriate. A bar chart nicely displays the percentages as relative heights of bars.
9. **Oil spills 2008.** The bar chart shows that grounding is the most frequent cause of oil spillage for these 319 spills, and allows the reader to rank the other types as well. If being able to differentiate between these close counts is required, use the bar chart. The pie chart is also acceptable as a display, but it’s difficult to tell whether, for example, there is a greater percentage of spills caused by grounding or collision. If you want to showcase the causes of oil spills as a fraction of all 319 spills, use the pie chart.

11. **Global warming.** Perhaps the most obvious error is that the percentages in the pie chart only add up to 92%, when they should, of course, add up to 100%. Furthermore, the three-dimensional perspective view distorts the regions in the graph, violating the area principle. The regions corresponding to No Solid Evidence and Due to Natural Patterns should be roughly the same size, at 20% and 21% of respondents, respectively. However, the angle for the 21% region looks much bigger. Always use simple, two-dimensional graphs.

13. **Complications**
   a) A bar chart is the proper display for these data. A pie chart is not appropriate, since these are counts, not fractions of a whole.
   b) The *Who* for these data is athletic trainers who used cryotherapy, which should be a cause for concern. A trainer who treated many patients with cryotherapy would be more likely to have seen complications than one who used cryotherapy rarely. We would prefer a study in which the *Who* referred to patients so we could assess the risks of each complication.
15. **Spatial distribution.**
   a) The relative frequency distribution of quadrant location is given below. Not all proportions are equal. In particular, the relative frequency for Quadrant 4 is approximately twice the other frequencies.

<table>
<thead>
<tr>
<th>Quadrant</th>
<th>Quadrant 1</th>
<th>Quadrant 2</th>
<th>Quadrant 3</th>
<th>Quadrant 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Frequency</td>
<td>0.18</td>
<td>0.21</td>
<td>0.22</td>
<td>0.39</td>
</tr>
</tbody>
</table>

b) The relative frequency distribution of quadrant location is given below. There seems to have some similarity with that in part a. For example, Quadrant 4 has the highest relative frequency and Quadrant 1 has the lowest.

<table>
<thead>
<tr>
<th>Quadrant</th>
<th>Quadrant 1</th>
<th>Quadrant 2</th>
<th>Quadrant 3</th>
<th>Quadrant 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Frequency</td>
<td>0.12</td>
<td>0.24</td>
<td>0.28</td>
<td>0.36</td>
</tr>
</tbody>
</table>

17. **More politics.**
   a) Marginal Distribution of Politics

   ![Marginal Distribution of Politics](image1)

   b) Conditional Distribution of Politics Among Males

   ![Conditional Distribution of Politics Among Males](image2)

   c) Conditional Distribution of Politics Among Females

   ![Conditional Distribution of Politics Among Females](image3)

   d) Distribution of Sex Across Political Categories

   ![Distribution of Sex Across Political Categories](image4)

   e) The percentage of males and females varies across political categories. The percentage of self-identified Liberals and Moderates who are female is about
twice the percentage of self-identified Conservatives who are female. This would suggest that sex and politics are not independent.

19. Canadian languages.
   a) 21,130,000 Canadians speak English only. 21,130,000/31,241,000 total Canadians $\approx 67.6\%$
   b) 4,142,000 Canadians speak French only and 5,449,000 speak both French and English, for a total of 9,591,000 French speakers. 9,591,000/31,241,000 total Canadians $\approx 30.7\%$
   c) 4,011,000 French and 3,018,000 French and English speakers yield a total of 7,029,000 French speakers in Quebec. 7,029,000/7,436,000 Quebec residents $\approx 94.5\%$
   d) 7,029,000 Quebec residents speak French and 9,591,000 Canadians speak French. The percentage of French-speaking Canadians who live in Quebec is 7,029,000/9,591,000 $\approx 73.3\%$
   e) If language knowledge were independent of province, we would expect the percentage of French-speaking residents of Quebec to be the same as the overall percentage of Canadians who speak French. Since 30.7% of all Canadians speak French while 94.5% of residents of Quebec speak French, there is evidence of an association between language knowledge and province.

   a) The table shows the marginal totals. It rained on 34 of 365 days, or 9.3% of the days.
   b) Rain was predicted on 90 of 365 days. 90/365 $\approx 24.7\%$ of the days.
   c) The forecast of rain was correct on 27 of the days it actually rained and the forecast of No Rain was correct on 268 of the days it didn’t rain. So, the forecast was correct a total of 295 times. 295/365 $\approx 80.8\%$ of the days.
   d) On rainy days, rain had been predicted 27 out of 34 times (79.4%). On days when it did not rain, forecasters were correct in their predictions 268 out of 331 times (81.0%). These two percentages are very close. There is no evidence of an association between the type of weather and the ability of the forecasters to make an accurate prediction.

<table>
<thead>
<tr>
<th></th>
<th>Actual Weather</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rain</td>
<td>No Rain</td>
</tr>
<tr>
<td>Forecast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rain</td>
<td>27</td>
<td>63</td>
</tr>
<tr>
<td>No Rain</td>
<td>7</td>
<td>268</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>331</td>
</tr>
</tbody>
</table>
23. **Working parents.**

The Gallup poll doesn’t provide strong evidence of a change in people’s opinions regarding the ideal family in today’s society between the years of 1991 and 2001. The conditional distributions of opinion by year appear roughly the same. For example, the percentage of respondents in 1991 who thought the ideal family had two parents that worked full time was 14%, and in 2001, the percentage was 13%.

25. **Blood pressure.**

a) The marginal distribution of blood pressure for the employees of the company is the total column of the table, converted to percentages. 20% low, 49% normal, and 31% high blood pressure.

b) The conditional distribution of blood pressure within each age category is:

- Under 30: 28% low, 49% normal, 23% high
- 30-49: 21% low, 51% normal, 28% high
- Over 50: 16% low, 47% normal, 37% high
c) A segmented bar chart of the conditional distributions of blood pressure by age category is at the right.

d) In this company, as age increases, the percentage of employees with low blood pressure decreases, and the percentage of employees with high blood pressure increases.

e) No, this does not prove that people’s blood pressure increases as they age. Generally, an association between two variables does not imply a cause-and-effect relationship. Specifically, these data come from only one company and cannot be applied to all people. Furthermore, there may be some other variable that is linked to both age and blood pressure.

27. Smoking gene.

a) The marginal distribution of genotype is given below:

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Marginal percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>GG</td>
<td>42.71</td>
</tr>
<tr>
<td>GT</td>
<td>45.08</td>
</tr>
<tr>
<td>TT</td>
<td>12.21</td>
</tr>
</tbody>
</table>

b) The conditional distributions of genotype for the four categories of smokers are given in columns 2–5 of the table below.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Cigarettes per day</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-10</td>
</tr>
<tr>
<td>GG</td>
<td>48.06</td>
</tr>
<tr>
<td>GT</td>
<td>42.96</td>
</tr>
<tr>
<td>TT</td>
<td>8.99</td>
</tr>
<tr>
<td>All</td>
<td>100.00</td>
</tr>
</tbody>
</table>
c) Though not a very noticeable difference, the percentages of smokers with genotype GT (also TT) are slightly higher among heavy smokers. However, this is only an observed association. This does not prove that presence of T increases susceptibility to nicotine addiction. We cannot conclude that this increase was caused by the presence of T. There can be many factors associated with the presence of T, and some of these factors might be the reason for the increase in susceptibility to nicotine addiction.

29. Antidepressants and bone fractures. These data provide evidence that taking a certain class of antidepressants (SSRI) might be associated with a greater risk of bone fractures. Approximately 10% of the patients taking this class of antidepressants experience bone fractures. This is compared to only approximately 5% in the group that were not taking the antidepressants.

a) The two-way table and the conditional distributions (percentages) of ‘car accident’ (crash or non-crash) for cell phone owners and non-cell phone owners are given below. The proportion of crashes is higher for cell phone owners than for non-cell phone owners.
b) On the basis of this study, we cannot conclude that the use of a cell phone increases the risk of a car accident. This is only an observed association between cell phone ownership and the risk of car accidents. We cannot conclude that the higher proportion of accidents was caused by the use of a cell phone. There can be lots of other factors common to cell phone owners, and some of those factors can be the reason for the accidents.

33. **Aboriginal identity.**

a) The second column includes some individuals in the 3rd, 4th, and 6th columns, so it is not a standard contingency table.

b) Use the label “Aboriginal population not included in columns 3, 4, and 5” (or call them “other Aboriginals”).

c) There are 22 560 Canadians who are Inuit from Nunavut. The Canadian population is 29 639 030, so the proportion of Canadians who are Inuit from Nunavut is 22 560/29 639 030 = 0.00076 = 0.08% (approx.).

d) There are 22 560 Canadians who are Inuit from Nunavut. The total Canadian Aboriginal population is 976 305, so the proportion of Canadian Aboriginals who are Inuit from Nunavut is 22 560/976 305 = 0.02311 = 2.31% (approx.).

e) The total Canadian Aboriginal population is 976 305 and of them 45 070 are Inuit. So 45 070/976 305 = 0.0462 = 4.62% of Aboriginals are Inuit.

f) The total Canadian Aboriginal population is 976 305 and of them 22 720 are from Nunavut. So 22 720/976 305 = 0.0233 = 2.33% of Aboriginals are from Nunavut.

g) The total population in Nunavut is 26 665 and 22 560 of them are Inuit. So 22 560/26 665 = 0.8461 = 94.61% of the people from Nunavut are Inuit.
h) There are 22,720 Nunavut Aboriginals and 22,560 of them are Inuit. So
\[
\frac{22,560}{22,720} = 0.9930 = 99.30\% \text{ of Nunavut Aboriginals are Inuit.}
\]
i) The total Inuit population is 45,070 and of them 22,560 are from Nunavut. So
\[
\frac{22,560}{45,070} = 0.5006 = 50.06\% \text{ of Inuit live in Nunavut.}
\]
j) The total number of Ontario Aboriginals is 188,315, and 188,315 – 131,560 – 48,340 – 1375 = 7040 of them are other Aboriginals (i.e., other than Inuit, Metis, or N.A. Indian) and so 7040/188,315 = 0.0374 = 3.74\% of Ontario Aboriginals could not be simply classified into either Inuit, Metis, or N.A. Indian.

k) A table of percentages of total provincial population for each Aboriginal identity group (Inuit, Metis, N.A. Indian) for Newfoundland, Ontario, Saskatchewan, and Alberta is given below. The second table is a bit easier if using MINITAB. The side-by-side bar charts below show that Saskatchewan has the highest proportion of N.A. Indian and Metis proportions. Ontario, Saskatchewan, and Alberta have very small proportions of Inuit.

<table>
<thead>
<tr>
<th>Region</th>
<th>Percent N.A. Indian</th>
<th>Percent Metis</th>
<th>Percent Inuit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newfoundland and Labrador</td>
<td>1.38561</td>
<td>1.07857</td>
<td>0.897496</td>
</tr>
<tr>
<td>Ontario</td>
<td>1.16574</td>
<td>0.42834</td>
<td>0.012184</td>
</tr>
<tr>
<td>Saskatchewan</td>
<td>8.69486</td>
<td>4.53665</td>
<td>0.024399</td>
</tr>
<tr>
<td>Alberta</td>
<td>2.88986</td>
<td>2.24606</td>
<td>0.037060</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>Newfoundland and Labrador</th>
<th>Ontario</th>
<th>Saskatchewan</th>
<th>Alberta</th>
</tr>
</thead>
<tbody>
<tr>
<td>N.A. Indian</td>
<td>1.38561</td>
<td>1.16574</td>
<td>8.96486</td>
<td>2.88986</td>
</tr>
<tr>
<td>Metis</td>
<td>1.07857</td>
<td>0.42834</td>
<td>4.53665</td>
<td>2.24606</td>
</tr>
<tr>
<td>Inuit</td>
<td>0.89750</td>
<td>0.01218</td>
<td>0.02440</td>
<td>0.03706</td>
</tr>
</tbody>
</table>

Chart of Newfoundland and, Ontario, Saskatchewan, Alberta vs. Group

Copyright © 2012 Pearson Canada Inc.
35. Hospitals.
   a) The marginal totals have been added to the table:

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Large Hospital</th>
<th>Small Hospital</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major surgery</td>
<td>120 of 800</td>
<td>10 of 50</td>
<td>130 of 850</td>
</tr>
<tr>
<td>Minor surgery</td>
<td>10 of 200</td>
<td>20 of 250</td>
<td>30 of 450</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>130 of 1000</strong></td>
<td><strong>30 of 300</strong></td>
<td><strong>160 of 1300</strong></td>
</tr>
</tbody>
</table>

   160 of 1300, or about 12.3%, of the patients had a delayed discharge.
   b) Yes. Major surgery patients were delayed 130 of 850 times, or about 15.3% of the time. Minor surgery patients were delayed 30 of 450 times, or about 6.7% of the time.
   c) Large Hospital had a delay rate of 130 of 1000, or 13%. Small Hospital had a delay rate of 30 of 300, or 10%. The small hospital has the lower overall rate of delayed discharge.
   d) Large Hospital: Major Surgery 15% delayed and Minor Surgery 5% delayed. Small Hospital: Major Surgery 20% delayed and Minor Surgery 8% delayed. Even though small hospital had the lower overall rate of delayed discharge, the large hospital had a lower rate of delayed discharge for each type of surgery.
   e) No. While the overall rate of delayed discharge is lower for the small hospital, the large hospital did better with both major surgery and minor surgery.
   f) The small hospital performs a higher percentage of minor surgeries than major surgeries. 250 of 300 surgeries at the small hospital were minor (83%). Only 200 of the large hospital’s 1000 surgeries were minor (20%). Minor surgery had a lower delay rate than major surgery (6.7% to 15.3%), so the small hospital’s overall rate was artificially inflated. Simply put, it is a mistake to look at the overall percentages. The real truth is found by looking at the rates after the information is broken down by type of surgery, since the delay rates for each type of surgery are so different. The larger hospital is the better hospital when comparing discharge delay rates.

37. Graduate admissions.
   a) 1284 applicants were admitted out of a total of 3014 applicants. 1284/3014 = 42.6%
b) 1022 of 2165 (47.2%) of males were admitted. 262 of 849 (30.9%) of females were admitted.

c) Since there are four comparisons to make, the table at the right organizes the percentages of males and females accepted in each program. Females are accepted at a higher rate in every program.

c| Program | Males | Females |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>61.9%</td>
<td>82.4%</td>
</tr>
<tr>
<td>2</td>
<td>62.9%</td>
<td>68.0%</td>
</tr>
<tr>
<td>3</td>
<td>33.7%</td>
<td>35.2%</td>
</tr>
<tr>
<td>4</td>
<td>5.9%</td>
<td>7%</td>
</tr>
</tbody>
</table>

d) The comparison of acceptance rate within each program is most valid. The overall percentage is an unfair average. It fails to take the different numbers of applicants and different acceptance rates of each program. Women tended to apply to the programs in which gaining acceptance was difficult for everyone. This is an example of Simpson’s Paradox.
Chapter 4 – Displaying and Summarizing Quantitative Data

1. **Statistics in print.** Answers will vary.

3. **In the news.** Answers will vary.

5. **Thinking about shape.**
   a) The distribution of the number of speeding tickets each student in the senior class of a college has ever had is likely to be unimodal and skewed to the right. Most students will have very few speeding tickets (maybe 0 or 1), but a small percentage of students will likely have comparatively many (3 or more) tickets.
   b) The distribution of player’s scores at the U.S. Open Golf Tournament would most likely be unimodal and slightly skewed to the right. The best golf players in the game will likely have around the same average score, but some golfers might be off their game and score 15 strokes above the mean. (Remember that high scores are undesirable in the game of golf!)
   c) The weights of female babies in a particular hospital over the course of a year will likely have a distribution that is unimodal and symmetric. Most newborns have about the same weight, with some babies weighing more and less than this average. There may be slight skew to the left, since there seems to be a greater likelihood of premature birth (and low birth weight) than post-term birth (and high birth weight).
   d) The distribution of the length of the average hair on the heads of students in a large class would likely be bimodal and skewed to the right. The average hair length of the males would be at one mode, and the average hair length of the females would be at the other mode, since women typically have longer hair than men. The distribution would be skewed to the right, since it is not possible to have hair length less than zero, but it is possible to have a variety of lengths of longer hair.

7. **Sugar in cereals.**
   a) The distribution of the sugar content of breakfast cereals is bimodal, with a cluster of cereals with sugar content around 10% sugar and another cluster of cereals around 48% sugar. The lower cluster shows a bit of skew to the right. Most cereals in the lower cluster have between 0% and 10% sugar. The upper cluster is symmetric, with centre around 45% sugar.
   b) There are two different types of breakfast cereals, those for children and those for adults. The children’s cereals are likely to have higher sugar contents, to make them taste better (to kids, anyway!). Adult cereals often advertise low sugar content.
9. **Test scores.**
   a) The number of students scoring 40 or higher is approximately \(16+5+3+1 = 25\) (These are only approximate heights of the bars in the histogram after 40) and the percentage = \(25/110 = 22.7\) percent.
   b) The number of students scoring between 25 and 35 is the sum of the heights of the two bars between 25 and 35. This is approximately \(21+31 = 52\), and so the percentage is \(52/110 = 47.3\) percent.
   c) The distribution is symmetric. The centre (the median) is around 32. The scores range from 0–60, but the few scores close to zero may be outliers (with a gap of just one we might not be able to conclude as a clear outlier, but they are somewhat unusual compared to the rest of the scores).

11. **Election 2000.**
   a) The distribution is right skewed and so it is logical to expect the mean to be greater than the median. The median is the \((301+1)/2\) th = 151st value in the ordered data set. This value must be in the third interval from the left of the histogram, i.e., the median must be in the interval 0.5–0.7.
   b) The distribution is right skewed. The median is between 0.5 and 0.7. The values of the percentage of rejected ballots range from 0–5 (approx.). The largest value(s) (after 4.8) looks like an outlier, so it warrants further investigation.

13. **Summaries.**
   a) The mean price of the electric smoothtop ranges is $1001.50.
   b) In order to find the median and the quartiles, the list must be ordered.
      
      \[
      565 \quad 750 \quad 850 \quad 900 \quad 1000 \quad 1050 \quad 1050 \quad 1200 \quad 1250 \quad 1400
      \]
      
      The median price of the electric smoothtop ranges is $1025.
      Quartile 1 = $850 and Quartile 3 = $1200.
   c) The range of the distribution of prices is Max – Min = $1400 – $565 = $835.
      The IQR = Q3 – Q1 = $1200 – $850 = $350.

15. **Mistake.**
   a) As long as the boss’s true salary of $200 000 is still above the median, the median will be correct. The mean will be too large, since the total of all the salaries will decrease by \(2 000 000 – 200 000 = 1 800 000\), once the mistake is corrected.
   b) The range will likely be too large. The boss’s salary is probably the maximum, and a lower maximum would lead to a smaller range. The IQR will likely be unaffected, since the new maximum has no effect on the quartiles. The standard deviation will be too large, because the $2 000 000 salary will have a large squared deviation from the mean.

17. **Standard deviation I.**
   a) Set 2 has the greater standard deviation. Both sets have the same mean (6), but set two has values that are generally farther away from the mean.
      \[
      SD(Set 1) = 2.24 \quad SD(Set 2) = 3.16
      \]
b) Set 2 has the greater standard deviation. Both sets have the same mean (15), maximum (20), and minimum (10), but 11 and 19 are farther from the mean than 14 and 16.
\[ \text{SD(Set 1)} = 3.61 \quad \text{SD(Set 2)} = 4.53 \]
c) The standard deviations are the same. Set 2 is simply Set 1 + 80. Although the measures of centre and position change, the spread is exactly the same.
\[ \text{SD(Set 1)} = 4.24 \quad \text{SD(Set 2)} = 4.24 \]

19. Payroll.

a) The mean salary is \[ \frac{1200 + 700 + 6(400) + 4(500))}{12} = \$525 \]
The median salary is the middle of the ordered list:
400 400 400 400 400 400 500 500 500 500 700 1200
The median is $450.
b) Only two employees, the supervisor and the inventory manager, earn more than the mean wage.
c) The median better describes the wage of the typical worker. The mean is affected by the two higher salaries.
d) The IQR is the better measure of spread for the payroll distribution. The standard deviation and the range are both affected by the two higher salaries.

21. The Great One.

a) The stemplot is shown below.

\[ \begin{array}{c}
0 & 4 \\
0 & 66 \\
0 & 99 \\
1 & 0 \\
1 & 233 \\
1 & 44 \\
1 & 666 \\
1 & 89 \\
2 & 0011 \\
\end{array} \]
points scored
(2 | 0 means 200 points)
b) The distribution looks slightly left skewed. The scores range from about 40–210 (note that these are truncated values). The median (the average of the 10th and the 11th values) is about 140. There are no outliers.
c) If we consider this as a left-skewed distribution, then the five-number summary is an appropriate summary and is given below.

\[ \begin{array}{ccccccc}
\text{Descriptive Statistics: points scored} \\
\text{Variable} & \text{Minimum} & Q1 & \text{Median} & Q3 & \text{Maximum} \\
\text{points scored} & 48.0 & 98.3 & 145.5 & 192.8 & 215.0 \\
\end{array} \]
23. **How tall?** The histogram shows some low outliers in the distribution of height estimates. These are probably poor estimates and will pull the mean down. The median is likely to give a better estimate of the professor’s true height.

25. **Wayne Gretzky.**
   a) The distribution of the points scored per season by Wayne Gretzky is slightly skewed to the left and has no outliers. The median is more resistant to the skewness than the mean.
   b) The median, or middle of the ordered list, is 145.5 points. This is the average of the 10th value (142) and the 11th value (149).
   c) The mean should be slightly lower as the distribution is slightly left-skewed.

27. **New homes.**
   a) The stem and leaf plot for rounded data is given below:
      -0 1
      0 112334444
      0 55566
      1 234
      1 5
      2
      2 8
      3
      3
      4
      4 6
   Percent Increase (rounded)
   (4|6 means 46 percent)
   b) Minimum = –1.0
      The first quartile (Q1) = (22/4)th value = average of the 5th and the 6th values in the ordered data = 3.0
      Median = (22/2)th value = 11th value = 5.0
      The third quartile (Q3) = average of the 5th and the 6th values from the end of the ordered data = (12+13)/2 = 12.5
      Maximum = 46.0
   c) The mean is 8.62 (or 8.57 with the rounded data). Yes, the mean must be larger than the median because the distribution is right-skewed.
   d) The percentage increases in price range from –1% to 15%. The largest value 46 (possibly 28 also) is an outlier. The median increase is 5%. The distribution is right-skewed.
   e) The data sorted by the percentage increase is shown below. This shows large increases in the Prairie and Atlantic regions.
## Data Display

<table>
<thead>
<tr>
<th>Row</th>
<th>Metropolitan Area</th>
<th>Percentage Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Windsor</td>
<td>-0.6</td>
</tr>
<tr>
<td>2</td>
<td>Victoria</td>
<td>1.2</td>
</tr>
<tr>
<td>3</td>
<td>Charlottetown</td>
<td>1.4</td>
</tr>
<tr>
<td>4</td>
<td>Saint John, Fredericton and Moncton</td>
<td>2.4</td>
</tr>
<tr>
<td>5</td>
<td>Ottawa–Gatineau</td>
<td>3.1</td>
</tr>
<tr>
<td>6</td>
<td>Kitchener</td>
<td>3.4</td>
</tr>
<tr>
<td>7</td>
<td>Québec</td>
<td>3.9</td>
</tr>
<tr>
<td>8</td>
<td>Hamilton</td>
<td>3.9</td>
</tr>
<tr>
<td>9</td>
<td>London</td>
<td>4.0</td>
</tr>
<tr>
<td>10</td>
<td>St. Catharines–Niagara</td>
<td>4.3</td>
</tr>
<tr>
<td>11</td>
<td>Montréal</td>
<td>4.5</td>
</tr>
<tr>
<td>12</td>
<td>Toronto and Oshawa</td>
<td>4.5</td>
</tr>
<tr>
<td>13</td>
<td>Calgary</td>
<td>5.3</td>
</tr>
<tr>
<td>14</td>
<td>Vancouver</td>
<td>6.1</td>
</tr>
<tr>
<td>15</td>
<td>Greater Sudbury and Thunder Bay</td>
<td>6.3</td>
</tr>
<tr>
<td>16</td>
<td>St. John's</td>
<td>12.0</td>
</tr>
<tr>
<td>17</td>
<td>Halifax</td>
<td>12.8</td>
</tr>
<tr>
<td>18</td>
<td>Edmonton</td>
<td>13.5</td>
</tr>
<tr>
<td>19</td>
<td>Winnipeg</td>
<td>15.0</td>
</tr>
<tr>
<td>20</td>
<td>Regina</td>
<td>27.8</td>
</tr>
<tr>
<td>21</td>
<td>Saskatoon</td>
<td>46.2</td>
</tr>
</tbody>
</table>

29. **Alberta slots.** The stem and leaf plot, a dotplot, and the five-number summary for these data are given below. The distribution looks slightly right-skewed. The median number of slot machines is 202. The interquartile range is 314 – 169 = 145. The maximum value 600 that comes after a long gap is an outlier. (1.5 x IQR rule also identifies this as an outlier).

```
0 5
1 44
1 588
2 000
2 578
3 4
3
4 02
4
5
5
6 0
```

numbers of slot machines
20  Part 1 Exploring and Understanding Data

(3 | 4 means 340 slot machines)

![Dotplot of numbers of slot machines]

Descriptive Statistics: numbers of slot machines

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>numbers of slot</td>
<td>50.0</td>
<td>169</td>
<td>202.0</td>
<td>314</td>
<td>600.0</td>
</tr>
</tbody>
</table>

   a) This is not a histogram. The horizontal axis should split the number of counts from each site into bins. The vertical axis should show the number of sites in each bin. The given graph is nothing more than a bar chart, showing the bird count from each site as its own bar. It is of absolutely no use for describing the shape, centre, spread, or unusual features of the distribution of bird counts.
   
   ![Christmas Bird Count]

33. Bi-lingual Ni-lingual.
   a) The histogram of the proportion of city residents who are bilingual and a bar chart showing the proportion of bilingual versus city are shown below. The highest proportion of bilinguals is in Montreal with more than 50 percent bilinguals, and the lowest proportion is in Brantford with less than 5 percent bilinguals. The median is about 8%. The distribution of the proportion of city residents who are bilingual is right skewed. This means a relatively smaller number of cities with high proportion of bilinguals and more cities with a low proportion of bilinguals. It looks like the distribution of a few cities (the ones with the proportion of bilinguals greater than 0.3) has a different shape than that of the other cities. Quebec, Greater Sudbury/Grand Sudbury, Sherbrooke, Ottawa–Gatineau, Moncton, and Montreal constitute this group. The five-number summary (disregarding this grouping) is given below. About 25% of the cities have less than 7% bilinguals.
Descriptive Statistics: Proportion Bi-lingual

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Min.</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion Bi-lin</td>
<td>33</td>
<td>0.1533</td>
<td>0.1388</td>
<td>0.0473</td>
<td>0.0694</td>
<td>0.0825</td>
<td>0.1730</td>
<td>0.5189</td>
</tr>
</tbody>
</table>

b) The bar chart below shows the cities in Quebec and New Brunswick in red. They have relatively high proportions of bilinguals.
c) The histogram of the proportion of city residents who speak neither English nor French and a bar chart showing this proportion versus city are shown below. The highest proportion of residents who speak neither English nor French is in Vancouver with more than 5%, and the lowest proportion is in Saguenay with about 0.05%. The median is about 0.43%. The distribution of the proportion of city residents who speak neither English nor French is right skewed. This means there is a relatively smaller number of cities with a high proportion of residents who speak neither English nor French and more cities with a low proportion. Three cities (Vancouver, Abbotsford, and Toronto, the bars above 0.036 on the histogram) appear to be outliers.
### Descriptive Statistics: Proportion Ni-lingual

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion Ni-lin</td>
<td>33</td>
<td>0.00998</td>
<td>0.01270</td>
<td>0.000501</td>
<td>0.00202</td>
<td>0.00434</td>
<td>0.01344</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion Ni-lin</td>
<td>0.05169</td>
</tr>
</tbody>
</table>

#### 35. Trimmed mean.

a) The sum of the 10 values given is 64, so the mean is $64/10 = 6.4$.

The data arranged in increasing order: 1, 5, 6, 6, 7, 7, 8, 8, 8, 8. The median is 7 (the average of the 5th and the 6th values).

There are 10 values in the data set and 10% of 10 is 1. To calculate the 10% trimmed mean, delete the smallest and the largest value in the sorted data set and calculate the average of the remaining values. The 10% trimmed mean is $55/8 = 6.875$.

b) The marijuana data of Exercise 34, arranged in increasing order:

2 5 6 6 6 7 10 10 12 15 17 19 19 21 23 27 31 37 40 53

There are 20 values in the data set and 5% of 20 is 1. To calculate the 5% trimmed mean, delete the smallest and the largest value in the sorted data set and calculate the average of the remaining values. The 5% trimmed mean is $311/18 = 17.28$.

c) The data in part (a) is left skewed. The median would be the most appropriate measure for centre, followed by the trimmed mean, then the mean. The data in
part (b) is right skewed. The median would be the most appropriate measure for centre, followed by the trimmed mean, then the mean.

37. Zip codes. Even though zip codes are numbers, they are not quantitative in nature. Zip codes are categories. A histogram is not an appropriate display for categorical data. The histogram the Holes-R-Us staff member displayed doesn’t take into account that some 5-digit numbers do not correspond to zip codes or that zip codes falling into the same classes may not even represent similar cities or towns. The employee could design a better display by constructing a bar chart that groups together zip codes representing areas with similar demographics and geographic locations.

   a) The five-number summary of the national averages is given below:
      
      Descriptive Statistics: Ave Score
      Variable   Minimum    Q1  Median   Q3  Maximum
      Avg. Score 406.00  485.50  499.50  520.00   548.00

      The IQR, mean, and the standard deviation of the national averages is given below:
      Descriptive Statistics: Ave Score
      Variable    Mean  StDev   IQR
      Avg. Score  497.22  30.93  34.50

      The two smallest values of 406 and 424 are outliers (using the 1.5 x IQR rule, they are smaller than Q1 - 1.5 x IQR, whereas the maximum value is not an outlier since it is not greater than Q3 + 1.5 x IQR). The outliers attract the mean toward them, whereas the median is resistant to outliers. Since this outlier is a smaller value compared to the rest, we can expect the mean to be smaller than the median.

   b) Since there are outliers in the data set, the five-number summary is better than mean and standard deviation. The values in the five-number summary are resistant to outliers. The mean and the standard deviation are not resistant measures.

   c) Thirty-two countries participated in the program. The highest national average is 548 and the lowest is 406. The median national average is 499.5. The interquartile range is 34.5. Twenty-five percent (i.e., 8 countries) of the participating countries had a national average 485.5 or below and at least 25% of the countries had a national average of 520 or above. Canada’s national average (which is 527) is in the top 25% of all participating countries, more specifically, the 5th highest of all participating countries.

   d) Using the 68-95-99.7 rule, the middle 68% of the students have their scores within one standard deviation from the mean, i.e., $527 - 86 = 441$ to $527 + 86 = 613$. 
The middle 95% of the students have their scores within two standard deviations from the mean, i.e., $527 - 2 \times 86 = 355$ to $527 + 2 \times 86 = 699$. The middle 99.7% of the students have their scores within three standard deviations from the mean, i.e., $527 - 3 \times 86 = 269$ to $527 + 3 \times 86 = 785$. Using this information we can fill in the blanks as shown below:

- About two-thirds of students scored between 441 and 613.
- Only about 5% scored less than 355 or more than 699.
- Only a real math genius could have scored above 785.

41. First Nations

a) A histogram (or a stemplot) is an appropriate graphical display. Both these displays are given below. The distribution is right skewed. This means the portion of large registry groups is relatively small. There are some (one or two) outliers (indicated by the gaps in the histogram or stemplot). The median size is 678. The interquartile range is 1180.

```
Size  Frequency
0 00011111222222223333333333333344444445555555555555555556666666+
1 00000011123344555567778999
2 0011122334589
3 12679
4 1
5 24
6
7 17
8
9
10 6
```
Descriptive Statistics: Size

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>138</td>
<td>40.0</td>
<td>434</td>
<td>678</td>
<td>1613</td>
<td>10607</td>
<td>1180</td>
</tr>
</tbody>
</table>

b) If we calculate the band sizes, most band sizes will be same as the registry group sizes since only some bands consist of more than one registry group. Some bands will be relatively large if it consists of many registry groups, like Six Nations of the Grand River in Ontario, which consists of 13 registry groups. This will increase the mean, median, standard deviation, and IQR. The histogram can have gaps due to a few bands with many registry groups.

43. Customer Database.

a) The mean of 54.41 is meaningless. The codes assigned to the titles are categories, even though the categories are represented by numbers. Averaging is only appropriate for quantitative data.

b) Typically, the mean and standard deviation are influenced by outliers and skewness.

c) There is no point in discussing the reason for the differences in the summary statistics, because the summary statistics are only appropriate for quantitative data. The title codes are categorical.

45. Election 2000 yet again.

a) The closeness of the mean and median suggests that the distribution of the percentage of voter turnout in the 2000 election is approximately symmetric.

b) The distribution of the percentage of voter turnout in the 2000 election is approximately symmetric, with mean 61.23 and standard deviation 5.244.

c) Mean + 2 Stdev = 61.231 + 2 x 5.244 = 71.719
   Mean -2 Stdev = 61.231 - 2 x 5.244 = 50.743
   From the histogram there appears to be about 13 observations below 51 and 7 observations above 71. This is about 20/301 = 6.6%. Thus, about 93.4% of the observations are within 2 standard deviations from the mean. This is pretty close to 95% as we would expect according to the empirical rule for bell-shaped distributions.

d) The overall percentage won’t be exactly the same as the mean of the percentages as the number of eligible voters is not exactly the same in each electoral district. However, since the number of eligible voters in each riding is pretty close, we should expect the overall percentage to be fairly close to 61.23%.

47. Grouped data.

a) The class midpoints and the mean and the standard deviation of the midpoints are given below:
Height (class midpoint)
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  64  64  64  64  64  64  64  64  64  64  64  64  64  64  64  64  64  64  64  64  64
64  64  64  64  64  67  67  67  67  67  67  67  67  67  67  67  67  67  67  67  67  67  67
67  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70
70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70
73  73  73  73  73  73  73  73  73  73  73  73  73  73  73  73  73  73  73  73  73  73  73

Descriptive Statistics: Height (class midpoint)
Variable       N   Mean  StDev
Height (class mid) 130  67.069  3.996

The actual heights and the mean and the standard deviation of the actual values are given below. The mean and the standard deviation of the actual data are very close to those calculated using the midpoints. For grouped data, we assume that all the values in a class are equal to the midpoint. This assumption is usually reasonable unless the class width is big.

Height (actual)
60  60  61  61  61  61  61  61  62  62  62  62  62  62
62  62  63  63  63  63  63  63  64  64  64  64  64  64  64
65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65
65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65
68  68  68  68  68  68  68  68  68  68  68  68  68  68  68  68  68  68  68  68  68  68  68
69  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70  70
71  71  71  71  71  71  71  71  72  72  72  72  72  72  72  72  72  72  72  72  72  72  72
73  73  73  74  74  74  75  75  75  75  75  75  75  75  75  75  75  75  75  75  75  75  75

Descriptive Statistics: Height (actual)
Variable       N   Mean  StDev
Height (actual) 130  67.115  3.792

b) 
\[ \text{Mean} = \text{Mean} \left( \bar{X} \right) = \frac{\sum_i f_i m_i}{\sum_i f_i}, \quad \text{Variance} = \frac{\sum_i f_i (m_i - \bar{X})^2}{\sum_i f_i - 1} \]