Chapter 2 – Describing Distributions with Numbers

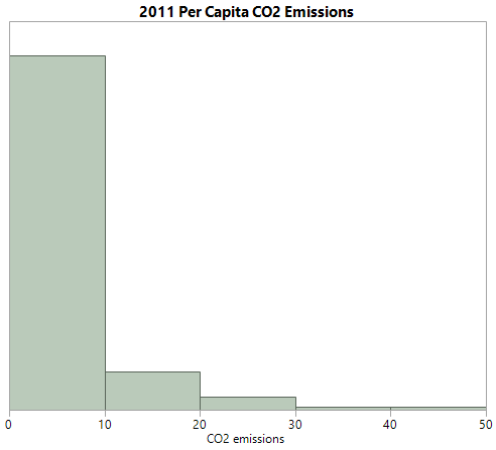
2.1 Mean *E. coli* level is $\bar{x} = \frac{291.0 + 10.9 + \ldots + 9.6}{16} = 56.28$ per milliliter. The mean is greater than most of the observations because of the two outliers (291.0 and 190.4).

2.2 The mean expenditure for all countries including the United States is $2808.66. The mean when the United States is excluded is $2622.26. The United States as an outlier increases the mean by $186.40, even with as many as 34 other countries.

2.3 The mean travel time is $\overline{x} = 31.25$ minutes. The median travel time is 22.5 minutes. The mean is significantly larger than the median due to the right-skew in the distribution of times.

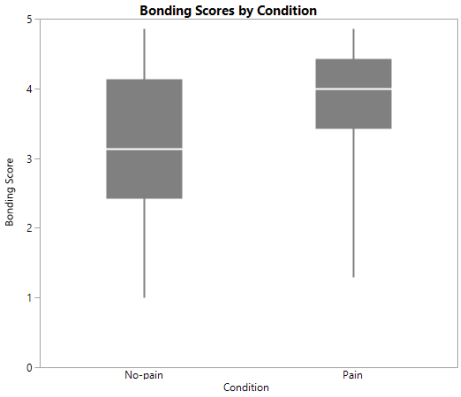
2.4 The mean is larger than the median; surely the distribution of home prices is right-skewed. This means that the mean is $348,900 and the median is $301,400.

2.5 A histogram is given. Note the right-skew. So, the mean is larger than the median. The mean is 4.866 and the median is 2.625 tons per person.

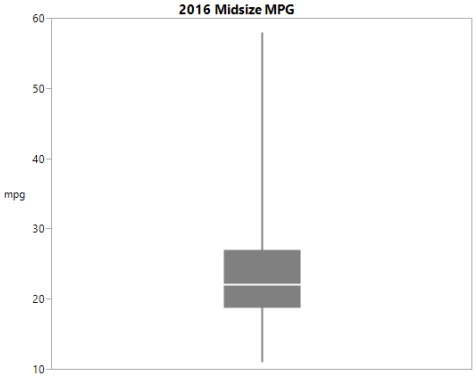


2.6 (a) and **(b)** The five-number summaries and boxplots for each group are provided. **(c)** The pain group tends to have higher bonding scores. There is less variability in the pain group versus the no-pain group. The pain group has at least one low outlier, which would be seen in a modified boxplot.

Group	Min	Q_1	Median	Q_3	Max
No-pain	1.00	2.43	3.14	4.14	4.86
Pain	1.29	3.43	4.00	4.43	4.86



2.7 (a) Minimum = 11, Q_1 = 18.75, median = 22, Q_3 = 27, and maximum = 58. **(b)** The boxplot provided shows right-skew in the distribution of mpg values. There are high outliers (most likely hybrid cars). A modified boxplot will explicitly show these outliers.



2.8 For these data, Q_1 = 10, Q_3 = 30, and so IQR = 30 – 10 = 20 minutes. $Q_1 - 1.5 \times IQR = 10 - 1.5 \times 20 = -20$ minutes. Obviously no times can be negative, so no outliers are in the left tail. $Q_3 + 1.5 \times IQR = 30 + 1.5 \times 20 = 60$ minutes. The "60" would not be considered an outlier, but it's close.

2.9 IQR = 27 – 19 = 8, so $Q_3 + 1.5 \times IQR = 27 + 1.5 \times 8 = 39$. There are nine values greater than 39 that would be identified as potential outliers (40, 40, 40, 40, 41, 43,

44, 54, 58). Since $Q_1 - 1.5 \times IQR = 19 - 1.5 \times 8 = 7$, there are no potential outliers on the low end of the distribution.

2.10 (a) $\overline{x} = (7.1 + 11.6 + 8.1 + 13.4)/4 = 40.2/4 = 10.05$ picocuries. **(b)** The standard deviation can be computed in steps:

x	7.1	11.6	8.1	13.4	Sum
$x - \overline{x}$	−2.95	1.55	−1.95	3.35	0
$(x - \overline{x})^2$	8.7025	2.4025	3.8025	11.2225	26.13

$s^2 = \frac{1}{n-1}\sum(x - \overline{x})^2 = \frac{1}{4-1}(26.13) = 8.71$. So $s = \sqrt{s^2} = \sqrt{8.71} = 2.95$ picocuries.

2.11 Both data sets have the same mean and standard deviation (about 7.5 and 2.0, respectively). However, simple stemplots (provided, with the data rounded to the nearest tenth) reveal that Data A has a very left-skewed distribution, while Data B has a slightly right-skewed distribution with a high outlier.

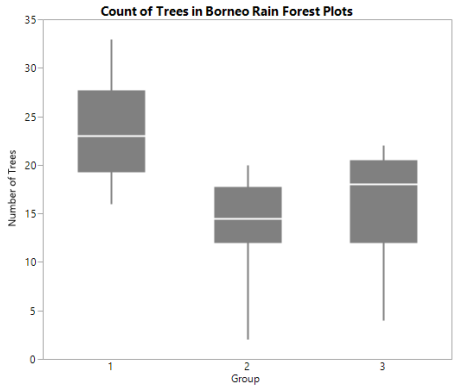
A		B
1	3	
7	4	
	5	568
1	6	69
3	7	079
8711	8	58
311	9	
	10	
	11	
	12	5

2.12 (a) No; the distribution isn't symmetric. **(b)** Yes; the distribution is symmetric and mound-shaped with no severe outliers. **(c)** No; the distribution is strongly right-skewed.

2.13 STATE: We'd like to know how logging impacts how many trees there are in 0.1 hectare plots in the rainforests of Borneo. PLAN: We'll create side-by-side boxplots for the three types of plots and compute appropriate summary statistics. SOLVE: According to the boxplots, none of the distributions are symmetric; Group 2 (logged one year earlier) has a low outlier and Group 3 (logged eight years earlier) is clearly left-skewed, while Group 1 (never logged) appears to be right-skewed. (Note: If modified boxplots are constructed, the outlier in Group 2 will be apparent.) Because of the non-symmetric shapes, we will compute the five-number summaries for each.

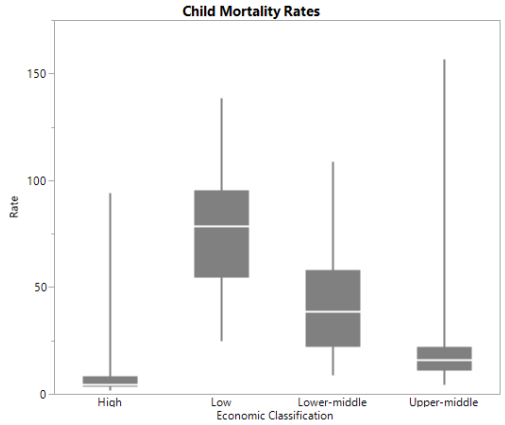
	Min	Q_1	M	Q_3	Max
Group 1 (never logged)	16	19.5	23	27.5	33
Group 2 (logged 1 year earlier)	2	12	14.5	17.5	20
Group 3 (logged 8 years earlier)	4	12	18	20.5	22

(If you compute the means and standard deviations, they are: Group 1: $\overline{x} = 23.75$, s = 5.07; Group 2: $\overline{x} = 14.08$, s = 4.98; and Group 3: $\overline{x} = 15.78$, s = 5.76.) CONCLUDE: It is clear from the boxplots and summary statistics that plots that have never been logged have more trees than either type of logged plot. Further, if we compare the distributions and summary statistics for the two different types of logged plots, we see it takes a long time for the rainforest to recover from having been logged; while the centers of the distribution for plots logged 8 years earlier indicate more trees per plot on average, the distribution of the number of trees for plots logged 8 years earlier had more variability.



2.14 STATE: Is child mortality rate related to the country's economic wealth? PLAN: Create side-by-side boxplots to compare the distributions for the various economy classifications. SOLVE: From the boxplots and five-number summary, we see that countries with greater economic wealth have lower rates of child mortality. The five-number summary was computed using technology. Q_1 and Q_3 will differ using the "by-hand" approach. All are right-skewed distributions. There are a few high outliers for countries with high and upper-middle wealth. With high wealth, the outliers correspond to Equatorial Guinea (94.1) and Trinidad and Tobago (20.4). Among upper-middle wealth countries, outliers correspond to Angola (156.9), Turkmenistan (51.4), Gabon (50.8), Namibia (45.4), Botswana (43.6), and South Africa (40.5). CONCLUDE: Economic classification does a good job of explaining differences in child mortality.

	Min	Q_1	M	Q_3	Max
High	1.9	3.475	4.45	8.225	94.1
Upper-middle	4.6	11.275	16.15	22.4	156.9
Lower-middle	9	22	38.75	58.15	108.8
Low	24.9	54.6	78.4	95.5	138.7



2.15 (a) 281.6.

2.16 (c) 285.5.

2.17 (b) 254.

2.18 (a) the mean is pulled toward the longer end of the distribution.

2.19 (c) 75%. Q_1 has 25% of observations equal to or less than its value, which implies 75% are greater than that value.

2.20 (c) the five-number summary.

2.21 (b) 28.6.

2.22 (a) $0 \leq s$.

2.23 (b) pounds.

2.24 (b) The mean

2.25 The distribution of incomes in this group is almost certainly right-skewed, so the mean is $51,754 and the median is $44,167.

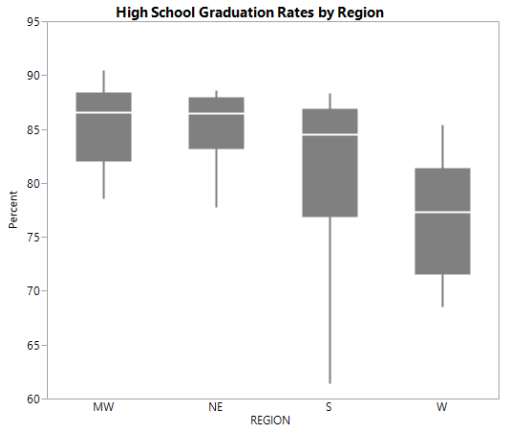
2.26 (a) The distribution of the value of transaction accounts is a highly right-skewed distribution, with a small percentage of accounts having extremely high values. The distribution being so right-skewed explains the mean being that much larger than the median. **(b)** A median of $0 says that at least half of households do not have a retirement account.

2.27 With 841 colleges (an odd number), the median location is (841 + 1)/2 = 421, so the median is the 421st ordered endowment. The first quartile, Q_1, is found by taking the median of the first 420 sorted endowments. This would be the (420 + 1)/2 = 210.5th endowment. Similarly, Q_3 is the 421 + 210.5 = 631.5th endowment.

2.28 (a) Min = 23.0 thousand pounds (23,000 as rounded), Q_1 = 30.35 thousand pounds (30,350 pounds), median = 31.95 thousand pounds, Q_3 = 32.7 thousand pounds, max = 33.7 thousand pounds. **(b)** Notice that the minimum is much farther from Q_1 (7.35 thousand pounds) than the maximum is from Q_3 (1 thousand pounds). This suggests a long left tail, consistent with a left-skewed distribution.

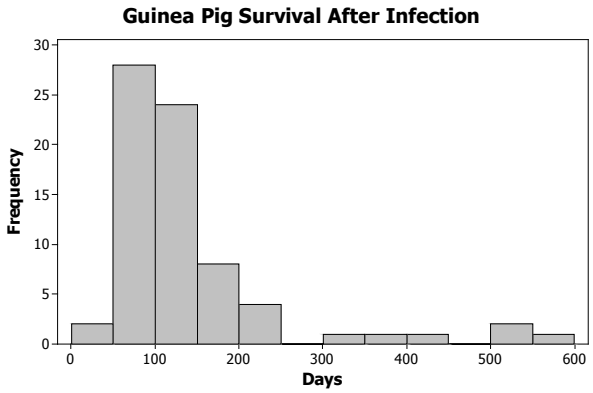
2.29 The boxplots do not reveal the gap in the South between the rates for Georgia and the District of Columbia, making the stemplot more useful for comparing regions. If you construct a modified boxplot, then you see the gap, and the boxplots essentially show the same information as the stemplot.

	Min	Q_1	M	Q_3	Max
MW	78.6	82.25	86.6	88.25	90.5
NE	77.8	83.15	86.5	87.95	88.6
S	61.4	76.85	84.5	86.95	88.3
W	68.5	71.55	77.3	81.4	85.4



2.30 There are $n = 74$ observations represented in the histogram; the median will be at position $(74 + 1)/2 = 37.5$ and the quartiles at position $(37 + 1)/2 = 19$ in each half. **(a)** Median = 2, $Q_1 = 1$, and $Q_3 = 4$. **(b)** $\overline{x}$ = [(15)(0) + (11)(1) + (15)(2) + (11)(3) + (8)(4) + (5)(5) + (3)(6) + (3)(7) + (3)(8)]/74 = 194/74 = 2.62 servings. This is larger than the median because the distribution is right-skewed. **(c)** The reason we can do so in this example is that we know all of the observations in the first bin correspond to 0 servings, the second bin is 1 serving, and so on. So we know exactly how many of each value (0, 1, …, 8) were observed.

2.31 (a) A histogram of the survival times is given. The distribution is strongly right-skewed, with the center around 100 days and a range from about 0 days to about 600 days.



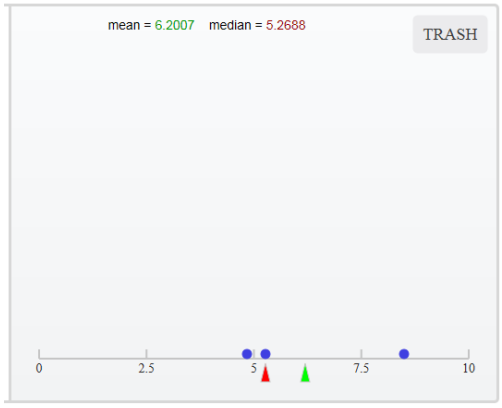
(b) Because of the extreme right-skew, we should use the five-number summary: 43, 82.5, 102.5, 151.5, 598 days. Notice that the median is closer to Q_1 than to Q_3.

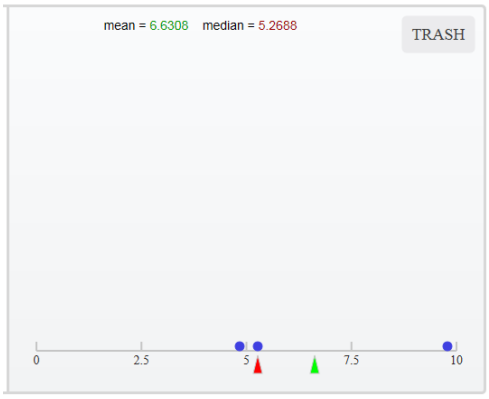
2.32 (a) If different years and countries have very different numbers of babies born, the distributions across age group will be difficult to compare because the scale will be determined by the larger counts. This makes it hard to visualize trends in countries with small counts, since all of those bars will appear small. **(b)** The total count in the data set is 1,550,274 births. Note that this ignores births for mothers under age 10 or above age 50, and it doesn't account for multiples (such as twins), so this is an undercount of the number of babies. **(c)** Draw a histogram by hand. Most software cannot be used since the data are aggregated. The distribution is right-skewed. **(d)** Using the total number of births from part (b), the median is the (1550274 + 1)/2 = 775137.5th ordered age. Thus, the median is between 25 and 29 years. The first quartile is the (775137 + 1)/2 = 387569th ordered age, so Q_1 is between 20 and 24 years. The third quartile is the 775137 + 387569 = 1162706th age, so Q_3 is between 30 and 34 years.

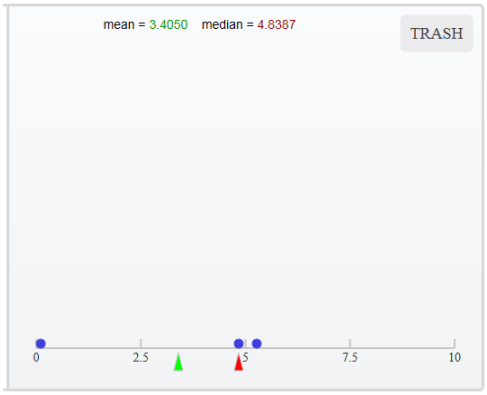
2.33 (a) Symmetric distributions are best summarized using $\overline{x}$ and s. The distribution for the treatment group was right-skewed. The control distribution could be called rather symmetric, but it has a high outlier. **(b)** Removing the outlier reduces all three statistics (there is one less observation). However, the mean decreased by 8.45 seconds, which is about double the decrease in the median (3.5 seconds). **(c)** The median decreased by 3.5. The median is less impacted by the outlier than the mean.

	With outlier	Without outlier
Mean	59.7	51.25
Standard deviation	63.0	50.97
Median	61	57.5

2.34 (a) The mean (green arrow) moves along with moving point. The median (red arrow) points to middle point (rightmost nonmoving point).

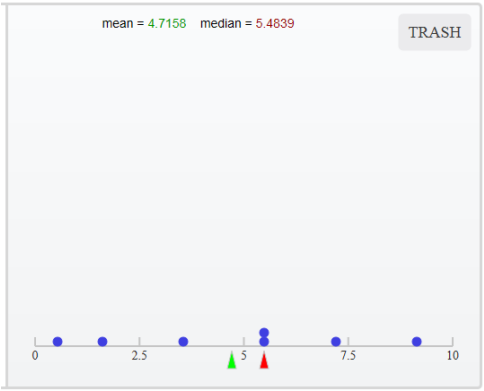






(b) The mean follows the moving point. When the moving point passes the rightmost fixed point (and moves to its left), the median moves with that point until it passes the leftmost fixed point—then the median stays there.

2.35 (a) The sixth observation must be placed at median for the original five observations. **(b)** No matter where you put the seventh observation, the median is one of the two repeated values above, because it will be the fourth (ordered) observation. The author's seventh point was the one at the extreme left.



2.36 Both distributions are very similar: On weekdays more babies are born, and there is variation from weekday to weekday, though Mondays appear to have slightly fewer births. On weekends, fewer births take place. Of course, many more births take place in the United States.

2.37 The mean for all 51 entries is 33.9%, far from the national percentage of 42.8%. You can't average averages. Some states, such as California and Florida, are larger and should carry more weight in the national percentage. In this case, the larger states also have a larger percentage of minority residents. By not accounting for population size, averaging the averages results in a value lower than the national percentage.

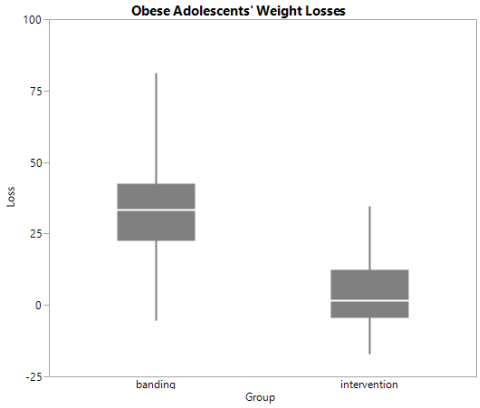
2.38 Answers will vary, but a raise in the minimum wage will probably have a greater impact on the median income. Most Americans earn "middle income" or less; a few people earn huge amounts each year. The few large amounts will still pull the mean toward that end of the distribution.

2.39 (a) The smallest possible standard deviation will come from choosing all four numbers to be the same; for example, choose the numbers (2, 2, 2, 2). **(b)** The largest possible standard deviation is with the four numbers (0, 0, 10, 10). **(c)** There is more than one choice in part (a) but not in part (b).

2.40 Many answers are possible. Start by ensuring that the median is 12 by "locking" 12 as the fourth smallest value. We also have 4 specified as the minimum and 19 as the maximum, so the seven numbers must be 4, ___, ___, 12, ___, ___, 19. With three numbers on either side of the median, the quartiles will be in positions 2 and 6. One set that works is 4, 8, 9, 12, 14, 15, 19.

2.41 Many answers are possible. One solution: (1, 2, 3, 4, 5, 6, 100). In general, a "large" high outlier will guarantee this condition.

2.42 (a) Weight losses that are negative correspond to weight *gains*. **(b)** A side-by-side boxplot is provided. Gastric banding seems to produce higher weight losses, typically. Because both distributions are somewhat right-skewed (and there is a high outlier in the banding group), the five-number summary would be appropriate. The summary statistics are given below. There is a high outlier of 81.4 for the gastric banding group. (Using technology, values of the quartiles may differ).

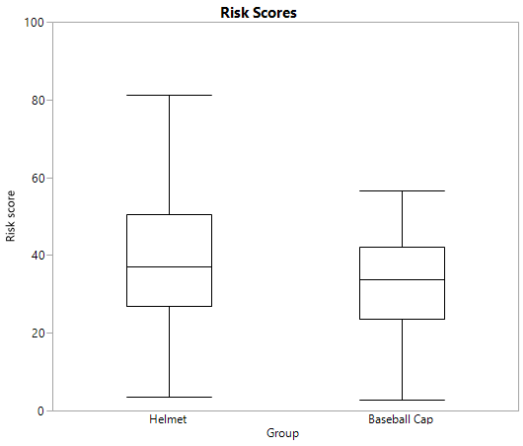


(c) It's better to measure weight loss relative to initial weight. A loss of 5 kg would not mean the same if individuals started at different weights. Percent of excess weight lost would be a good measure. *Percent* reduction in BMI would also be good. **(d)** If the subjects that dropped out had continued, the difference between these groups would be as great or greater because many of the "lifestyle" dropouts had negative weight losses (weight gains), which would pull that group down.

	Min	Q_1	M	Q_3	Max
Banding	−5.40	23.4	33.35	42.35	81.40
Intervention	−17.00	−4.3	−0.20	11.6	34.60

2.43 STATE: We want to determine how wearing a helmet relates to the measure of risk behavior. PLAN: We will make side-by-side boxplots and compute five-number summaries of the scores for the helmet and baseball cap groups. We will compare the distributions for each group to make a conclusion on how wearing a helmet relates to the average number of pumps (risk taking). SOLVE: The boxplots are provided. The summaries were computed using technology. Values of Q_1 and Q_2 may differ using the "by-hand" approach given in this chapter. The minimum, first quartile, and median of the helmet group is only slightly larger than of the baseball cap group. There is a greater discrepancy between the third quartile and maximum, with those being much larger for the helmet group than for the baseball cap group. There is greater variability in the helmet group. CONCLUDE: Wearing a helmet appears to be related to risk behavior. Although it didn't increase the behavior by much for most individuals, some helmet-wearing individuals displayed much riskier behavior.

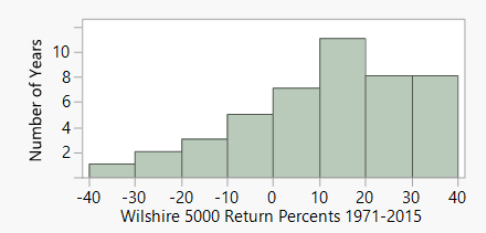
	Min	Q_1	M	Q_3	Max
Helmet	3.67	27.015	37	50.2	81.29
Baseball cap	2.68	23.935	33.635	42.27	56.58

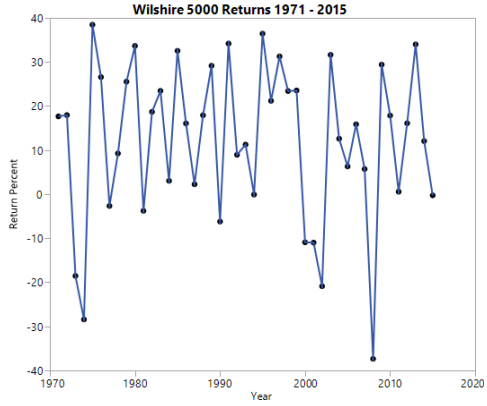


2.44 STATE: We'd like a description of the 2015–2016 Canadiens' salaries for management; are there any interesting features in the distribution? PLAN: We'll graph the data with a stemplot after rounding to the nearest $100,000. Based on the shape of that graph, we'll compute appropriate summary statistics (based on the actual salaries). SOLVE: The stemplot shows a right-skewed distribution. The median salary is $1,000,000 (while the mean is $2,359,354.8, consistent with a strong right-skew). The middle half of players earn between $800,000 and $3,900,000, although six players earn at least $5,000,000. CONCLUDE: The team salaries range from about $600,000 to about $7 million. The median salary is $1 million. There must be some differential in pay for team "stars," because more than half the players earn less than $2 million per season, whereas a second group earns between $2 million and $4 million. A third group of players makes $5 million or more.

```
0 | 66667778889999
1 | 0003
1 |
2 | 3
2 | 558
3 |
3 | 59
4 | 0
4 |
5 | 0
5 | 5
6 | 0
6 | 5
7 | 00
```

2.45 STATE: We'd like to describe the distribution of Wilshire 5000 stock index returns over the period from 1971 through 2015. PLAN: We'll graph the return with a histogram and a time plot. Based on what is seen there, we'll compute and report appropriate summary statistics. SOLVE: The histogram, time plot, and summary statistics are given. CONCLUDE: The distribution of average returns is left-skewed. Most years, the average return is positive. Returns range from about –40% to about 40%, with the median return about 16%.



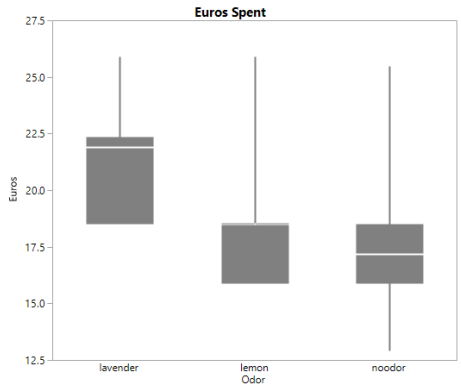


Mean	St. Dev.	Minimum	Q_1	Median	Q_3	Maximum
12.12	17.99	-37.34	0.265	16.09	26.075	38.47

2.46 STATE: Does the presence of a lavender or a lemon odor in a pizza restaurant lead to customers spending more? PLAN: We'll compare side-by-side boxplots for the distributions of amount spent (in euros), as well as compute appropriate summary statistics. SOLVE: Boxplots are given. The summary statistics are given in the table.

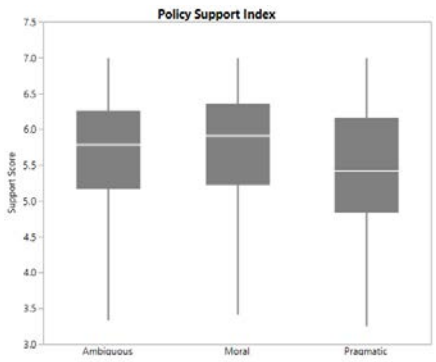
	Mean	St. Dev.	Minimum	Q_1	Median	Q_3	Maximum
Lavender	21.123	2.345	18.5	18.5	21.9	22.35	25.9
Lemon	18.157	2.218	15.9	15.9	18.5	18.5	25.9
No odor	17.513	2.359	12.9	15.9	17.2	18.5	25.5

CONCLUDE: All three distributions are right-skewed. We note that the minimum and Q_1 are equal for both the lavender and lemon odors. With the lemon odor, the median is equal to Q_3. Both the lemon and control (no odor) have high outliers. Because of the shapes, using the five-number summary to describe these distributions is more appropriate than the mean and standard deviation. Lavender seems to produce the highest customer expenditures; its median is 21.9 euros, which is above Q_3 (18.5 euros) for both other conditions.



2.47 STATE: We want to know how a leader's justification affects support for the policy. PLAN: Create side-by-side boxplots of the support index for the three different justifications and compute summary statistics. SOLVE: Side-by-side boxplots and summary statistics are given. The distributions are very similar, with a pragmatic approach yielding slightly less support. All three distributions are left-skewed. If modified boxplots are constructed, we see a single low outlier for both ambiguous and moral. CONCLUDE: An ambiguous or moral justification of a policy tends to have slightly more support than a pragmatic justification. There is little difference between ambiguous and moral justifications.

	Mean	St. Dev.	Minimum	Q_1	Median	Q_3	Maximum
Ambiguous	5.71	0.77	3.33	5.17	5.79	6.25	7
Moral	5.81	0.82	3.42	5.25	5.92	6.5	7
Pragmatic	5.44	0.89	3.25	4.83	5.42	6.17	7

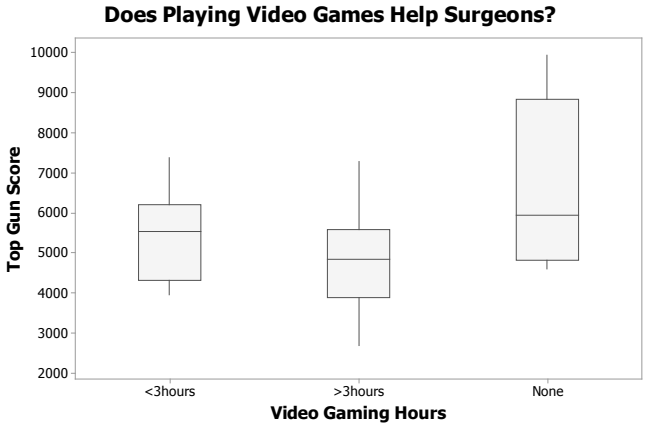


2.48 STATE: Does playing video games improve the skills needed by a surgeon for laparoscopic surgery? PLAN: We will compare the Top Gun scores for three groups of surgeons, categorized by their video gaming hours per day at the height of their video game use. SOLVE: Boxplots that display the distributions are given. The boxplots for those who played video games are close to symmetric, while the boxplot for surgeons who never played video games is right-skewed (indicating poorer performance, since low scores are better). Descriptive statistics from Minitab follow.

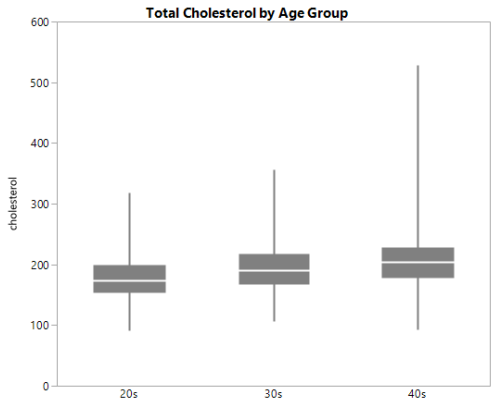
Descriptive Statistics: topgun

category	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
<3hours	9	5420	1106	3968	4308	5540	6204	7367
>3hours	9	4787	1313	2703	3884	4845	5596	7288
None	15	6793	1947	4605	4828	5947	8837	9930

CONCLUDE: Based on the side-by-side boxplots and summary statistics, surgeons who have done a lot of video gaming do better at the Top Gun program. Their median score is more than 600 points lower than those who played less than three hours, and more than 1100 points lower than the median for those who never played video games. The most consistent group (as measured by the standard deviation and range) was the group who played some video games, but not intensely; the least consistent group was those who never played video games.



2.49 (a) With three age groups to compare and a large data set, we'll do side-by-side boxplots. All three distributions are right-skewed with high outliers. The median increases slightly with increasing age (from 173 to 190 to 204). We also see an increase in variability, which is mostly due to the outliers as people age, although the *IQR*s are relatively the same.



(b) We'll note here that 25% or more of the individuals in each age group had total cholesterol levels above 200 (Q_3 = 199 for the people in their 20's, and 218 and 229 for the other two). Unless their original cholesterol levels were *extremely* high, the 4 or 24 people on medication in their 20s and 30s, respectively, probably wouldn't affect these distributions a great deal because there were roughly 950 people in each group (and these become small fractions of the total). However, there were 1139 people in their 40s; 117 of those on medication is more than 10% of this group. If those 117 had not been on medication, that distribution would likely show more variability and higher cholesterol readings (that might make the box longer, for example).

2.50 (a) The five-number summary is 8, 22, 31, 48, 75. **(b)** The $1.5 \times IQR$ rule says that a high outlier is any observation larger than $48 + 1.5(48 - 22) = 87$. By this rule, there are no high outliers.

2.51 (a) With all observations, the mean and median of the bonding scores for the pain group are 3.71 and 4, respectively. With the two smallest observations omitted, the mean and median are 3.9 and 4.14. Omitting the two smallest observations had a greater impact on the mean than the median. The median is more robust to a small number of unusually small or large values. **(b)** The $1.5 \times IQR$ rule identifies low outliers as being smaller than $Q_1 - 1.5(Q_3 - Q_1) = 3.43 - 1.5(4.43 - 3.43) = 1.93$. Yes, the rule does identify these two scores as suspected outliers. **(c)** It is reasonable that there exist subjects who experience little bonding regardless of the group they are in. It is possible that, after randomization, subjects of this sort were assigned to the pain group. In this instance, these will appear as "outliers."

2.52 (a) Min = 139.4, Q_1 = 149.4, median = 182, Q_3 = 286.6, and max = 485.7. Notice that the maximum is much farther from Q_3 than the minimum is from Q_1. This suggests right-skew. **(b)** $IQR = 286.6 - 149.4 = 137.2$. $1.5 \times IQR = 205.8$. Now $Q_1 - 1.5 \times IQR = -56.4 < 0$, so there are no low outliers. Also, $Q_3 + 1.5 \times IQR = 286.6 + 205.8 = 392.4$. This is larger than the maximum value, so there are no high outliers according to this rule. A stemplot is provided. I agree that there are no high outliers. Rather, it appears we may have a bimodal distribution. There is a group of observations with revenues between \$140 billion and \$300 billion, and another group with revenues between \$350 billion and \$500 billion. **(c)** Opinions will vary. The 30 companies account for 6903.3/31200 = 0.2213, or about 23% of total Global 500 revenues.

Stem	Leaf	Count
4	59	2
4	33	2
3	68	2
3	4	1
2	57	2
2	0012	4
1	555566667889	12
1	44444	5

2.53 The five-number summary of cholesterol levels for people in their 20s is 92, 154, 173, 199, 318. We have IQR = 199 − 154 = 45. Outliers would be values smaller than $154 - 1.5 \times 45 = 86.5$ or larger than $199 + 1.5 \times 45 = 266.5$. Using this criterion, there are no low-end outliers, but there are high-end outliers (we saw these in the boxplots in the solution to Exercise 2.49).

2.54 and **2.55** are Web-based exercises.