Chapter 1

Data Collection and Exploring Univariate Distributions

1.2 Types of Data and Frequency Distribution Tables

1.1 **a** Qualitative
b Quantitative
c Quantitative
d Quantitative
e Qualitative
f Quantitative
g Quantitative
h Qualitative
i Quantitative
j Qualitative

1.2 **a** Percent deviation in ozone levels (Quantitative)
Square miles of ozone hole size (Quantitative)

b Incidence of kidney failure (Qualitative)
Amount of blood loss (Quantitative)
Length of recovery period (Quantitative)
Incidence of complications (Qualitative)
Incidence of side effects (Qualitative)

c Amount of damage (Quantitative)
Type of damage (Qualitative)
Insurance status (Qualitative)

1.3 **a**

	Years of Formal Education				
Income Category	12	14	16	18	20 +
$0-29,999$	0.620	0.473	0.323	0.220	0.148
$30,000-59,999$	0.313	0.408	0.397	0.417	0.298
$60,000-89,999$	0.054	0.093	0.174	0.227	0.270
$90,000$ and above	0.014	0.025	0.106	0.136	0.284

b The relative frequency of higher-income categories increased with the increasing number of years of formal education

1.4

Capital Value (£million)	Number of Projects Completed	Relative Freq.	Cumulative Relative Freq.
Less than 50	145	0.580	0.580
50 – 100	58	0.232	0.812
101 – 150	20	0.080	0.892
151 – 200	10	0.040	0.932
201 – 250	5	0.020	0.952
251 – 300	3	0.012	0.964
301 – 350	2	0.008	0.972
351 – 400	1	0.004	0.976
401 – 450	3	0.012	0.988
451 – 500	2	0.008	0.996
501 – 700	1	0.004	1.000
	Total= 250		

a 95.20%

b 100% − 98.80% = 1.2%

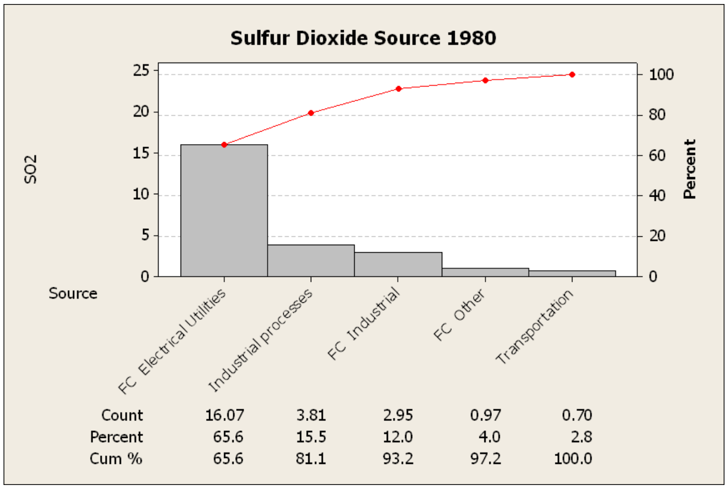
c 99.6% − 96.4% = 3.2%

1.3 Tools for Describing Data: Graphical Methods

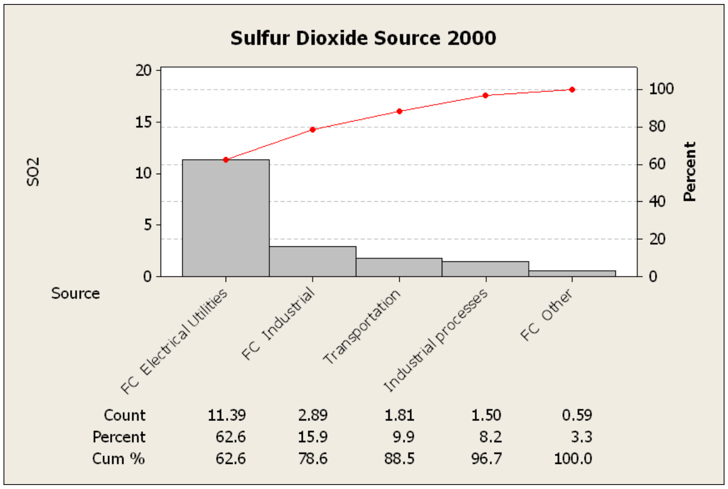
1.5 **a** Complaints 5, 12, and 10 (cleaning of public highways, working hours, and screening/fencing respectively) each comprised at least 10% of the total number of complaints.

b Complaints 5, 12, 10, 4, 7, 8, 9, and 1 (cleaning of public highways, working hours, screening/fencing, water courses affected by construction, blue routes and restricted times of use, temporary and permanent diversions, TMP, and property damage) comprise a cumulative total of 80% of the complaints.

1.6 **a** Pareto charts for SO_2 sources in 1980 and 2000:

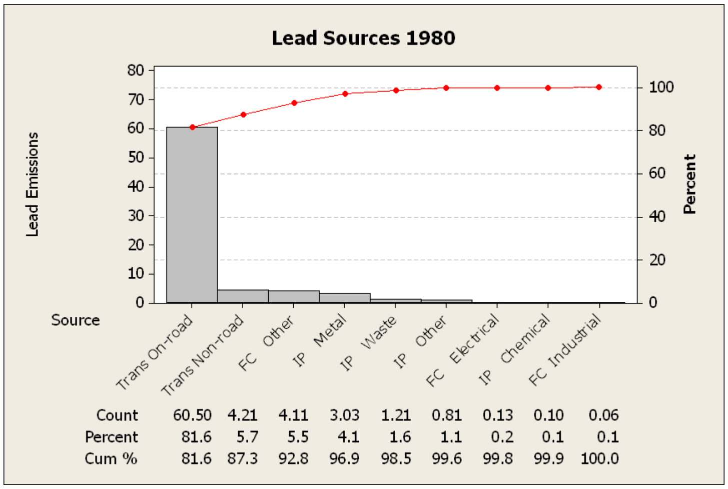


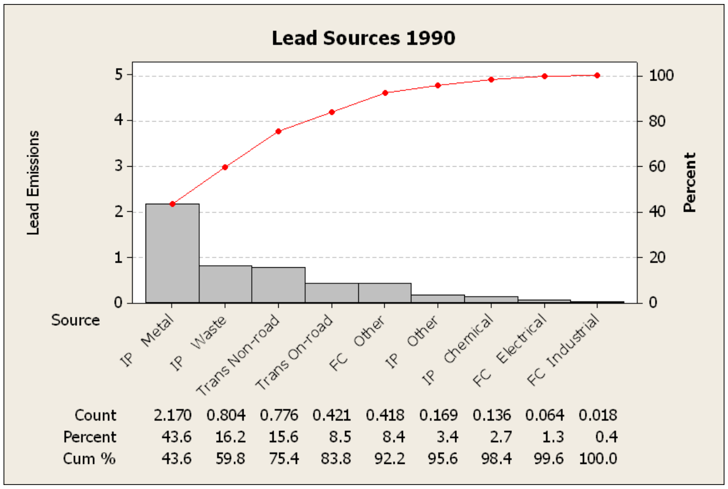
Count	16.07	3.81	2.95	0.97	0.70
Percent	65.6	15.5	12.0	4.0	2.8
Cum %	65.6	81.1	93.2	97.2	100.0



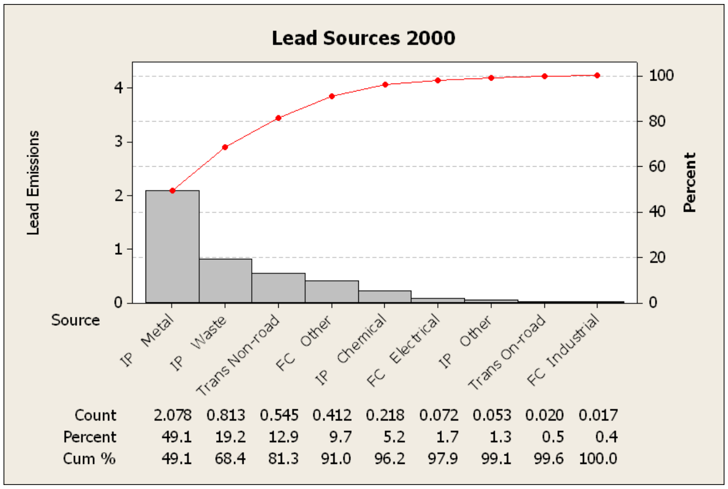
b Industrial processes have a decreased SO_2 contribution from 1980 to 2000, while trasportaion has an increased SO_2 contribution.

1.7 **a** Pareto charts for lead pollution sources in 1980, 1990 and 2000:





Source	IP Metal	IP Waste	Trans Non-road	Trans On-road	FC Other	IP Other	IP Chemical	FC Electrical	FC Industrial
Count	2.170	0.804	0.776	0.421	0.418	0.169	0.136	0.064	0.018
Percent	43.6	16.2	15.6	8.5	8.4	3.4	2.7	1.3	0.4
Cum %	43.6	59.8	75.4	83.8	92.2	95.6	98.4	99.6	100.0

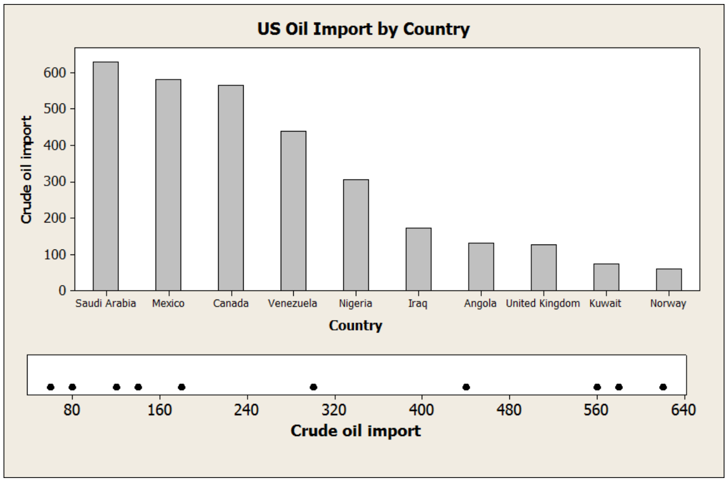


Source	IP Metal	IP Waste	Trans Non-road	FC Other	IP Chemical	FC Electrical	IP Other	Trans On-road	FC Industrial
Count	2.078	0.813	0.545	0.412	0.218	0.072	0.053	0.020	0.017
Percent	49.1	19.2	12.9	9.7	5.2	1.7	1.3	0.5	0.4
Cum %	49.1	68.4	81.3	91.0	96.2	97.9	99.1	99.6	100.0

b Lead emissions seem to have decreased since 1980, especially in the areas of transportaion and miscellaneous fuel combustion sources.

c The evidence seems to suggest that we are releasing lead pollutants into our environment at a decreased rate since 1980.

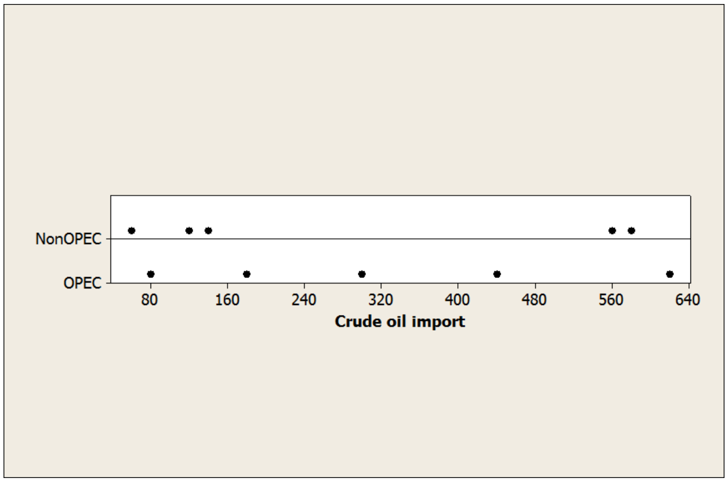
1.8 **a** Bar chart and dot plot of top 10 suppliers of US crude oil:



US crude oil import ranged from 60 to 629 million barrels. Saudi Arabia, Venezuela, Mexico and Canada are the largest exporters.

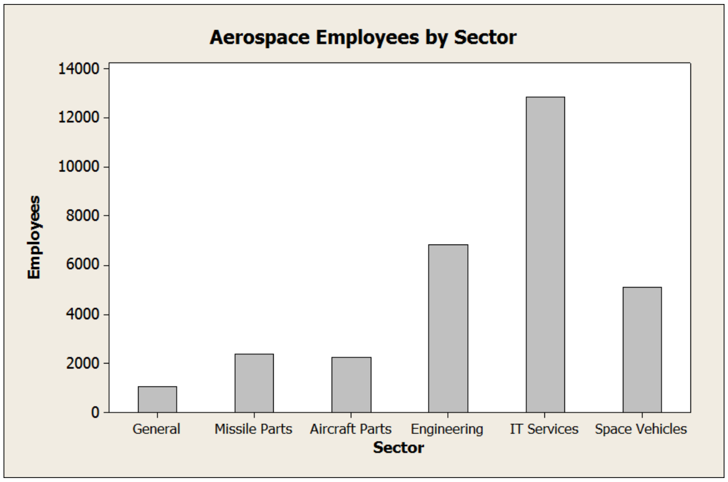
b The bar chart specifies which country each figure comes from, the dot plot merely gives the numbers.

c Bar Chart of OPEC vs. Non-OPEC suppliers:



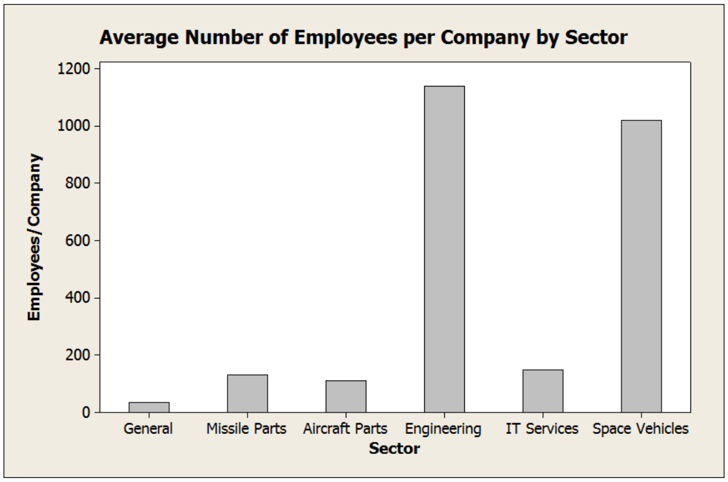
In general, OPEC countries supplied more oil to the US than non-OPEC countries.

1.9 **a** Bar Chart of Alabama Aerospace Employment:



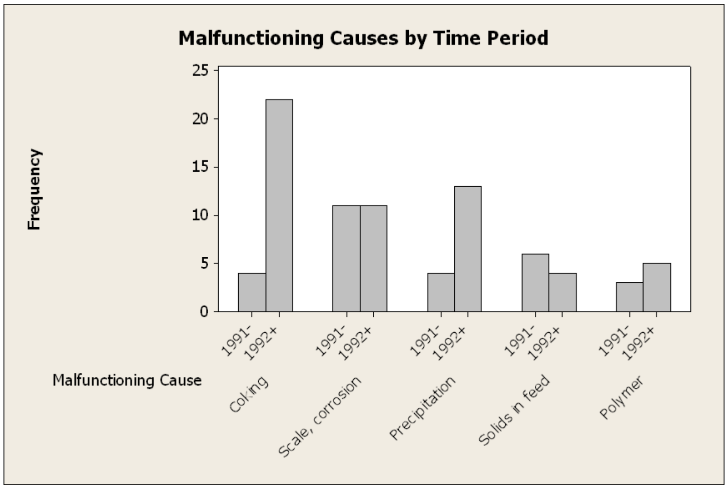
The largest number were employed by the information technology services, followed by engineering and RFD services, and missile space vehicle manufacturing.

b Bar Chart of number of employees per company amongst Alabama Aerospace fields:



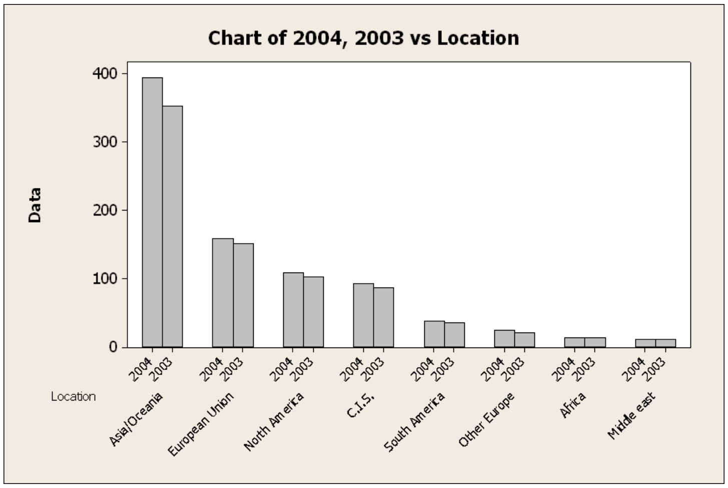
Although information technology services employed the largest number of employees, they were not, on average, large employers. Engineering RFD services and missile space vehicle manufacturing employed fewer people than the information technology services, yet they employed far more people on average per company.

1.10 **a** Bar chart of causes of distillation tower malfunction:



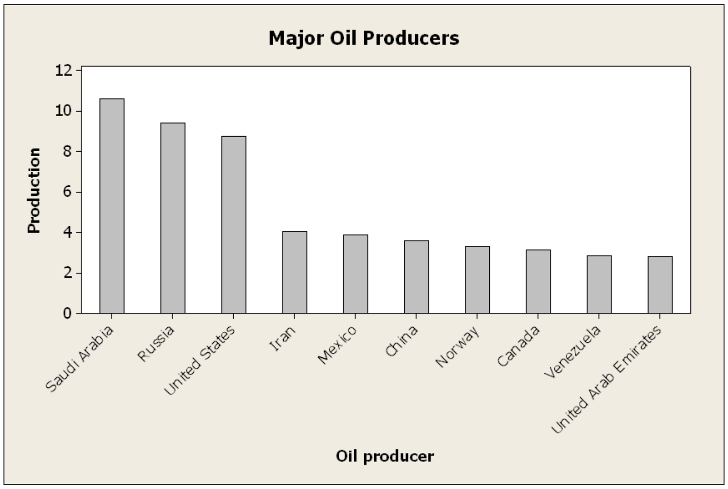
b Prior to 1991 scale and corrosion was a major cause of tower malfunction. Coking and precipitation have become much more prevalent causes of distillation tower malfunction since 1991.

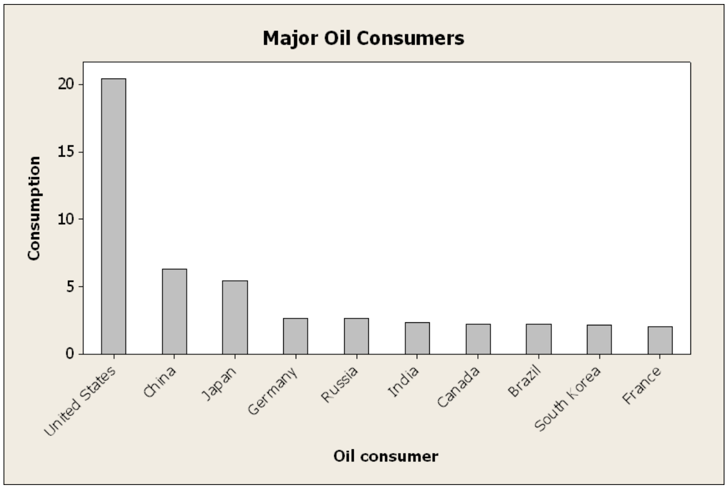
1.11 **a** Bar chart of crude Steel production by region in 2004 and 2003:



b In general, the crude oil production has increased from 2003 to 2004.

1.12 **a** Bar charts of oil consumption and oil production in millions of barrels per day:

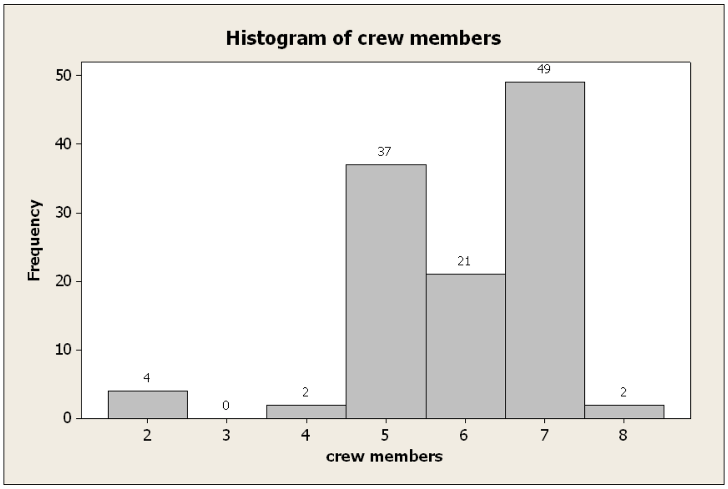




b Several countries consume more oil than they produce. The United States consumes over three times as much as any other nation and almost 12 million more barrels daily than it produces.

1.13 The data are grouped together in a histogram, losing identity of individual observations, which are still retained by dotplot. A small number of observations makes it difficult to notice any patterns. Gaps in the data are visible from a dotplot but are not identified from a histogram.

1.14 **a** Histogram for numbers of crew members on orbiter missions:



b $21 + 49 + 2 = 72$ missions

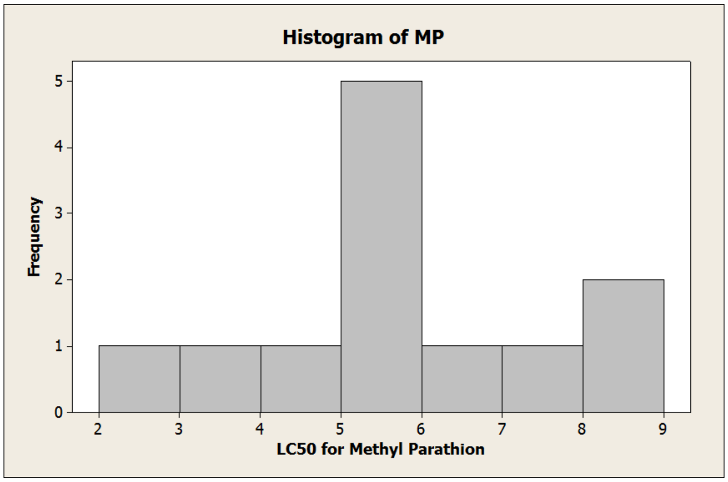
c $\dfrac{4+0}{4+0+2+37+21+49+2} = .035 = 3.5\%$

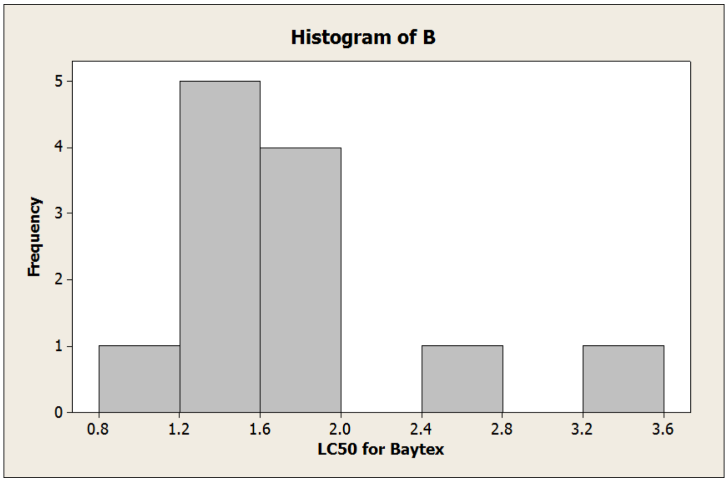
d The average number of crew per flight seems to have increased slightly since 1981

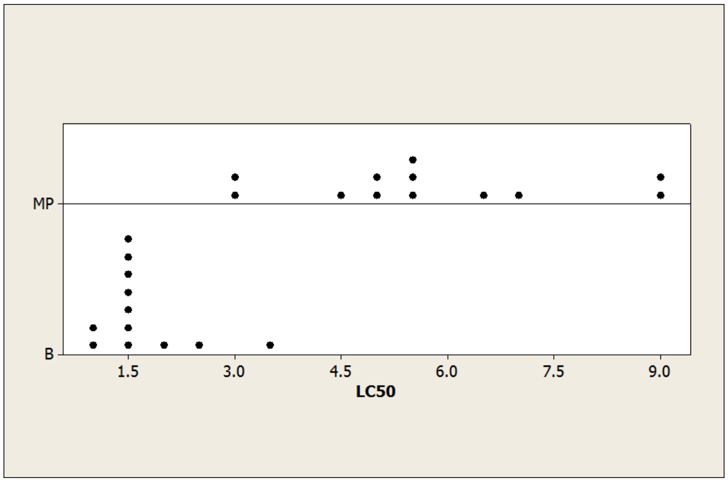
1.15 **a** $\dfrac{0+27+12+0}{500} = .078 = 7.8\%$

b There were no rods of length .999, and an abnormally large amount of rods of length 1.000. This may indicate that someone may have been inappropriately placing .999 rods into the 1.000 category to prevent them from being declared defective.

1.16 **a** Histograms and dotplots for LC50 of Methyl Parathion and Baytex in water samples:







b The LC50 distribution for Methyl Parathion seems to be more or less symmetrical, the distribution for Baytex seems skewed to the right. There also seems to be much more variability in the distribution of LC50s for Methyl Parathion.

1.17 **a** Yes, in 1890, most of the population was in the younger age ranges. In 2005, a larger percentage of the population are in the upper age ranges. This might suggest that there have been some sort of medical advances to improve life expectancy and quality of life over time.

b Percent of population under 30 in 1890
$= 25\% + 22\% + 18\% = 65\%$.
Percent of population under 30 in 2005
$= 13.3\% + 14.5\% + 13.4\% = 41.2\%$

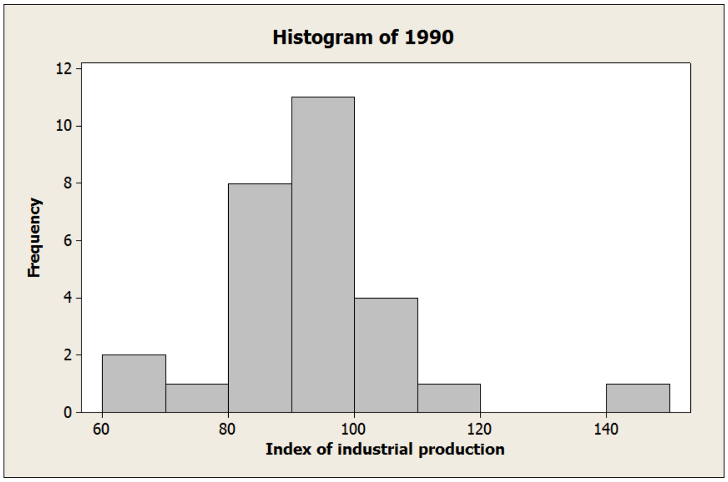
c The percentage of older population has increased. In 1890, the percentage of population in different age categories decreased steadily with the increasing age. In 2005, it is fairly evenly distributed across different age groups except for the two oldest age groups.

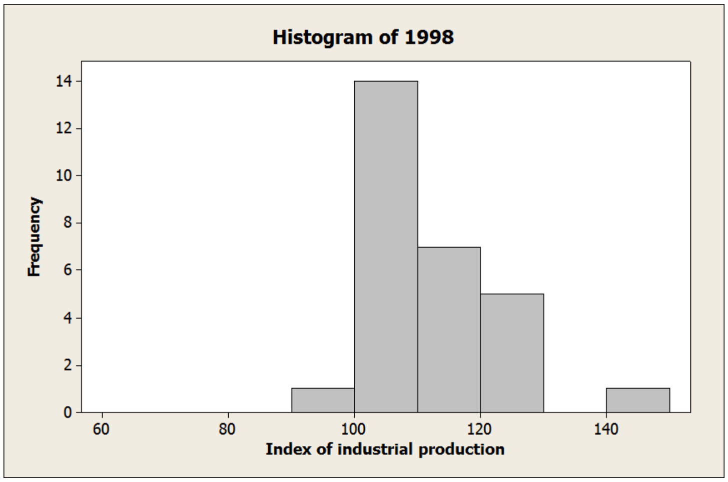
1.18 **a** Yes

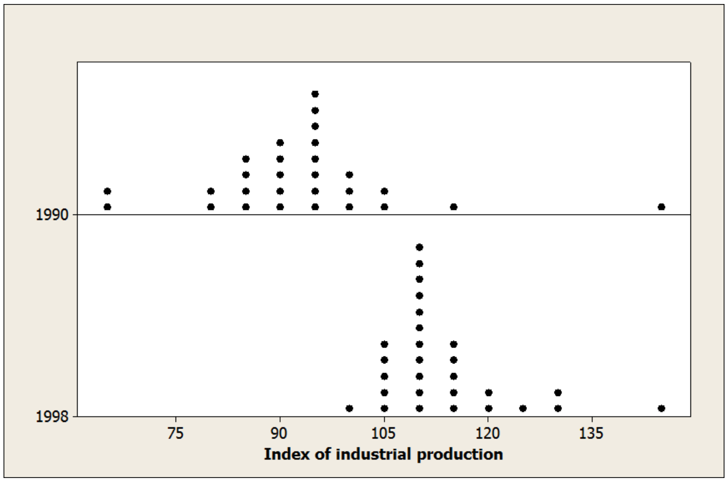
b The distribution of desired work start times is more spread out than the arrival times, and much more spread out than the official start times.

c This plot shows that when start times are staggered throughout the morning, workers' official start times tend to bunch up around 8am.

1.19 **a** Histograms and dotplots displaying indices of industrial production in 1990 and 1998:

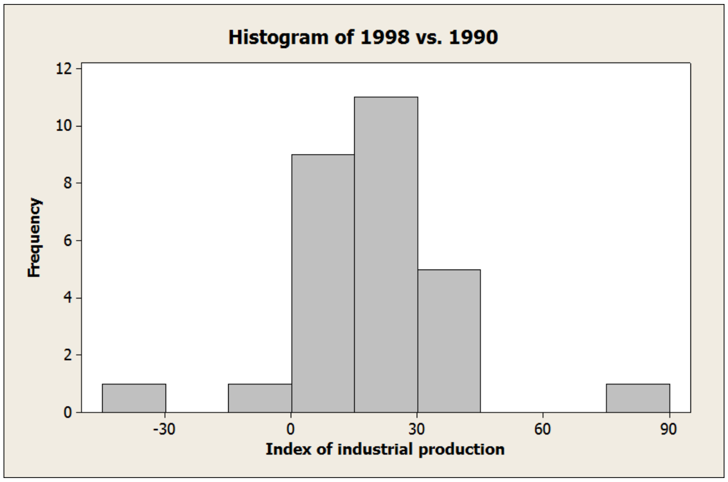


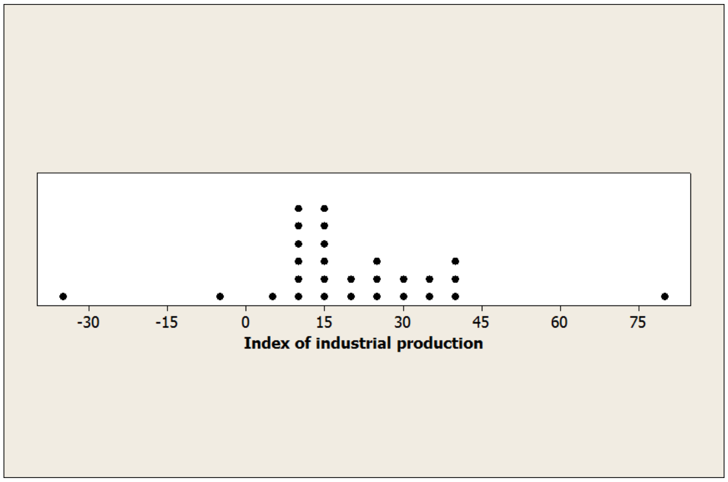




There is a considerable increase in average index from 1990 to 1998, indicating an increase in industrial product by most of the countries in general. The indexes were more spread out in 1990 compared to 1998. The distribution is right-skewed in 1998. There might be an outlier on the upper end in 1990.

b Histogram and dotplot of the difference in production from 1990 to 1998:





Most countries showed improvement (an increase of up to 40 points) in the industrial production; one country in particular showed a tremendous amount of improvement. Only two countries showed a decrease.

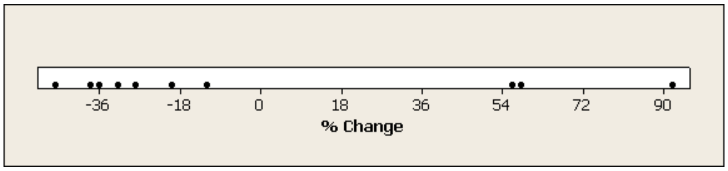
1.4 Tools for Describing Data: Numerical Measures

1.20 **a** In 1890, $24 + 22 = 46 < 50$ and $24 + 22 + 18 = 64 > 50$, so the median is in the age range $20 - 29$.

b In 2010, $13.2 + 13.9 + 13.7 = 40.8 < 50$ and
$13.2 + 13.9 + 13.7 + 12.7 = 53.5 > 50$ so the median is in the age range $30 - 39$

c The median age in 2010 is greater than the median age in 1890, indicating that more people fall into upper age ranges in 2010 than in 1890.

1.21 **a** Dotplot for percentage change in crude oil import:



b

$$\bar{x} = \frac{(-38.39) + (-20.34) + (-46.42) + \ldots + (-11.54)}{10}$$
$$= \frac{-6.99}{10} = -.699$$
$$s = \sqrt{\frac{((-38.39 - (-.699))^2 + \ldots + (-11.54 - (-.699))^2)}{9}}$$
$$= \sqrt{\frac{22376.5}{9}} = 49.86$$

c Minitab output follows:

```
Descriptive Statistics: % Change
  Variable    N    Mean   SE Mean   StDev   Minimum     Q1   Median     Q3   Maximum
% Change     10  -0.699      15.8    49.9     -46.4  -36.9    -24.5   56.7      92.4
```

$$\text{Median} = \frac{(-20.34) + (-28.59)}{2} = -24.47$$

$$\text{IQR} = 56.7 - (-36.9) = 93.6$$

d Probably not, because the distribution is skewed with outliers on the higher end that affect the values of mean and standard deviation. Minitab output follows:

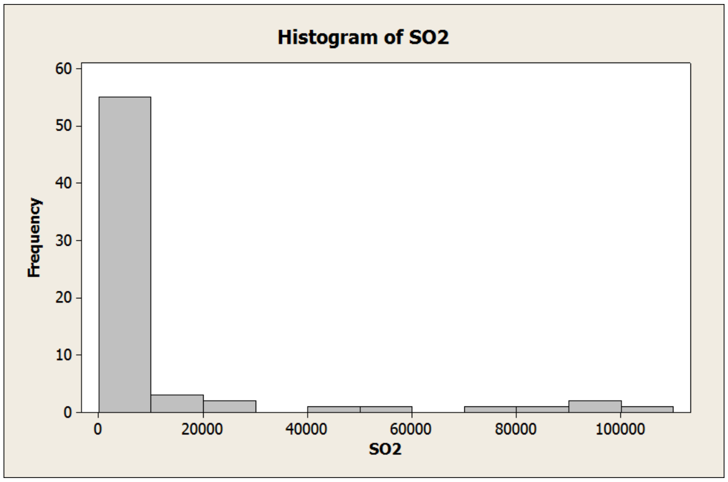
```
Descriptive Statistics: % Change
  Variable   N   Mean  SE Mean  StDev  Minimum     Q1  Median    Q3  Maximum
% Change     9  -11.0     13.3   39.9    -46.4  -37.4   -28.6  22.4     57.9
```

$$\begin{aligned}
\bar{x} &= \frac{(-38.39) + (-20.34) + (-46.42) + \ldots + (-11.54)}{9} \\
&= \frac{-99.3402}{9} = -11.0378 \\
s &= \sqrt{\frac{(-38.39 - (-.699))^2 + \ldots + (-11.54 - (-.699))^2}{8}} \\
&= \sqrt{\frac{12756.4}{8}} = 39.93 \\
\text{Median} &= -28.59 \\
\text{IQR} &= 22.4 - (-37.4) = 59.8
\end{aligned}$$

1.22 The word ‘average’ here probably refers to the mean, in which case a few older students (right skewed data) would have made the mean larger than the median age of 24.

1.23 The word ‘average’ here probably refers to the mean, in which case a few winters with very deep snow pack (right skewed data) would have made the mean larger than the median (and hence more than 50% of the data would be below the mean)

1.24 **a** Histogram for SO_2 levels in various counties:



b

$$
\begin{aligned}
\bar{x} &= \frac{2447 + 1586 + 410 + \ldots + 290}{67} = 10591.77 \\
s &= \sqrt{\frac{(2447 - 10591.77)^2 + (1586 - 10591.77)^2 + \ldots + (290 - 10591.77)^2}{66}} \\
&= \sqrt{\frac{41222887019.64}{66}} = 24991.78
\end{aligned}
$$

c Median = 926.48, IQR = 4073 − 254 = 3819

d The data is heavily skewed to the right. There are several counties that have abnormally high SO_2 levels. At least half of the counties have reported SO_2 levels less than or equal to 926.68. The SO_2 levels of the middle 50% of counties are between 254 and 4073.

1.25 **a** Composite Mean $= \dfrac{(4.0)(30) + (4.2)(33)}{63} = 4.10476$

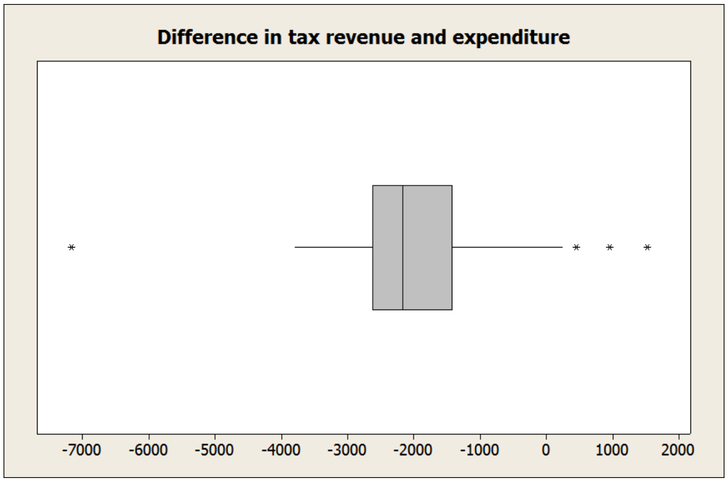
b Composite Mean $= \dfrac{(4.2)(30) + (2.7)(29) + (3.0)(29) + (4.2)(30) + (3.0)(30)}{30 + 29 + 29 + 30 + 30} = 3.4277$

1.26 **a** Because of outliers, the median and IQR may better describe the 'average' state and the spread of most of the data set.

b Because of outliers, the median and IQR may better describe the 'average' state and the spread of the data set.

c Total revenue values span from 2,880 to 176,081. Most values lie below about 30,000, but there are several states with very large revenue. Per capita revenue values span from 589 to 7,109. Most values are between 1,000 and 4,000, but there are some states with abnormally large tax revenue.

d Boxplot of difference between per capita tax revenue and per capita expenditure:



e Because of the existence of several large outliers, the median and IQR may provide a better description of the data.

1.27 **a**

$$
\begin{aligned}
\bar{x}_{\mathrm{MP}} &= \frac{2.8+3.1+4.7+\ldots+9.0}{12} = 5.78\\
s_{\mathrm{MP}} &= \sqrt{\frac{(2.8-5.78)^2+(3.1-5.78)^2+\ldots+(9.0-5.78)^2}{11}}\\
&= \sqrt{\frac{40.6625}{11}} = 1.923
\end{aligned}
$$

$$
\begin{aligned}
\bar{x}_{\mathrm{B}} &= \frac{0.9+1.2+1.3+\ldots+3.4}{12} = 1.68\\
s_{\mathrm{B}} &= \sqrt{\frac{(0.9-1.68)^2+(1.2-1.68)^2+\ldots+(1.3-1.68)^2}{11}}\\
&= \sqrt{\frac{4.77667}{11}} = 0.659
\end{aligned}
$$

b Because the value 3.4 is high compared to the rest of the measurements for the Baytex LC50, the mean and standard deviation become abnormally large by the inclusion of this data point.

1.28 One could compute the mean and standard deviation of the two data sets to see if the average time interval has decreased from group 1 to group 2, and the standard deviation of the two groups to determine if the difference in the two means can be explained by the natural variation in the data.

$$
\begin{aligned}
\bar{x}_{\mathrm{G1}} &= \frac{23+261+87+\ldots+42}{15} = 83.4\\
s_{\mathrm{G1}} &= \sqrt{\frac{(23-83.4)^2+(261-83.4)^2+\ldots+(42-83.4)^2}{14}}\\
&= \sqrt{\frac{111395.6}{14}} = 89.201
\end{aligned}
$$

$$
\begin{aligned}
\bar{x}_{\mathrm{G2}} &= \frac{12+120+11+\ldots+95}{14} = 37.5\\
s_{\mathrm{G2}} &= \sqrt{\frac{(12-37.5)^2+(120-37.5)^2+\ldots+(95-37.5)^2}{13}}\\
&= \sqrt{\frac{20877.5}{13}} = 40.074
\end{aligned}
$$

The means seem to suggest that the average time interval for group 2 is smaller. Data possibly indicates wearing out of the systems over time.

1.29 **a**

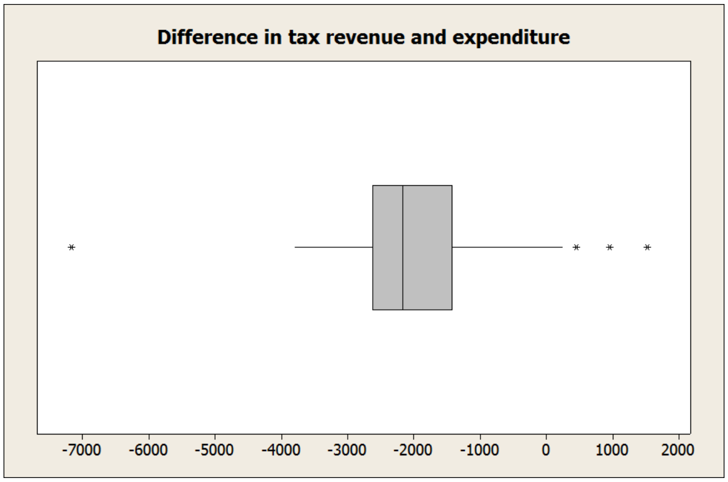
$$\begin{aligned}
\bar{x}_{\text{SDB}} &= \frac{12+8+3+\ldots+15}{12} = 10.75 \\
s_{\text{SDB}} &= \sqrt{\frac{(12-10.75)^2+(8-10.75)^2+\ldots+(15-10.75)^2}{11}} \\
&= \sqrt{\frac{192.25}{11}} = 4.181
\end{aligned}$$

$$\begin{aligned}
\bar{x}_{\text{FOB}} &= \frac{14+15+15+\ldots+22}{12} = 15 \\
s_{\text{FOB}} &= \sqrt{\frac{(14-15)^2+(15-15)^2+\ldots+(22-15)^2}{11}} \\
&= \sqrt{\frac{182}{11}} = 4.068
\end{aligned}$$

b The variation in percent bridges recorded among southeastern states seem to be comparable for structurally deficient and functionally obsolete bridges, however, there are a higher mean percentage of functionally obsolete bridges than structurally deficient ones.

1.5 Summary Measures and Decisions

1.30 **a** Boxplot of difference of per capita tax revenue and per capita expenditure:



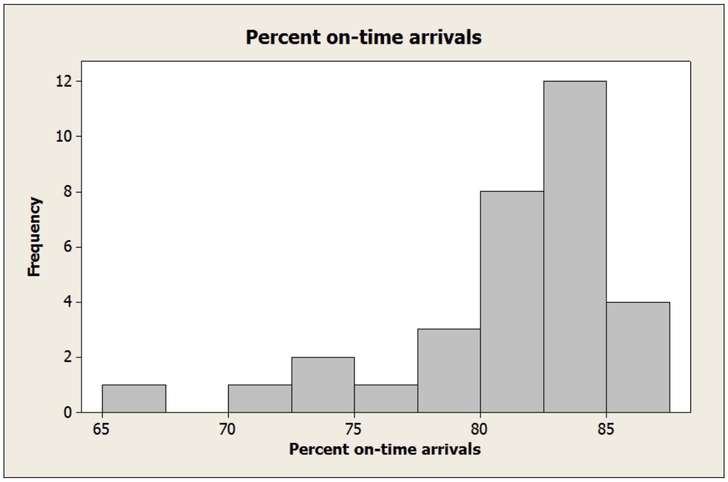
AK is a very extreme outlier on the lower end, indicating that per capita expenditure is much higher that per capita tax revenue. MA, NH and RI are less extreme outliers on the upper end of the dataset, indicating that their per capita tax revenue is greater than their per capita expenditure.

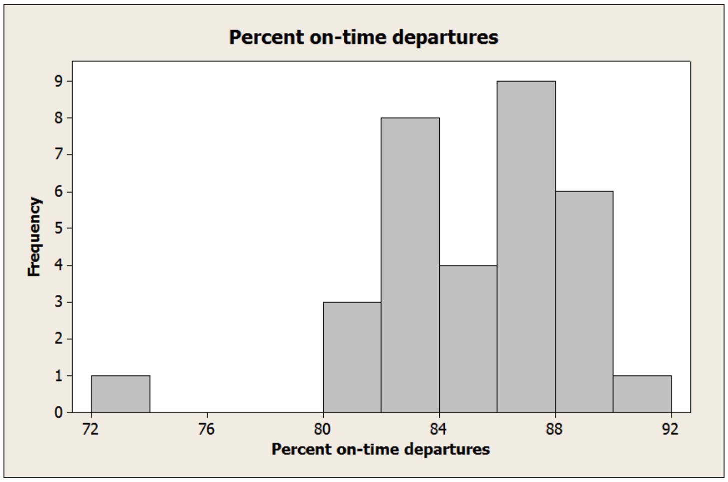
b Alaska is the lower outlier with a difference of -7161. Using the formulas for mean and standard deviation we find that $\bar{x} = -2009.14$ and $s = 1288.622$. We can then use the formula for z-score to find:

$$z = \frac{-7161 - (-2009.14)}{1288.622} = -3.998$$

It's z-score of -3.998 reveals that it is almost 4 standard deviations below the mean.

1.31 **a** Histograms and boxplots of percent on-time arrivals and departures:





Both, the arrival and departure time distributions are left-skewed, arrival times more so than the departure times. The median percentage of on-time departures is higher than the median percentage of on-time arrivals. Both the distributions have about the same range. Both the distributions have outliers on the lower end, indicating a low-performing airport (or airports).

b $\frac{1}{32} = 3.125\%$

c $\frac{0}{32} = 0\%$

d For arrival data, we find that $\bar{x} = 81.33$ and $s = 4.558$. For departure data, we find that $\bar{x} = 85.23$ and $s = 3.417$.

	Arrivals		Departures	
k	$(\bar{x} - ks, \bar{x} + ks)$	% Data in Interval	$(\bar{x} - ks, \bar{x} + ks)$	% Data in Interval
1	$(76.772, 85.888)$	78.1%	$(81.813, 88.647)$	71.9%
2	$(72.214, 90.446)$	93.75%	$(78.396, 92.064)$	96.9%

Departure data seems to agree more strongly with the empirical rule, which says that around 68% should lie within 1 standard deviation of the mean and 95% should lie within 2 standard deviations of the mean.

e The range representing values within 5% of the mean is $(81.33 - 0.05(81.33), 81.33 + .05(81.33)) = (77.2635, 85.3965)$. 24 or 75% of airports have percent on-time arrivals in this range.

f The range representing values within 5% of the mean is $(85.23 - 0.05(85.23), 85.23 + .05(85.23)) = (80.9685, 89.4915)$. 28 or 87.5% of airports have percent on-time arrivals in this range.

g Looking at the boxplots, we see that for arrival times, the three lowest: Chicago O'Hare, Newark Int and New York LaGuardia qualify as outliers, and for departure times, Chicago O'Hare is an outlier.

h
$$z_{\text{ATLarrivals}} = \frac{80.5 - 81.33}{4.558} = -0.1821$$

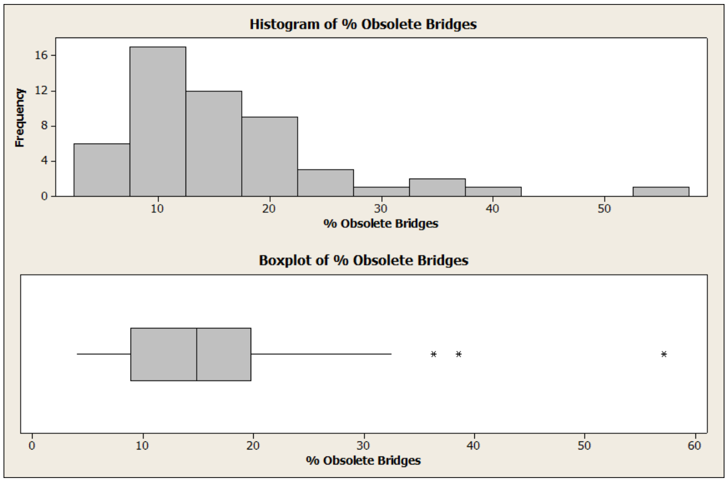
$$z_{\text{ATLdepartures}} = \frac{83.5 - 85.23}{3.417} = -0.5063$$

$$z_{\text{CHIarrivals}} = \frac{67 - 81.33}{4.558} = -3.1439$$

$$z_{\text{CHIdepartures}} = \frac{73.4 - 85.23}{3.417} = -3.4621$$

Atlanta is -0.1821 standard deviations below the mean for percent on-time arrivals and -0.5063 standard deviations below the mean for percent on-time departures. Atlanta is better with on-time arrivals than on-time departures. Chicago O'Hare is -3.1439 standard deviations below the mean for percent on-time arrivals and -3.4621 standard deviations below the mean for percent on-time departures. O'Hare is also better with on time arrivals than on-time departures.

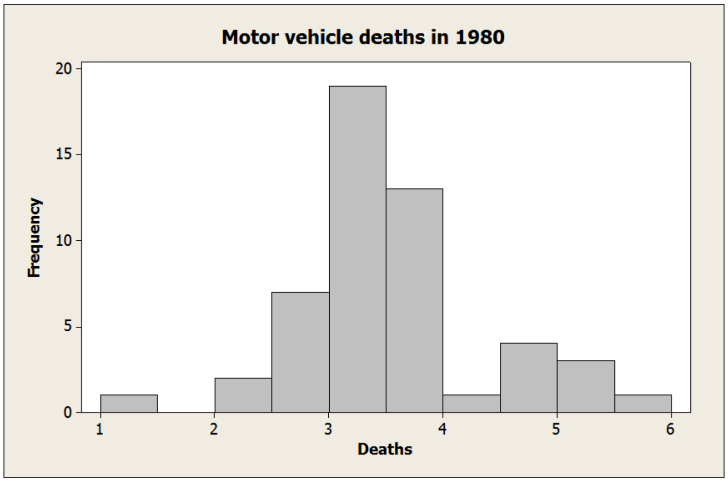
1.32 **a** Histogram and Boxplot for percent obsolete bridges in US:

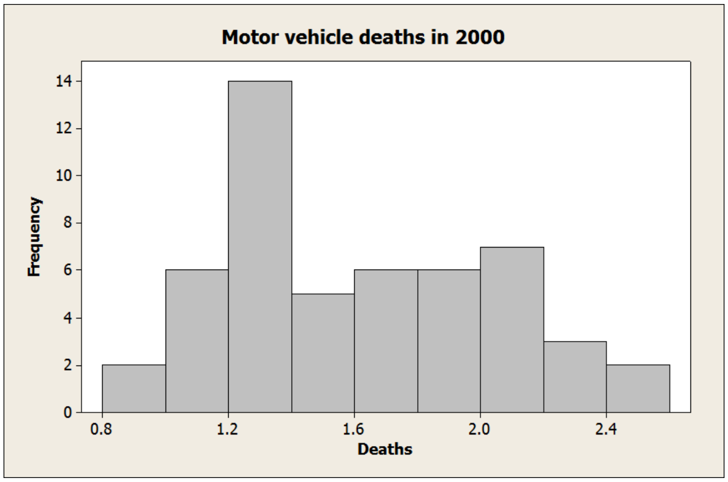


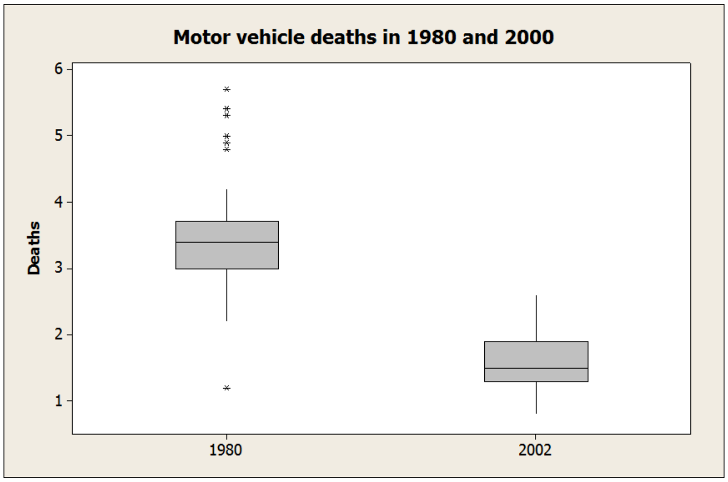
b The data are right skewed, with three outlier. The outliers are DC, Puerto Rico, and HI. The % of obsolete bridges ranged from 4% to 57% with a median of about 15%.

c If the outliers Puerto Rico and DC were removed from the dataset, then the mean and the standard deviation would become smaller.

1.33 **a** Histograms and Boxplots for motor vehicle deaths in 1980 and 2002:







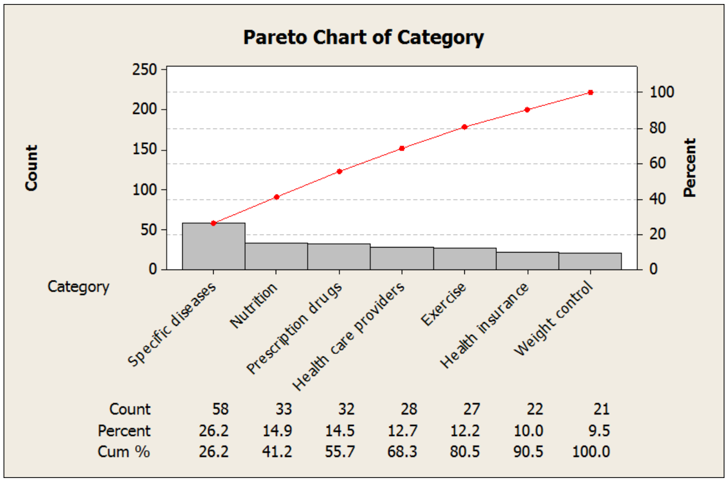
The skewness in the distribution and the abundance of outliers in the 1980 data indicate that the median and IQR will describe these datasets better than the mean and median.

b Washington, DC; Idaho; Montana; West Virginia; Wyoming; Arizona; New Mexico; Louisiana; and Nevada. These states, except for DC, have low population densities, which may mean that medical teams must travel large distances to provide help to accident victims. In DC, medical teams should be able to arrive at accidents much more quickly.

c Based on the data, even though more vehicles are probably using the highways in 2002 than in 1980, the median rate of motor vehicle deaths has decreased, which may indicate that safety measures have improved in that time.

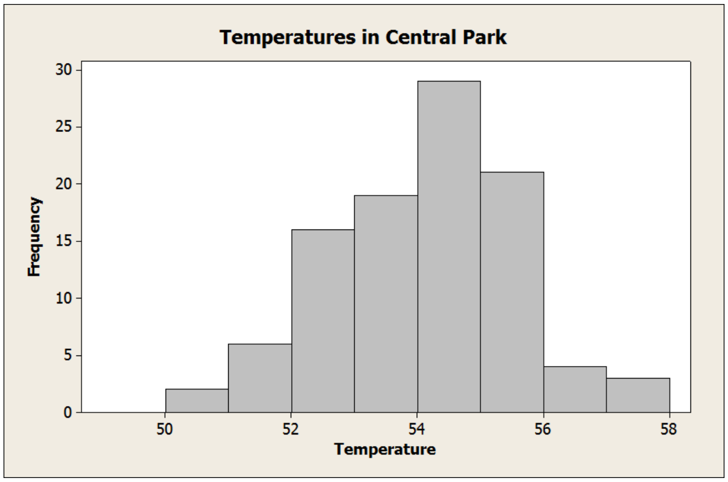
1.6 Supplementary Exercises

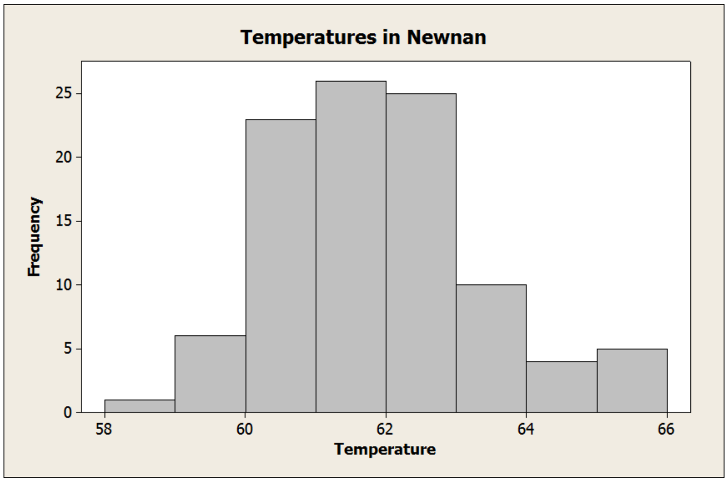
1.34 Pareto Chart of internet medical research:



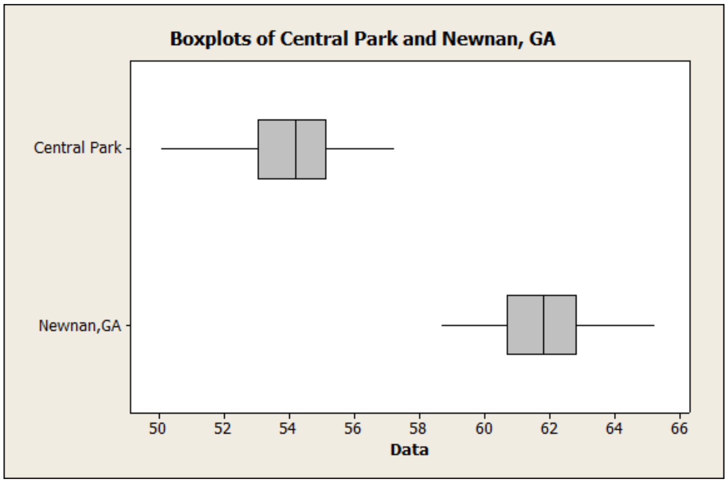
More people searched for information on specific diseases than on any other category, almost twice as much as the next largest category, nutrition. The other categories were selected by a fairly similar percent of respondents that ranged from 21% to 33%.

1.35 **a** Histograms of temperatures for Central Park and Newnan:





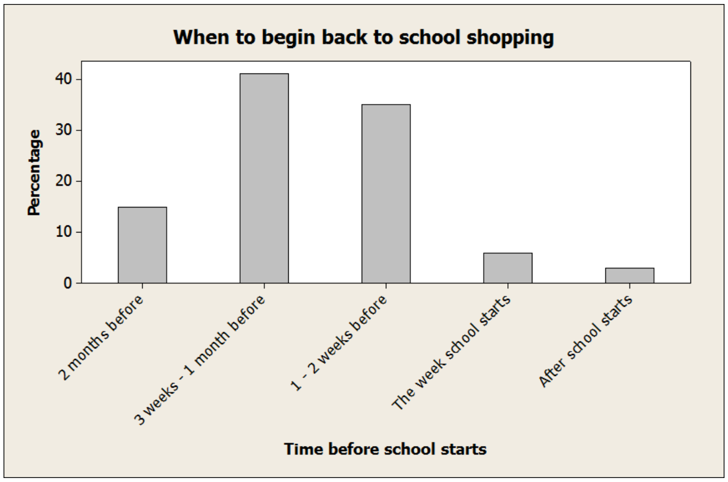
b Boxplots of temperatures for Central Park and Newnan:



The distribution of annual temperatures in Central Park is slightly left-skewed. The temperatures ranged from about 50°F to 57°F with a mean about 54°F. There are no outliers. The distribution of temperatures at Newnan is slightly right-skewed. The temperatures ranged from about 58°F to 66°F with a mean about 62°F. There are no outliers.

c The shapes of the two distributions indicate that Central Park has seen more years with warmer temperatures and Newnan more years with cooler temperatures during the last century. On the average, Newnan is warmer than the Central Park. The range of temperatures is about the same at both locations.

1.36 **a** Bar chart of when consumers begin back-to-school shopping:

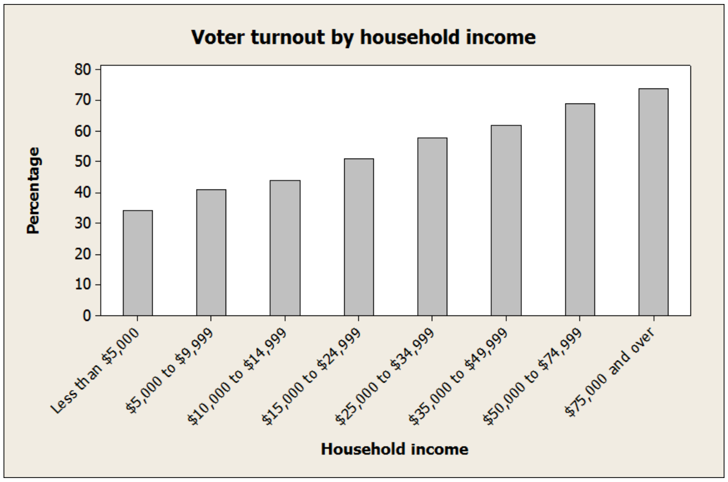


A vast majority of consumers begin shopping at least a week before school starts, with a few shoppers starting that week or after.

b $15 + 41 + 35 + 6 = 97$ percent of consumers begin shopping before school starts.

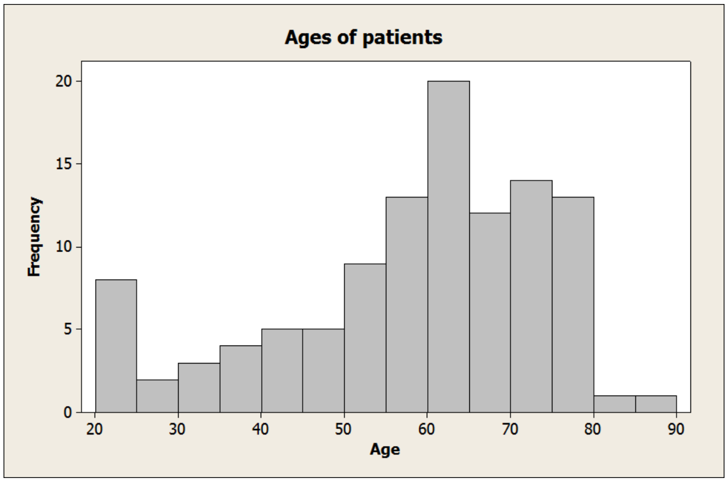
c About 6% of $8,453$ which is around 507 consumers.

1.37 Bar chart of classification of voters by income:



The percentage of eligible voters who voted in the 2000 presidential election increased steadily with the household income group. From the lowest income group, the lowest percentage of voters voted, whereas from the highest income group the highest percentage of voters voted in this election.

1.38 **a** Histogram of ages of patients:

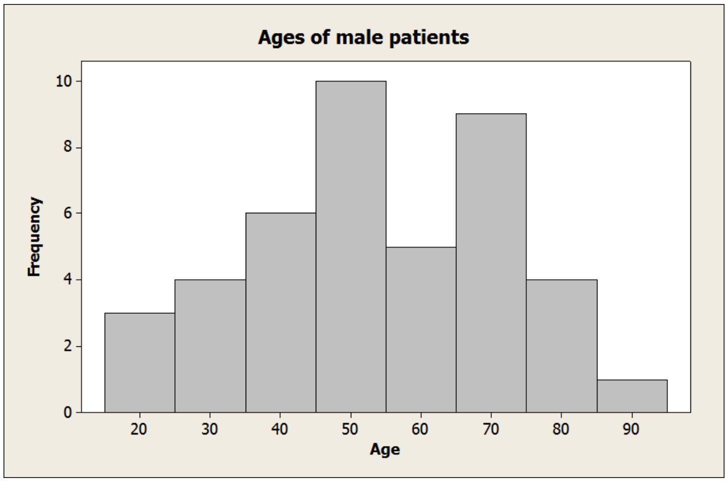


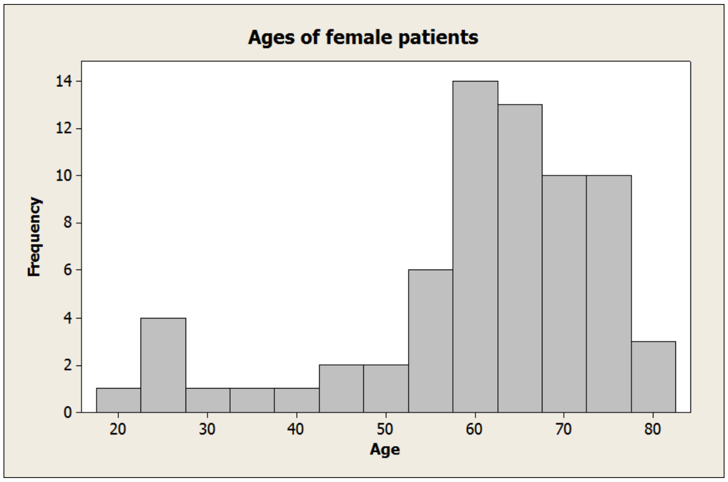
Data are skewed to the left, with a majority of patients coming from age groups between 50 and 80. The average age of patients is about 55. A large number of patients are from the age group 20-25 compared to the immediately following groups.

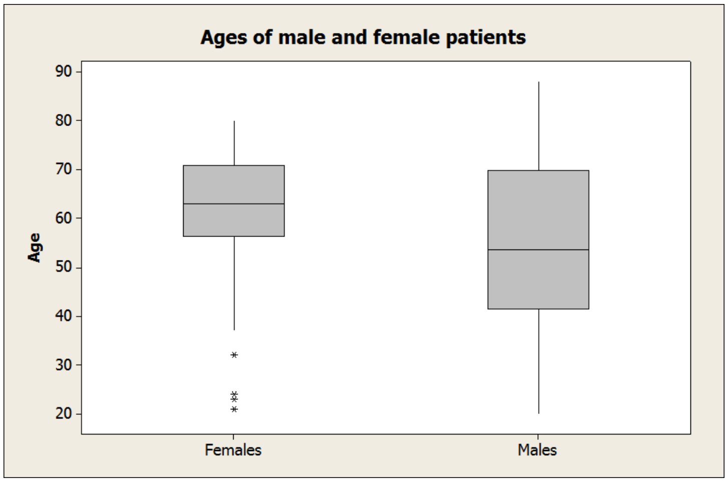
b We find that $\bar{x} = 57.96$ and $s = 16.058$. The empirical rule states that almost all of the data should lie between $57.96 - 3(16.058) = 9.786$ and $57.96 + 3(16.058) = 106.134$. All of the data lie within this interval. 95% of the data should lie within $57.96 - 2(16.058) = 25.844$ to $57.96 + 2(16.058) = 90.076$. 92.7% of the data lie within this interval. The empirical rule works tolerably well with this data.

c The data contain no outliers, so no.

d Histograms and Boxplots by gender:

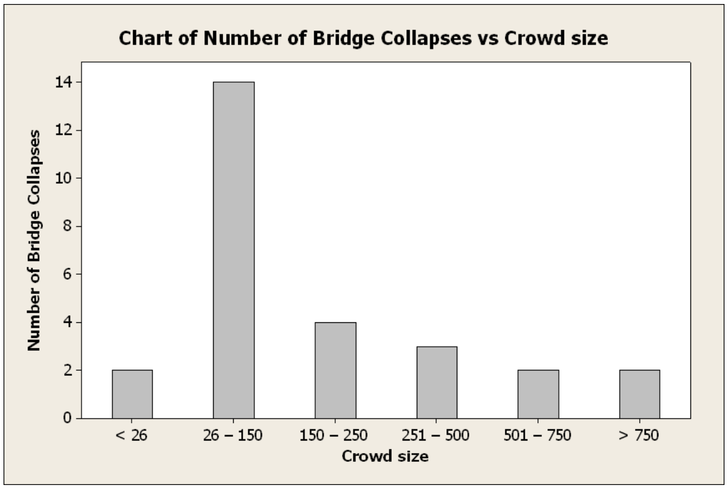






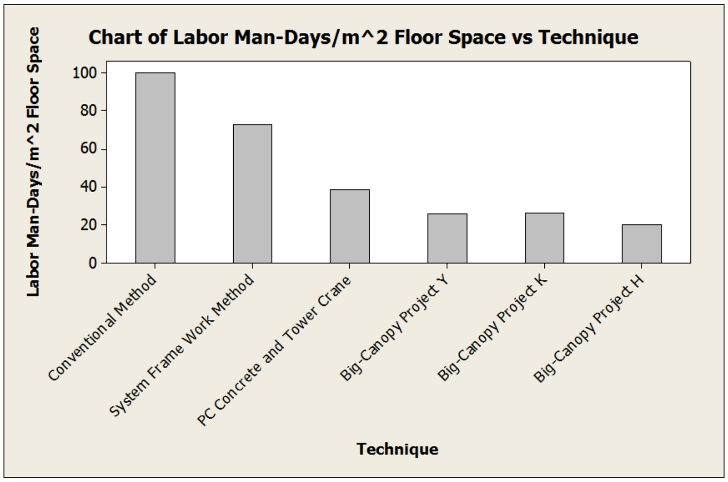
The age distribution of female patients is left skewed whereas that of male patients is more mound-shaped. For male patients age ranged from about 20 to 90 and for female patients age ranged from about 20 to 80. The average age of male patients is about 50 and that of female patients is closer to 60. There are a few female patients that are much younger than the rest of the female patients.

1.39 Bar chart of bridge collapses by size of crowd:



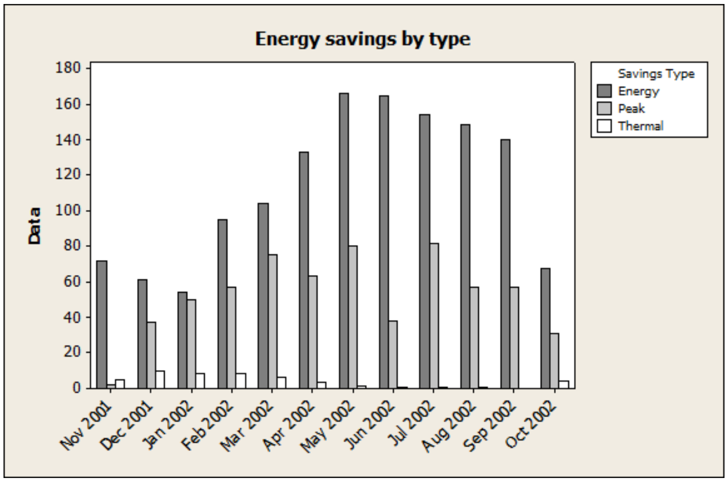
The median number of people on collapsing bridges was between 26 and 150; the data are skewed to the right, so most of the bridges had a relatively small crowd when they collapsed; the spread is small, a vast majority of the collapses occurred with a relatively small number of people on the bridge. The crowd size on collapsing bridges ranged from less than 26 to more than 750.

1.40 Bar chart of different construction methods:



Big-Canopy methods resulted in a 74% to 80% reduction in labor time compared to the conventional methods, and at least a $\frac{38.6 - 26}{38.6} = 32.6\%$ decrease in labor time from the next most efficient method.

1.41 **a** Bar Chart showing energy, max peak demand, and thermal savings over time:



b Every month the energy savings are the highest and the thermal savings are the lowest. The energy savings show a cycle with highest savings during the summer months and lowest savings during the winter months. On the other hand, thermal savings are highest during the winter months and lowest during the summer months, showing exactly opposite cycles. The maximum peak demand savings are higher in general during summer months and lower in the winter months.

c

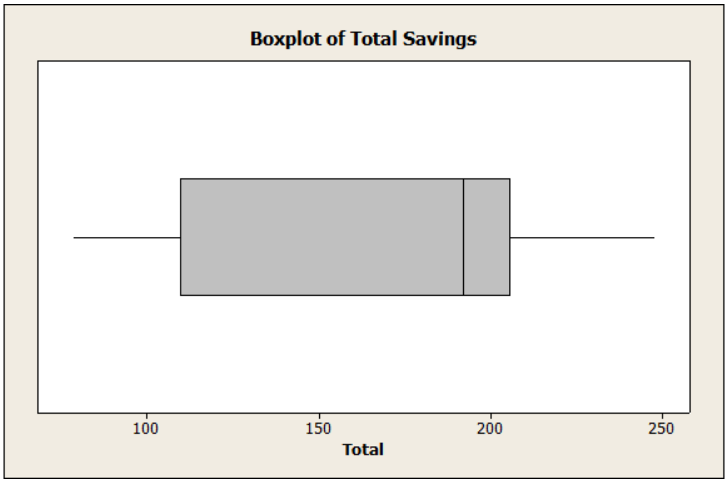
Month	Energy	Peak	Thermal	Total
Nov 2001	71.49	1.77	5.09	78.35
Dec 2001	61.43	37.39	9.57	108.39
Jan 2002	54.47	50.10	8.52	113.09
Feb 2002	94.84	56.71	8.47	160.02
Mar 2002	104.19	75.28	6.56	186.03
Apr 2002	132.77	63.33	3.17	199.27
May 2002	166.18	79.92	1.66	247.76
Jun 2002	164.24	38.40	0.60	203.24
Jul 2002	154.17	81.12	0.87	236.16
Aug 2002	148.62	56.71	0.81	206.14
Sep 2002	140.58	56.97	0.16	197.71
Oct 2002	67.35	31.09	4.35	102.79

$$\bar{x}_{\text{Total}} = \frac{78.35 + 108.38 + 113.09 + \ldots + 102.79}{12} = 169.91$$

$$s_{\text{Total}} = \sqrt{\frac{(78.35 - 169.91)^2 + (108.38 - 169.91)^2 + \ldots (102.79 - 169.91)^2}{11}}$$

$$= \sqrt{\frac{34768.484}{11}} = 56.221$$

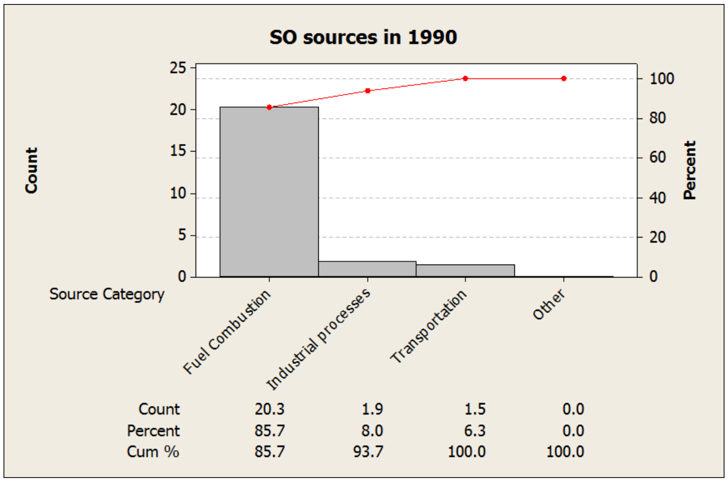
d Boxplot of total savings:



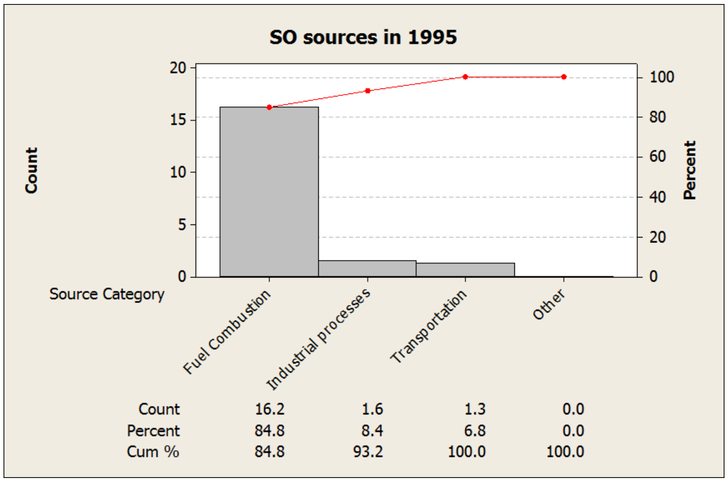
No outliers.

1.42 $\frac{2.9(27) + 2.6(16)}{43} = 2.79$

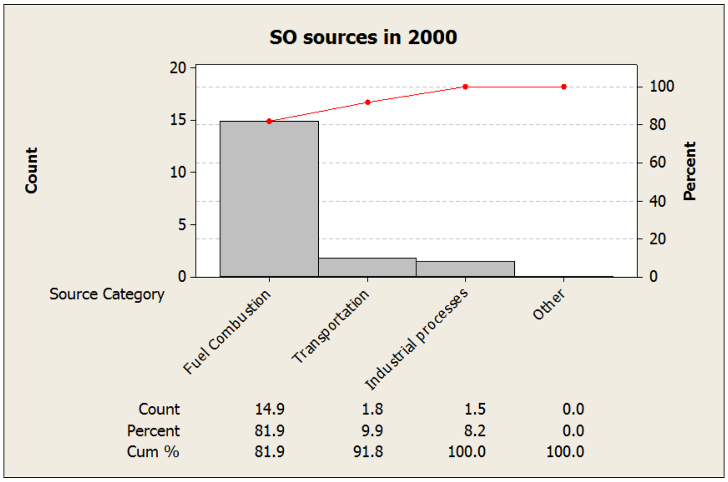
1.43 Pareto charts of SO pollution sources:



Source Category	Fuel Combustion	Industrial processes	Transportation	Other
Count	20.3	1.9	1.5	0.0
Percent	85.7	8.0	6.3	0.0
Cum %	85.7	93.7	100.0	100.0

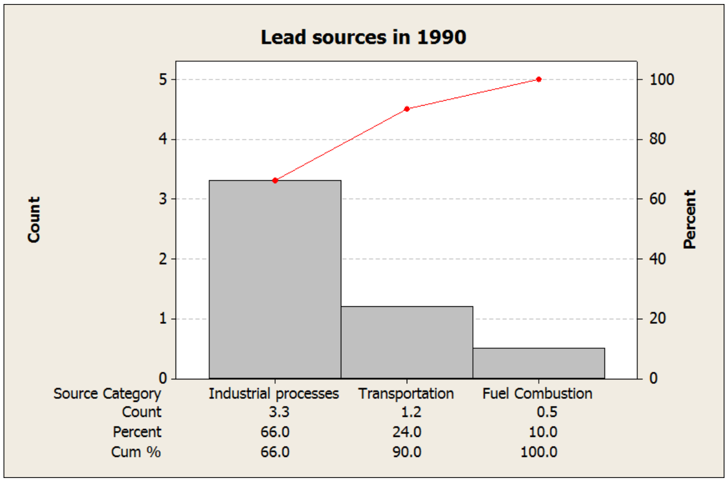


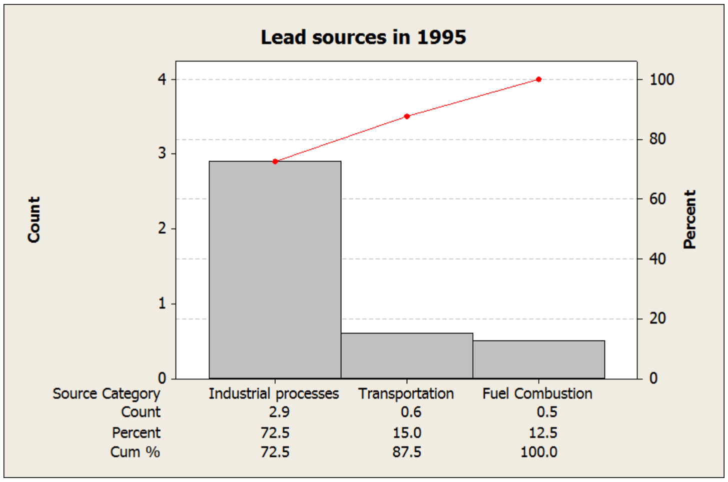
Source Category	Fuel Combustion	Industrial processes	Transportation	Other
Count	16.2	1.6	1.3	0.0
Percent	84.8	8.4	6.8	0.0
Cum %	84.8	93.2	100.0	100.0



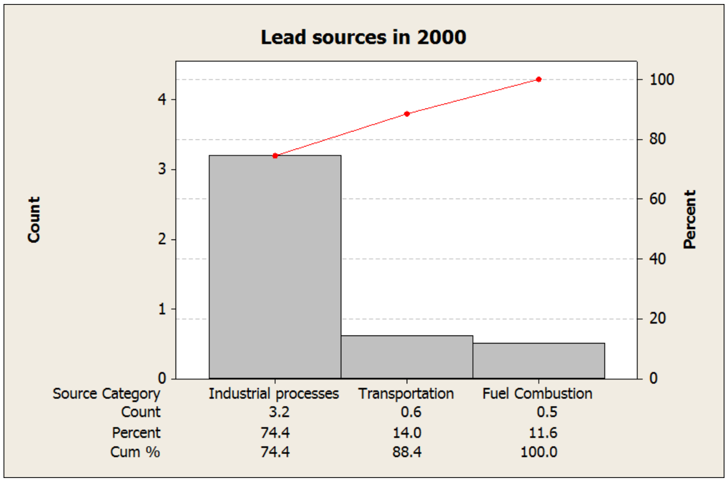
Fuel combustion is the largest contributor of sulfur dioxide emissions. Although the amount of contribution decreased over the years, it is still a major contributor. Amount of contribution by industrial processes decreased over the years, but the percentage of total emission increased over the years. The percent contribution by transportation increased slightly.

1.44 **a** Pareto charts of lead pollution sources:





Source Category	Industrial processes	Transportation	Fuel Combustion
Count	2.9	0.6	0.5
Percent	72.5	15.0	12.5
Cum %	72.5	87.5	100.0



Source Category	Industrial processes	Transportation	Fuel Combustion
Count	3.2	0.6	0.5
Percent	74.4	14.0	11.6
Cum %	74.4	88.4	100.0

Industrial processes are the main contributor to lead pollution and, though the amount of lead pollution from industrial processes only decreased slightly from 1990 to 2000, it contributes a larger percentage in 2000. Fuel combustion lead pollution remained constant throughout the decade and transportation lead pollution saw a large reduction from 1990 to 1995.

b It seems that although there were improvements in lead pollution from 1990 to 1995, lead pollution is either rising or remaining constant in all areas since 1995.