# CHAPTER 2. DATA DESCRIPTION AND TREATMENT



The following table provides a summary of the problems with their appropriate sections:

| Section | Problems |
|---|---|
| 2.2 | 1 to 3 |
| 2.3 | 4 to 22 |
| 2.4 | 23 to 32 |
| 2.4 and 2.5 | 33 to 49 |

## 2.2. Classification of Data

---

**Problem 2-1.**

| Scale | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 |
|---|---|---|---|---|---|
| Nominal | Color | Citizenship | Ethnic origin | Religion | Race |
| Ordinal | Hardness | Texture | Hazard | | |
| Interval | Elevation | Velocity | Acceleration | Temperature | Yield Strength |
| Ratio | Life expectancy | Flow volume | Coefficient of variation | Standard deviation | Reynolds number |

---

**Problem 2-2.**

Age is a variable of interest.

| Scale | Function |
|---|---|
| Nominal | |
| Ordinal | child, adult, senior citizen |
| Interval | date of birth |
| Ratio | life expectancy |

---

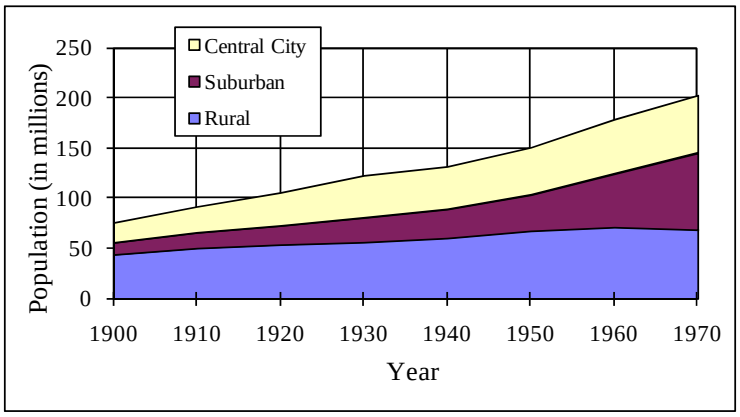
**Problem 2-3.**

Copper content of steel is a variable of interest.

| Scale | Function |
|---|---|
| Nominal | presence of copper |
| Ordinal | small, medium, high |
| Interval | weight |
| Ratio | percent by weight of steel |

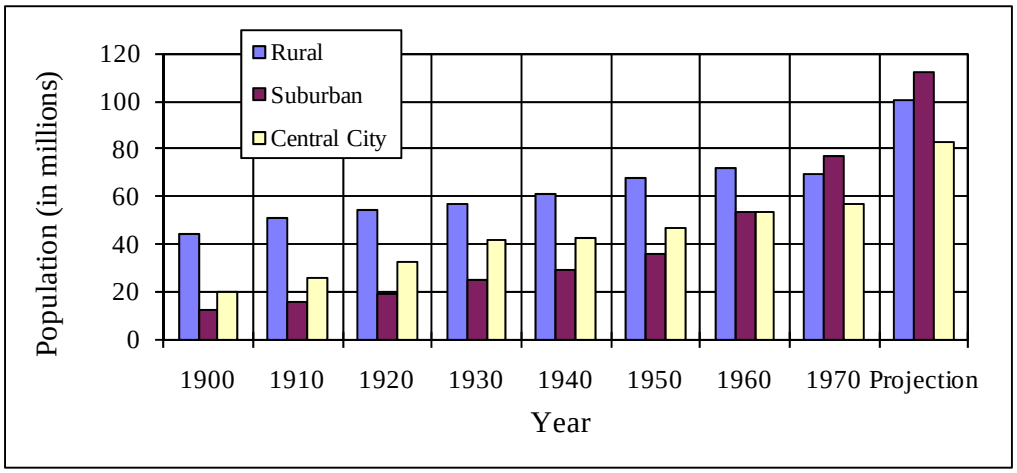
## 2.3. Graphical Description of Data

**Problem 2-4.**



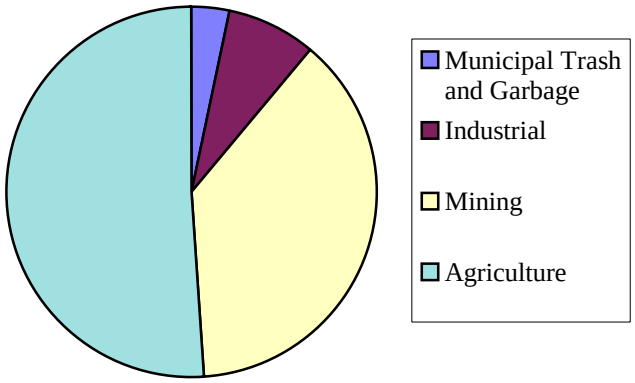

**Problem 2-5.**

See Problem 2-4 for the area chart. The column chart is as follows:



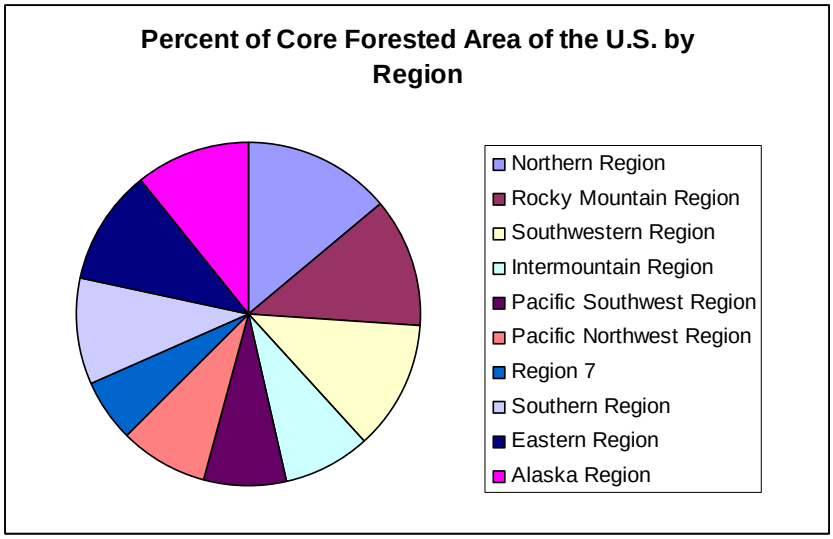

Assuming that the 295 is future projection, then the best estimates of the proportions would be those from the last year of record, 1970, which are 34, 38, and 28. This would lead to 100.3, 112.1, and 82.6 for rural, suburban, and central city. It appears from the data of Problem 2-18 that the proportion of rural is decreasing, the proportion of suburban is increasing, and the central city is remaining constant. Regression lines by curve fitting could be used to enhance prediction. The stacked columns figures show the trends side-by-side; whereas the area chart shows the relative proportions and total numbers.

**Problem 2-6.**




The pie chart shows clearly the amounts as fractions of the total. The visual image gives a more lasting sense of the proportions than a tabular summary.

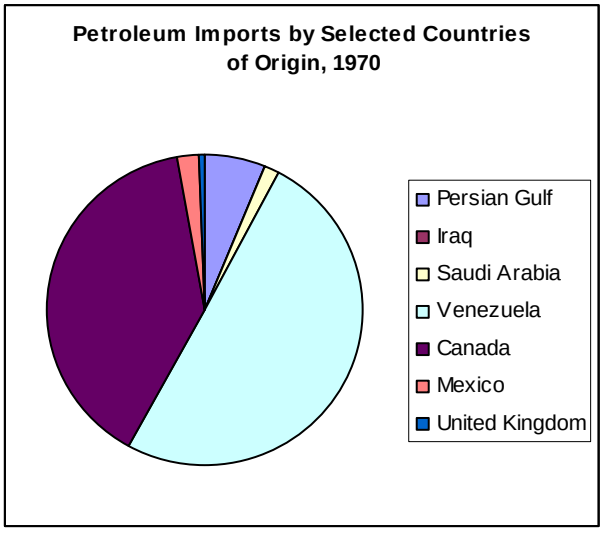
**Problem 2-7.**

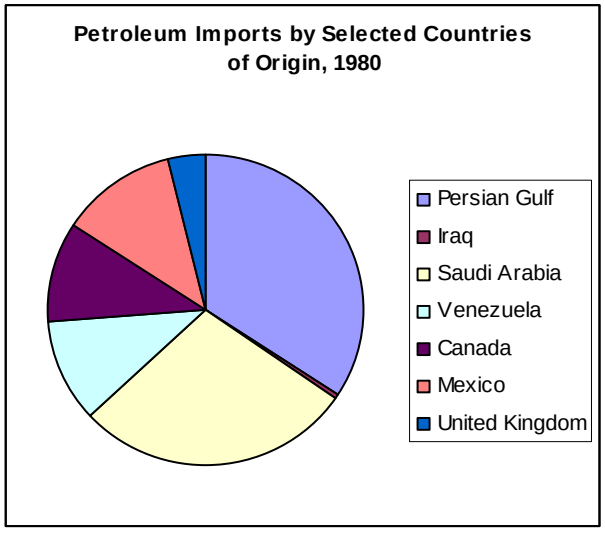


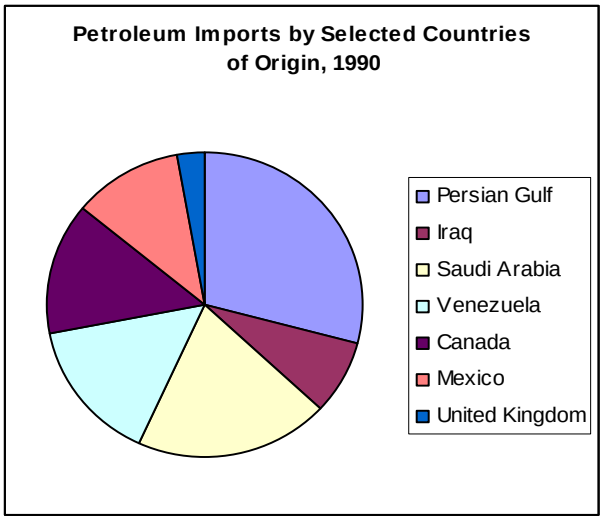

| Region | Percent Core Forested of Region | Percent of all US |
|---|---|---|
| Northern Region | 38.0 | 13.89% |
| Rocky Mountain Region | 33.5 | 12.24% |
| Southwestern Region | 33.3 | 12.17% |
| Intermountain Region | 22.1 | 8.06% |
| Pacific Southwest Region | 21.4 | 7.80% |
| Pacific Northwest Region | 23.0 | 8.40% |
| Region 7 | 15.6 | 5.70% |
| Southern Region | 27.8 | 10.14% |
| Eastern Region | 29.7 | 10.86% |
| Alaska Region | 29.4 | 10.74% |

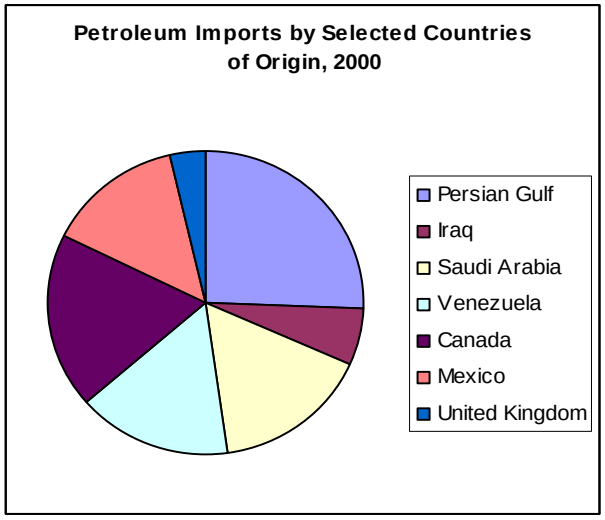
Data source (accessed in 2009):
http://cfpub.epa.gov/eroe/index.cfm?fuseaction=detail.viewInd&ch=50&subtop=210&lv=list.listByChapter&r=188266

**Problem 2-8.**



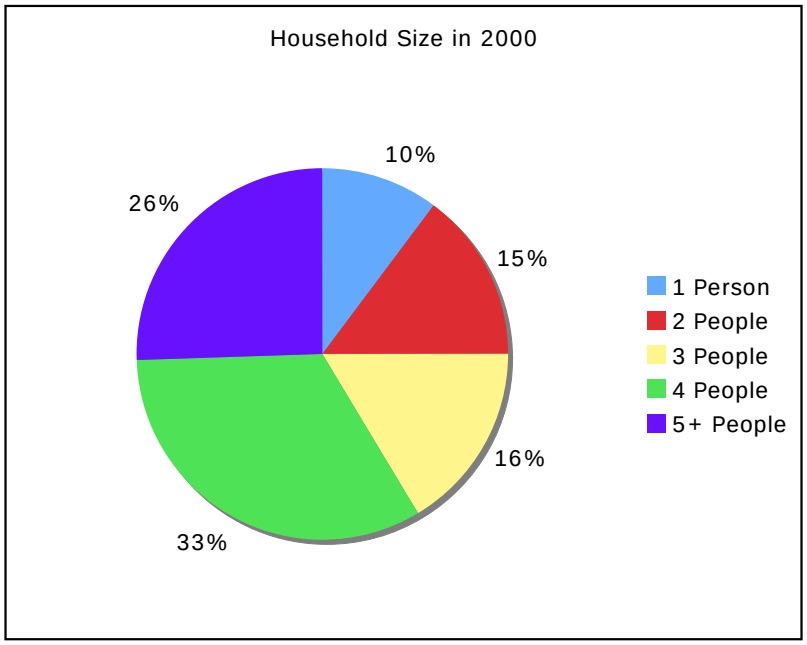




Data (Source accessed in 2009: http://www.eia.doe.gov/emeu/aer/txt/ptb0504.html)

| Petroleum Imports by Country of Origin, 1970-2000 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Selected OPEC Countries | | | Selected Non-OPEC Countries | | | |
| Year | Persian Gulf | Iraq | Saudi Arabia | Venezuela | Canada | Mexico | United Kingdom | Total Imports |
| | Thousand Barrels per Day | | | | | | | |
| 1970 | 121 | 0 | 30 | 989 | 766 | 42 | 11 | 1,959 |
| 1980 | 1,519 | 28 | 1,261 | 481 | 455 | 533 | 176 | 4,453 |
| 1990 | 1,966 | 518 | 1,339 | 1,025 | 934 | 755 | 189 | 6,726 |
| 2000 | 2,488 | 620 | 1,572 | 1,546 | 1,807 | 1,373 | 366 | 9,772 |

Through these pie charts, one can clearly see different trends in imports from various countries. For example, in 1970 the U.S. relied heavily on Venezuela for petroleum imports, but overall there has been a leveling of imports from various countries (it has become slightly more equal).

**Problem 2-9.**

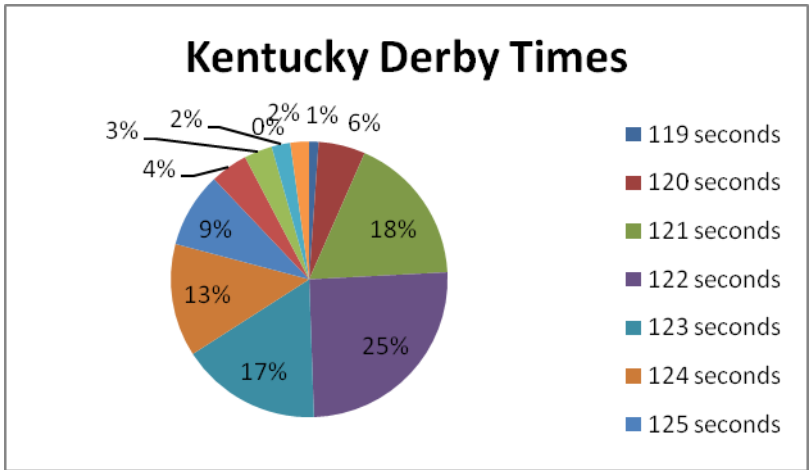



*Note: The U.S. Family Size was unable to be found because the U.S. Census Bureau finds family size through household size. Therefore, I used household size.
Source accessed in 2009: http://www.census.gov/prod/2001pubs/p20-537.pdf

Excel Table Used for Pie Chart:

| Household Size in 2000 | |
|---|---|
| Size | % |
| 1 Person | 10.4 |
| 2 People | 14.6 |
| 3 People | 16.4 |
| 4 People | 33.1 |
| 5+ People | 25.5 |

**Problem 2-10.**

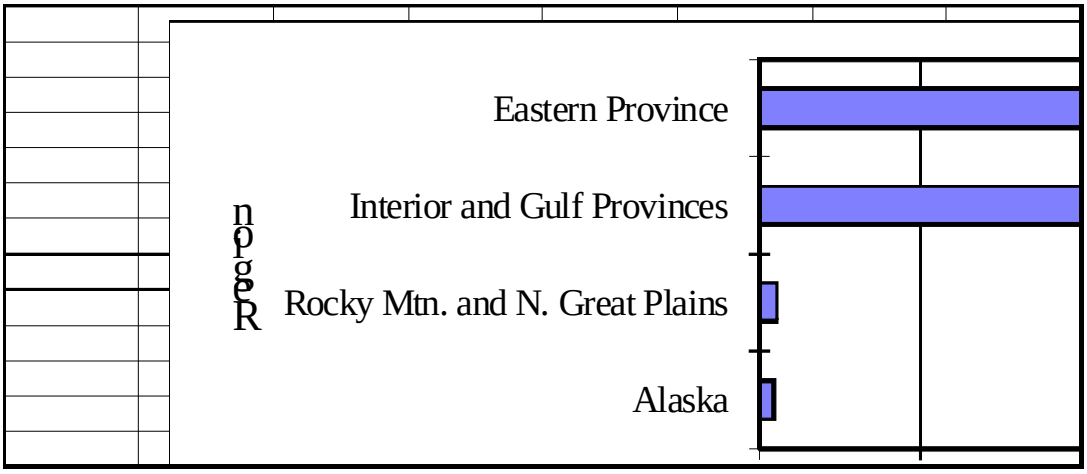



Frequency of Kentucky Derby times from 1919 to present:

| Time | Frequency |
|---|---|
| 119 seconds | 1 |
| 120 seconds | 5 |
| 121 seconds | 16 |

| | |
|---|---|
| 122 seconds | 23 |
| 123 seconds | 15 |
| 124 seconds | 12 |
| 125 seconds | 8 |
| 126 seconds | 4 |
| 127 seconds | 3 |
| 128 seconds | 0 |
| 129 seconds | 2 |
| 130 seconds | 2 |

**Problem 2-11.**




**Problem 2-12.**

Figure 2-3

Bar chart: This chart nicely displays information of total steel production between each quarter and steel type. This does not necessarily show the total steel production for each quarter but can be used to compare the total production of each type for each quarter. From this chart it is clear the total production of each steel type remains close to each other for each type except in the 3rd quarter when 40 ksi steel is produced at least five times as much as usual.

Figure 2-5a

Column chart: Steel production is shown as a percentage of total steel produced for each quarter. This chart uses a percentage for comparison and will not show totals. It can be used to show which steel is produced the most for each quarter and this information can be used to allocate resources depending on the quarter.

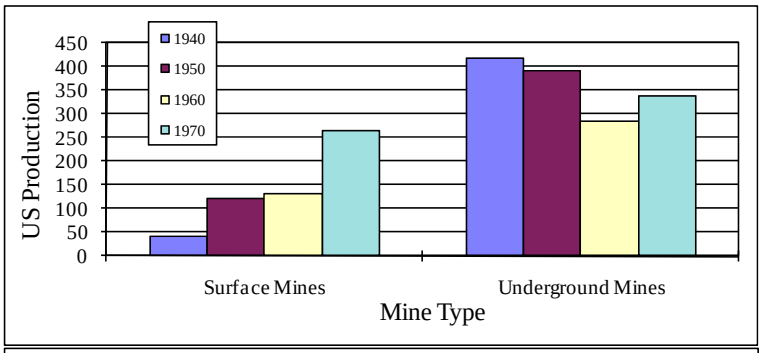
Figure 2-5b

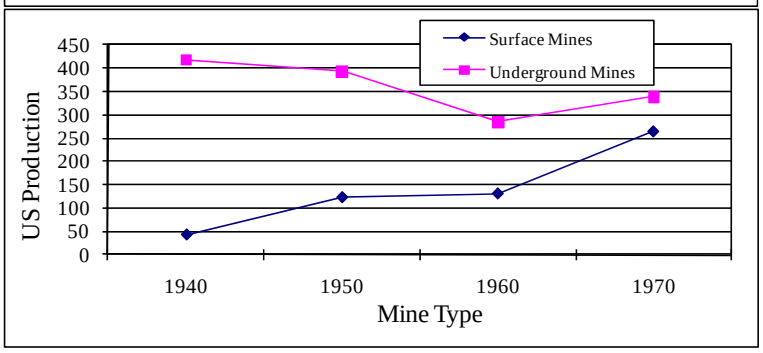
Column chart steel production shows total steel produce for each quarter. This chart quickly shows which quarter has the greatest total steel produced of all the types of steel combined. This is useful in determining the most active quarter during the year in terms of total steel produced.

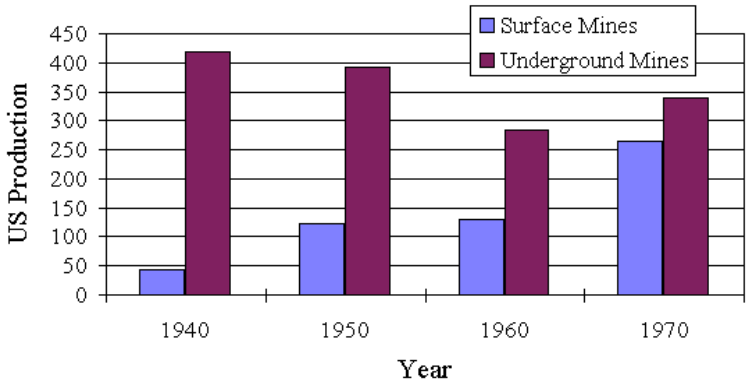
Comparison:

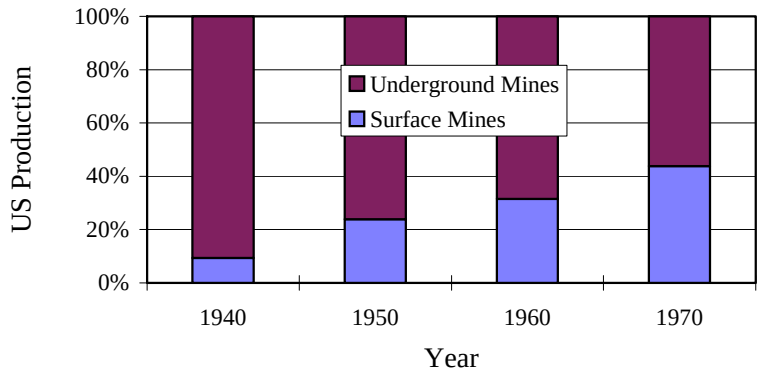
Figure 2-3 is a bar chart showing the steel production by yield strength and quarter. The emphasis in this chart is on the production for steel type. This is useful to keep track of the production for both the type and the quarter. Figures 2-5a and 2-5b also show the steel production by yield strength and quarter. However, the data in these two figures are presented in column charts where the steel production (dependent variable) is expressed as a percentage of the a total in the first figure and in tons in the second.

**Problem 2-13.**


1940
1950
1960
1970
US Production
450
400
350
300
250
200
150
100
50
0
Surface Mines
Underground Mines
Mine Type

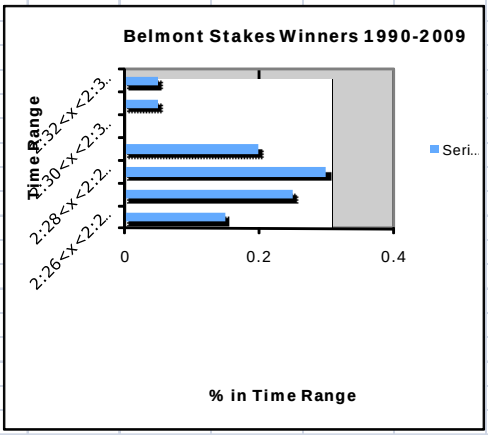

Surface Mines
Underground Mines
US Production
450
400
350
300
250
200
150
100
50
0
1940
1950
1960
1970
Mine Type


Surface Mines
Underground Mines
US Production
450
400
350
300
250
200
150
100
50
0
1940
1950
1960
1970
Year

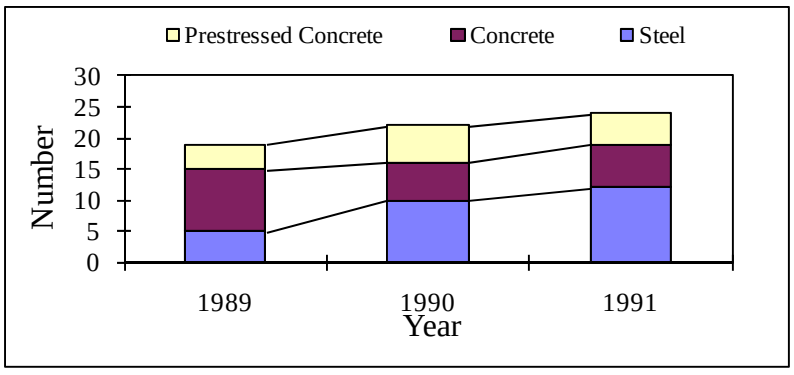

Underground Mines
Surface Mines
US Production
100%
80%
60%
40%
20%
0%
1940
1950
1960
1970
Year

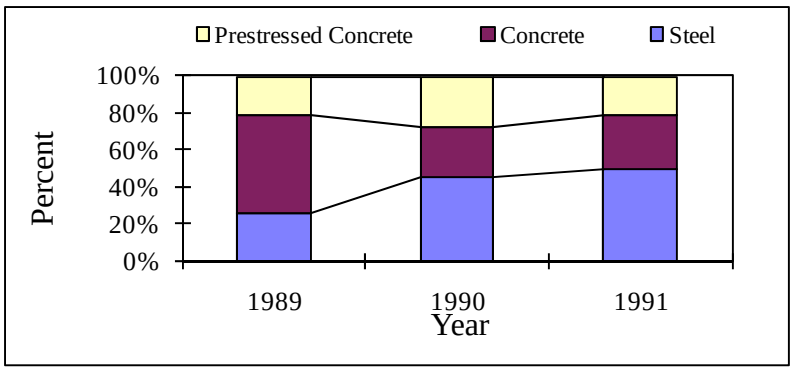
## Problem 2-14.

| | | | | | |
|---|---|---|---|---|---|
| **Data:** | | | | | % in Range |
| Year | Time | | Time Range | # in Range | 0.15 |
| 2009 | 02:27.5 | | 2:26<x<2:26 | 3 | 0.25 |
| 2008 | 02:29.7 | | 2:27<x<2:27 | 5 | 0.3 |
| 2007 | 02:28.7 | | 2:28<x<2:28 | 6 | 0.2 |
| 2006 | 02:27.8 | | 2:29<x<2:29 | 4 | 0 |
| 2005 | 02:28.7 | | 2:30<x<2:30 | 0 | 0.05 |
| 2004 | 02:27.5 | | 2:31<x<2:31 | 1 | 0.05 |
| 2003 | 02:28.3 | | 2:32<x<2:32 | 1 | |
| 2002 | 02:29.7 | | | | |
| 2001 | 02:26.8 | | | | |
| 2000 | 02:31.2 | | | | |
| 1999 | 02:27.8 | | | | |
| 1998 | 02:29.0 | | | | |
| 1997 | 02:28.8 | | | | |
| 1996 | 02:28.8 | | | | |
| 1995 | 02:32.0 | | | | |
| 1994 | 02:26.8 | | | | |
| 1993 | 02:29.8 | | | | |
| 1992 | 02:26.1 | | | | |
| 1991 | 02:28.0 | | | | |
| 1990 | 02:27.2 | | | | |
| | | | | | |
| **Column Chart:** | | | | | |



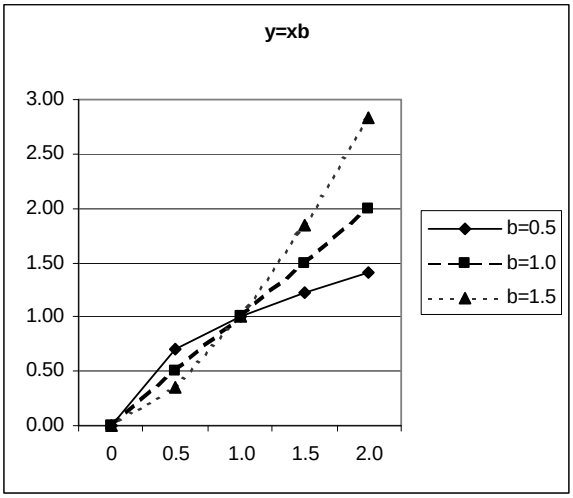

## Problem 2-15.

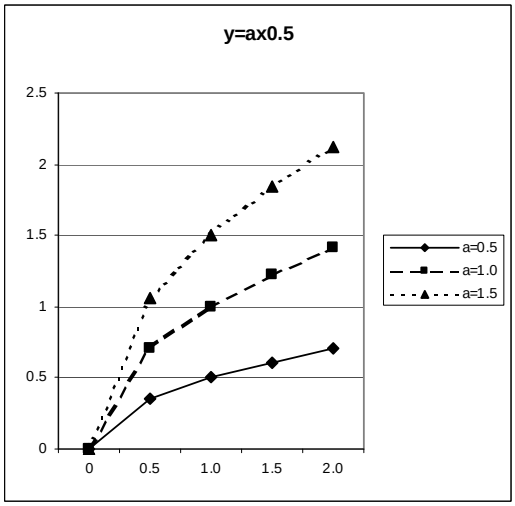





**Problem 2-16.**

| $y=x^b$ | | | | $y=ax^{0.5}$ | | | |
|---|---|---|---|---|---|---|---|
| *X* | *b=0.5* | *b=1.0* | *b=1.5* | *X* | *a=0.5* | *a=1.0* | *a=1.5* |
| 0 | 0.00 | 0.00 | 0.00 | 0 | 0 | 0 | 0 |
| 0.5 | 0.707 | 0.50 | 0.35 | 0.5 | 0.35 | 0.707 | 1.06 |
| 1.0 | 1.00 | 1.00 | 1.00 | 1.0 | 0.5 | 1 | 1.5 |
| 1.5 | 1.22 | 1.50 | 1.84 | 1.5 | 0.61 | 1.22 | 1.84 |
| 2.0 | 1.41 | 2.00 | 2.83 | 2.0 | 0.71 | 1.41 | 2.12 |



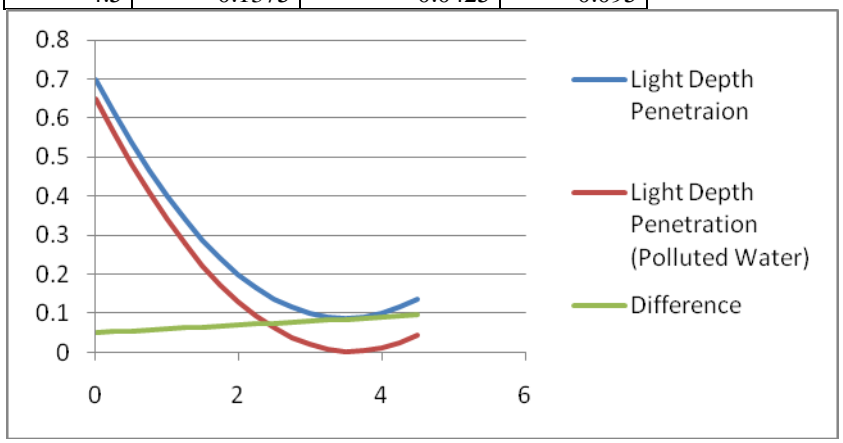




Observations: (1) the coefficient *a* scales the *y*-axis, with the magnitude increasing as *a* increases. (2) *b* controls the shape, with $b=1$ being linear, $b > 1$ being concave up (increasing rate), $b < 1$ being concave down (decreasing slope).

**Problem 2-17.**

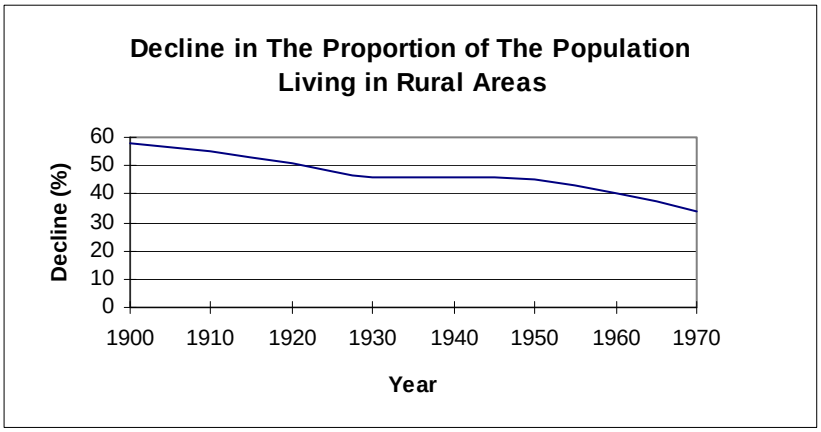
| Depth | Clean Water | Polluted Water | Difference |
|---|---|---|---|
| 0 | 0.7 | 0.65 | 0.05 |
| 0.25 | 0.615625 | 0.563125 | 0.0525 |
| 0.5 | 0.5375 | 0.4825 | 0.055 |
| 0.75 | 0.465625 | 0.408125 | 0.0575 |
| 1 | 0.4 | 0.34 | 0.06 |
| 1.25 | 0.340625 | 0.278125 | 0.0625 |

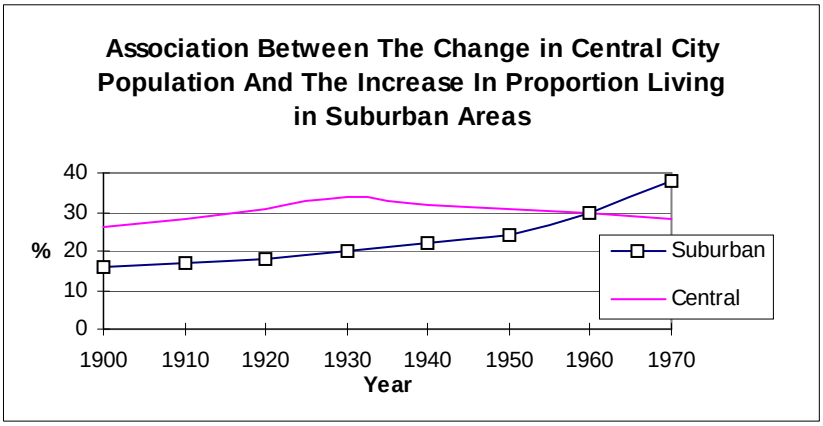
| | | | |
|---|---|---|---|
| 1.5 | 0.2875 | 0.2225 | 0.065 |
| 1.75 | 0.240625 | 0.173125 | 0.0675 |
| 2 | 0.2 | 0.13 | 0.07 |
| 2.25 | 0.165625 | 0.093125 | 0.0725 |
| 2.5 | 0.1375 | 0.0625 | 0.075 |
| 2.75 | 0.115625 | 0.038125 | 0.0775 |
| 3 | 0.1 | 0.02 | 0.08 |
| 3.25 | 0.090625 | 0.008125 | 0.0825 |
| 3.5 | 0.0875 | 0.0025 | 0.085 |
| 3.75 | 0.090625 | 0.003125 | 0.0875 |
| 4 | 0.1 | 0.01 | 0.09 |
| 4.25 | 0.115625 | 0.023125 | 0.0925 |
| 4.5 | 0.1375 | 0.0425 | 0.095 |




The light penetrates less polluted water. The difference increases as the depth of the water increases.

**Problem 2-18.**

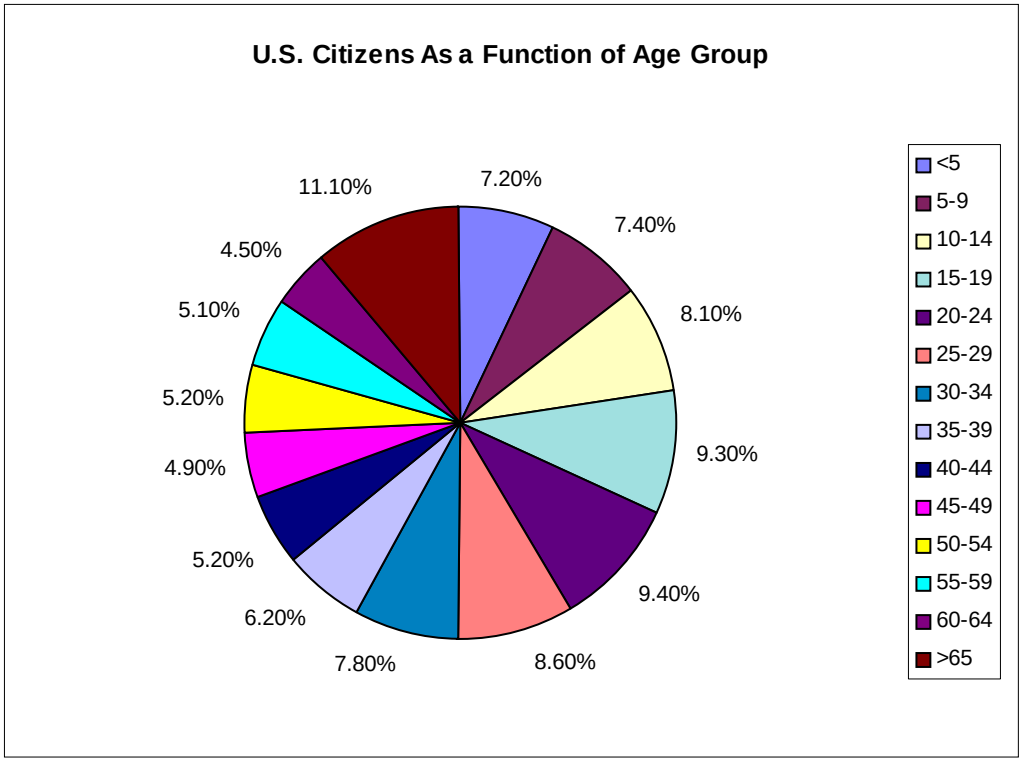





**Problem 2-19.**

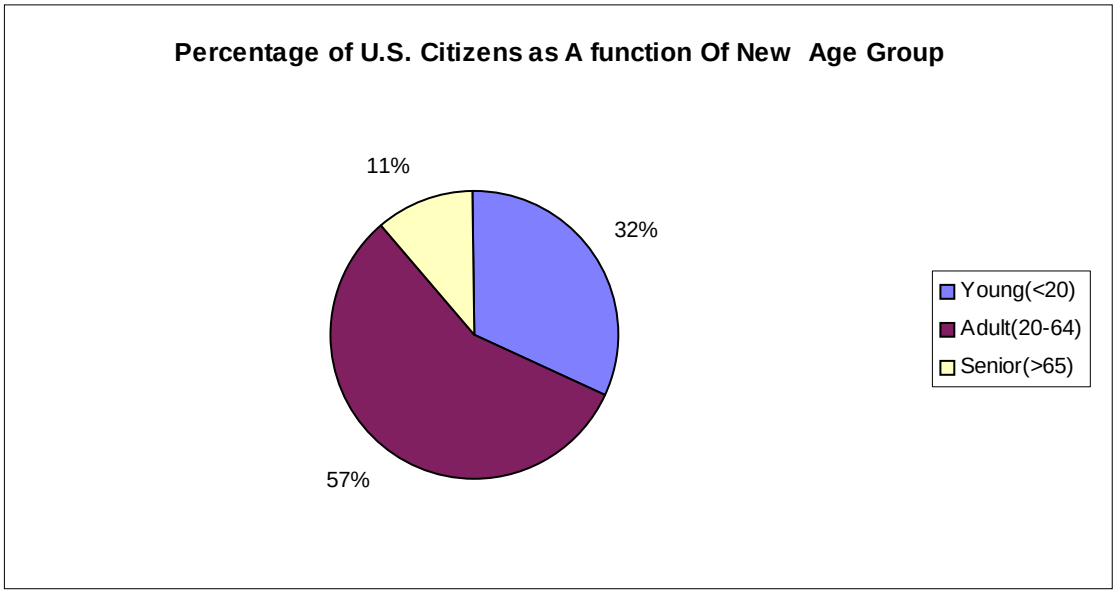
| Type | **Use** | **Independent Variable** | **Dependent Variable** | **Example** |
|---|---|---|---|---|
| **Area charts** | Three-dimensional data that include both nominal and interval-independent variables. | Measured on an interval scale and shown on the abscissa | Measured on the interval scale and cumulated over all values of the nominal variable | Analyzing the traffic at an intersection |
| **Pie charts** | Graphically present data recorded as fractions, percentages, or proportions | Measured on an interval scale | | Breakdown according to form of transportation in a shipping company |
| **Bar charts** | Data recorded on an interval scale | One or more recorded on nominal or ordinal scales | A magnitude or a fraction | Reinforcing steel production |
| **Column charts** | Similar to bar charts | Used for the abscissa. | Expressed as a percentage (or fraction) of a total. It is shown as the ordinate. | Capacity of desalination plants |
| **Scatter diagrams** | When both variables are measured on interval or ratio scales. | Shown on the abscissa | Shown on the ordinate | Yield strength and carbon content |
| **Line graphs** | Illustrate mathematical equations | Measured on interval or ratio scales | Measured on interval or ratio scales. Shown as the ordinate | Peak discharge rates |
| **Combination charts** | Experimental data and theoretical (or fitted) prediction equations. Two or more of the above mentioned methods are used to present data. | | | Operation of a marine vessel |
| **Three dimensional charts** | Describe the relationships among three variables. | | | Any of the above mentioned examples can be displayed. |

**Problem 2-20.**



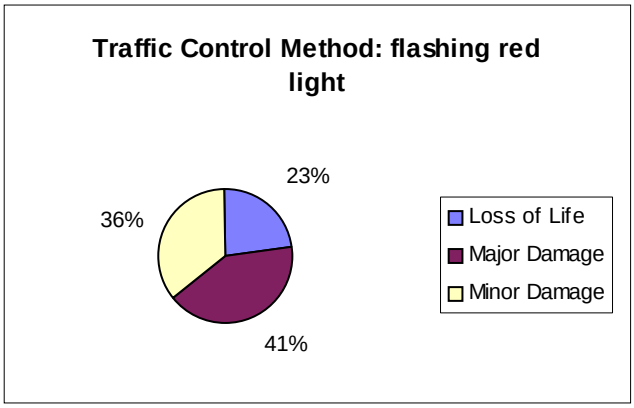

The pie chart shows the percentage of the U.S. citizens with respect to their age group. It would be also meaningful to classify the citizens as young, adult, and senior. The following table shows the distribution of the citizens as a function of the new classification:

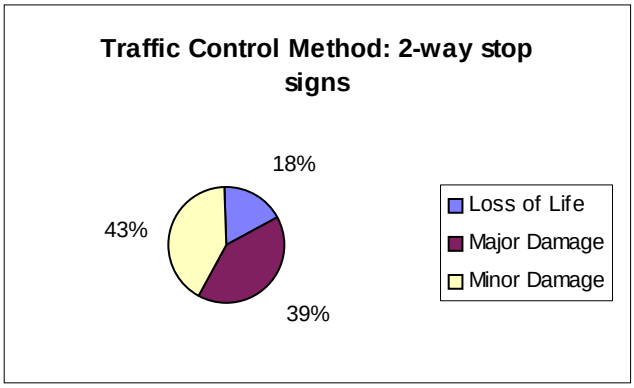
| Age group | Young(<20) | Adult(20-64) | Senior(>65) |
|---|---|---|---|
| Percentage | 32 | 56.9 | 11.1 |

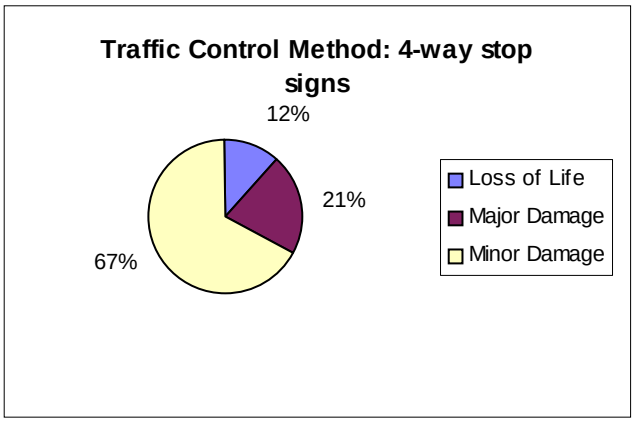


**Problem 2-21.**

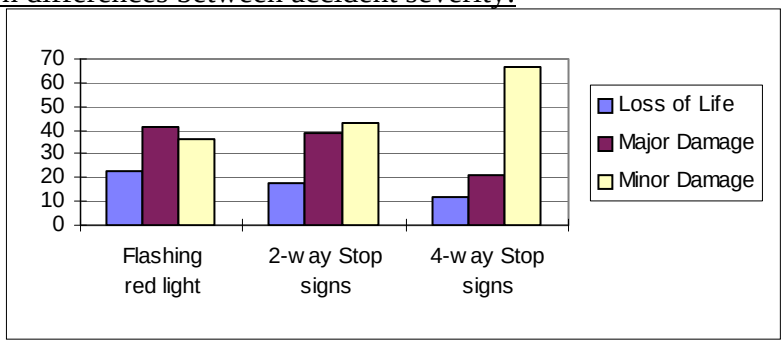
(a). Pie charts:



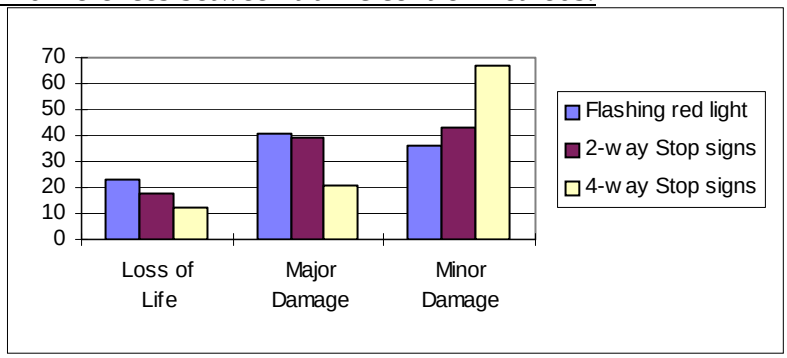



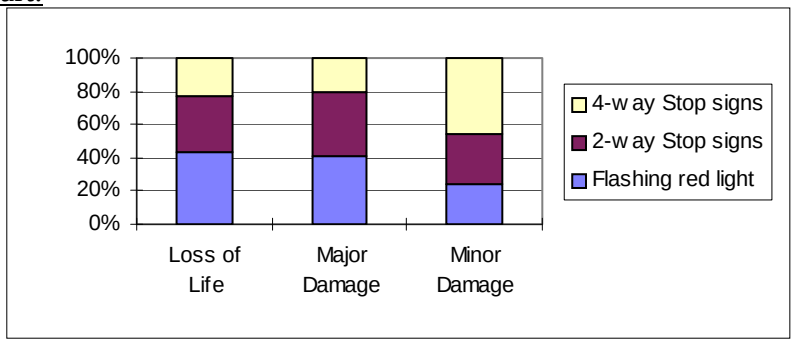

(b). Emphasis on differences between accident severity:



(c). Emphasis on differences between traffic control methods:



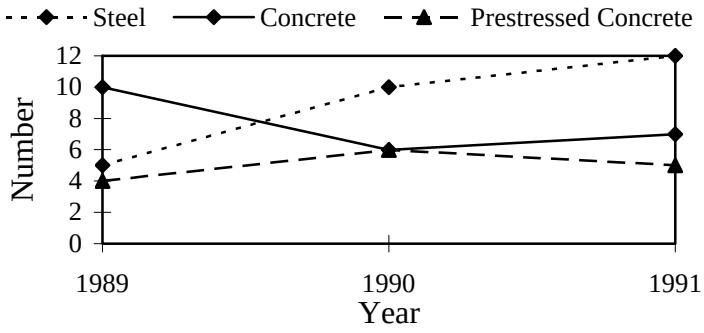

(d). Column chart:

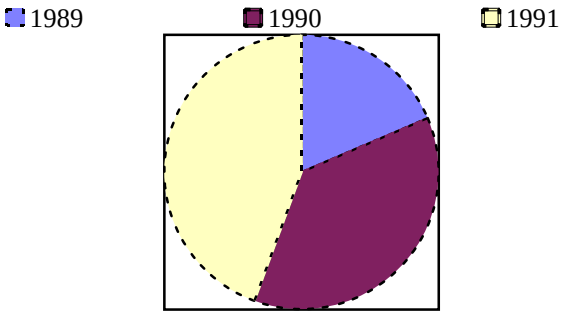



(e). The advantage of pie charts over bar charts is that they show the breakdown of accident severity as a proportion or percentage of 100%. On the other hand, bar charts clearly illustrate the increase or decrease of the rate of accident severity type from year to year. Column charts also show the breakdown of severity as a proportion of the whole, but for example, it is unclear as to the exact percentage of major damages in a 2-way stop sign control method.

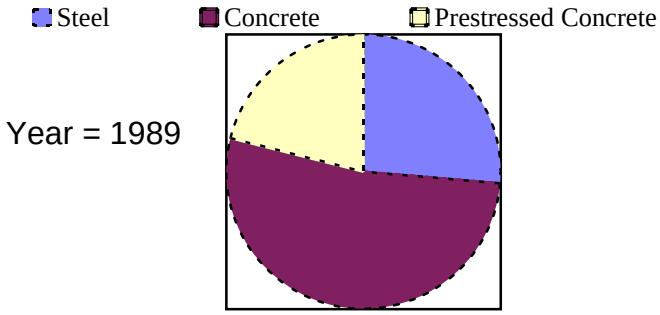
---

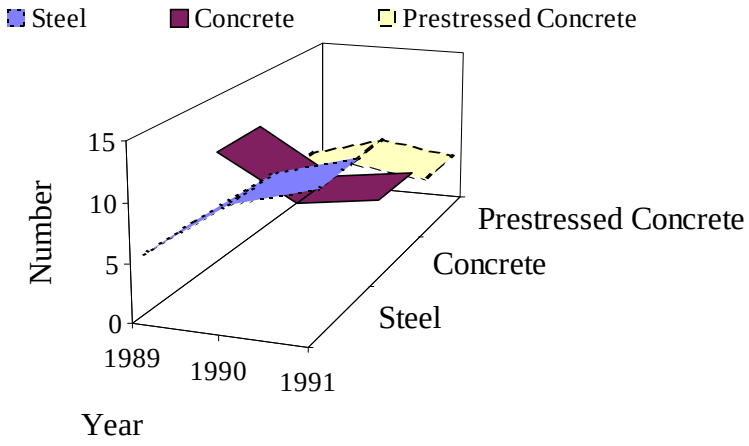
**Problem 2-22.**






For an example year,



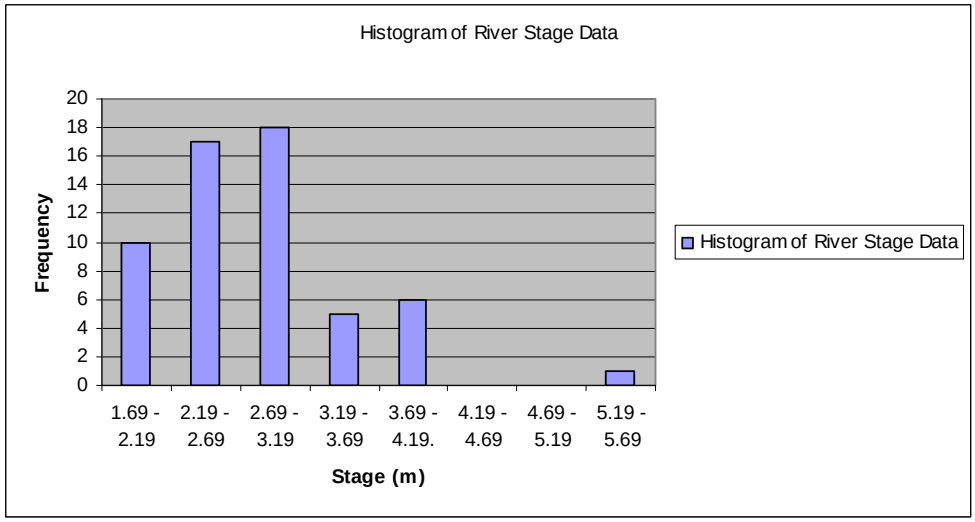



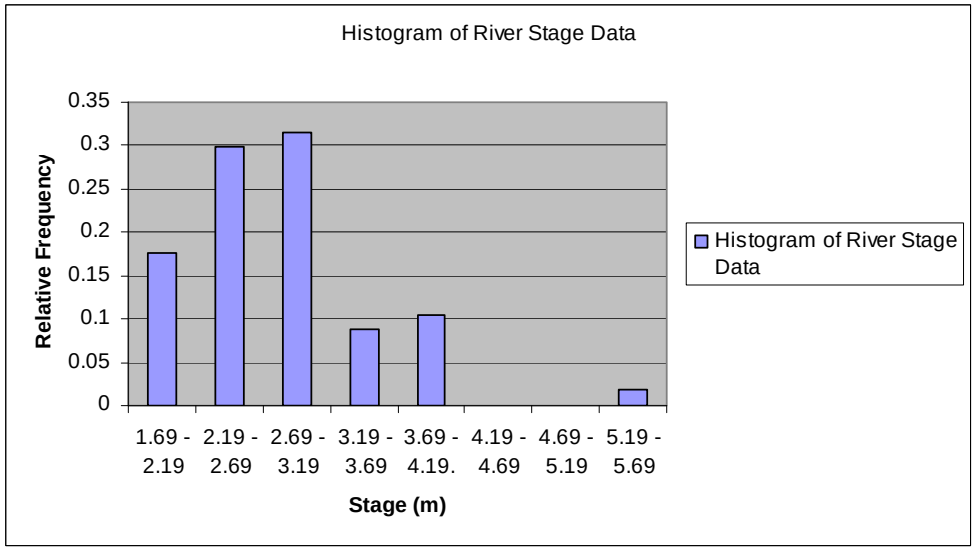

## 2.4. Histograms and Frequency Diagrams

**Problem 2-23.**

0.5m intervals

| Interval | Frequency | Rel. Freq. |
|---|---|---|
| 1.69 - 2.19 | 10 | 0.175439 |
| 2.19 - 2.69 | 17 | 0.298246 |
| 2.69 - 3.19 | 18 | 0.315789 |
| 3.19 - 3.69 | 5 | 0.087719 |
| 3.69 - 4.19. | 6 | 0.105263 |
| 4.19 - 4.69 | 0 | 0 |
| 4.69 - 5.19 | 0 | 0 |
| 5.19 - 5.69 | 1 | 0.017544 |
| Total | 57 | 0.982456 |

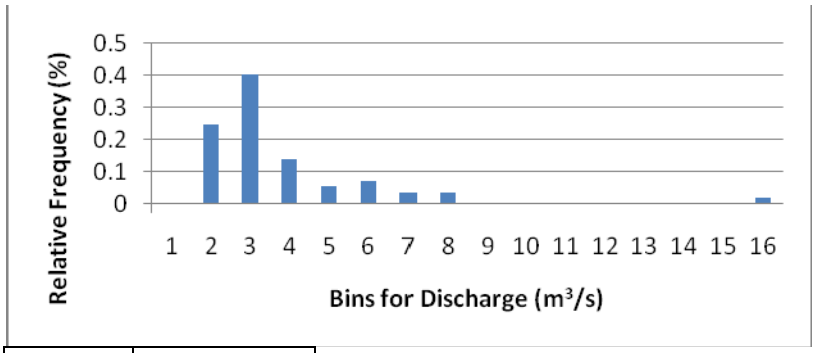



The difference between this relative frequency histogram and Figure 2-19 is that the relative frequencies are much smaller due to a smaller interval size.

---

**Problem 2-24.**



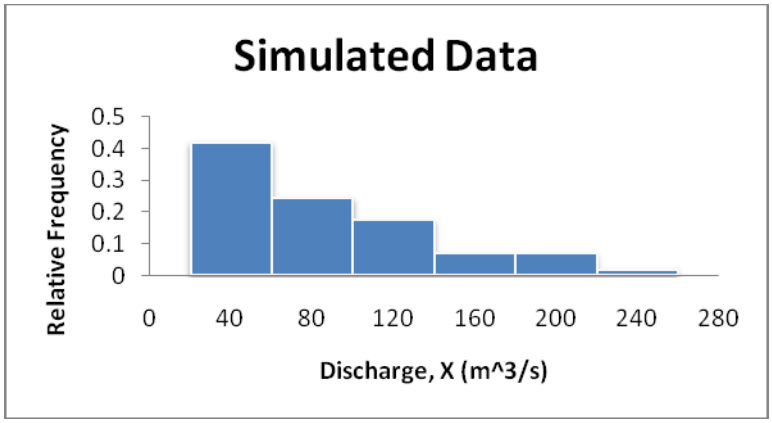

| Bin | Range (m^3/s) |
|---|---|
| 1 | <0 |
| 2 | 0-25 |
| 3 | 25-50 |
| 4 | 50-75 |
| 5 | 75-100 |

| | |
|---|---|
| 6 | 100-125 |
| 7 | 125-150 |
| 8 | 150-175 |
| 9 | 175-200 |
| 10 | 200-225 |
| 11 | 225-250 |
| 12 | 250-275 |
| 13 | 275-300 |
| 14 | 300-325 |
| 15 | 325-350 |
| 16 | 350-375 |

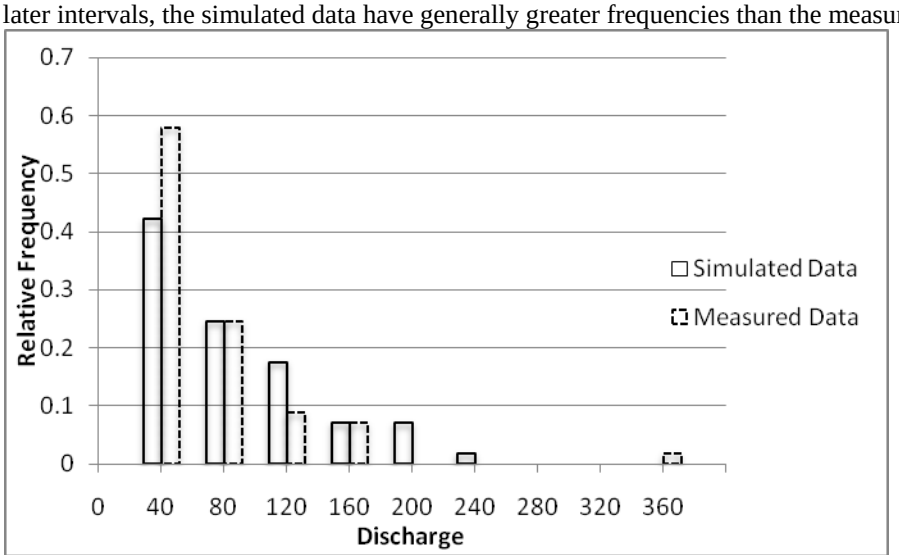
The shape of this histogram is slightly different than Figure 2-20 in the book. Both histograms are highly skewed. With the smaller bin sizes in this histogram, you are able to see more variations in the data and the shape looks more bell-like. Instead of the frequency constantly decreasing it goes up and down and up and down but maintains its overall shape.

---

**Problem 2-25.**




The data vary. In the first interval, the measured frequencies are more numerous than the simulated data; but in the later intervals, the simulated data have generally greater frequencies than the measured data.




The simulated data can has some differences with the measured data and the usage of the simulation should be based on the reason why the simulation is done in the first place.

---

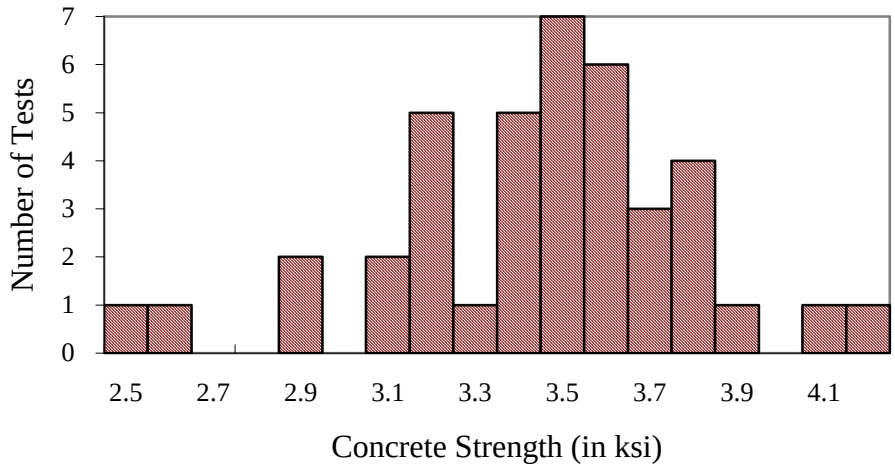
**Problem 2-26.**

Using an interval of 0.1 ksi, the following table can be constructed.

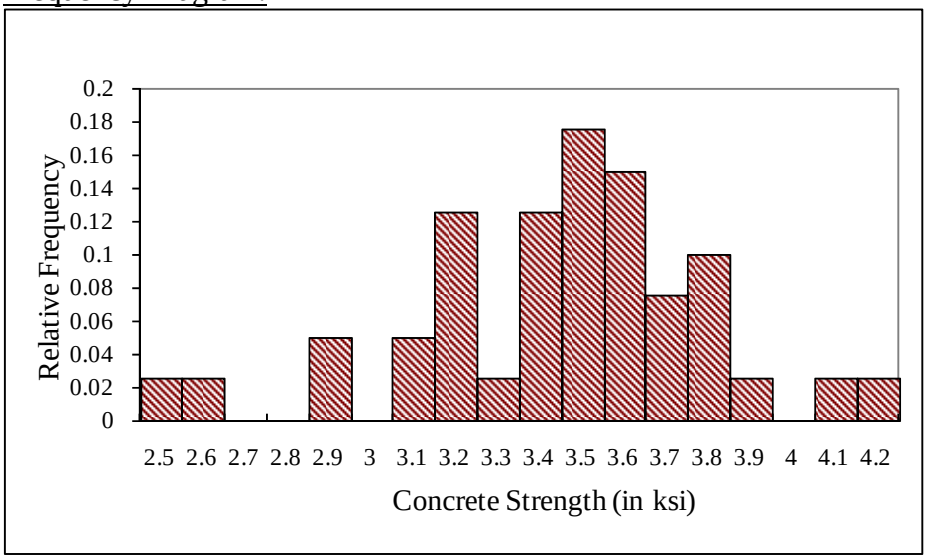
| Mid-interval | Count | Frequency (f) | Cumulative value | x(count) | $x^2$(count) |
|---|---|---|---|---|---|
| 2.5 | 1 | 0.025 | 1 | 2.5 | 6.25 |
| 2.6 | 1 | 0.025 | 2 | 2.6 | 6.76 |
| 2.7 | 0 | 0 | 2 | 0 | 0 |
| 2.8 | 0 | 0 | 2 | 0 | 0 |
| 2.9 | 2 | 0.05 | 4 | 5.8 | 16.82 |
| 3 | 0 | 0 | 4 | 0 | 0 |
| 3.1 | 2 | 0.05 | 6 | 6.2 | 19.22 |
| 3.2 | 5 | 0.125 | 11 | 16 | 51.2 |
| 3.3 | 1 | 0.025 | 12 | 3.3 | 10.89 |
| 3.4 | 5 | 0.125 | 17 | 17 | 57.8 |
| 3.5 | 7 | 0.175 | 24 | 24.5 | 85.75 |
| 3.6 | 6 | 0.15 | 30 | 21.6 | 77.76 |
| 3.7 | 3 | 0.075 | 33 | 11.1 | 41.07 |
| 3.8 | 4 | 0.1 | 37 | 15.2 | 57.76 |
| 3.9 | 1 | 0.025 | 38 | 3.9 | 15.21 |
| 4 | 0 | 0 | 38 | 0 | 0 |
| 4.1 | 1 | 0.025 | 39 | 4.1 | 16.81 |
| 4.2 | 1 | 0.025 | 40 | 4.2 | 17.64 |
| Total | 40 | 1 | | 138 | 480.94 |

Note: A larger interval can be used and might produce better results than the interval of 0.1 ksi used herein.

Histogram:

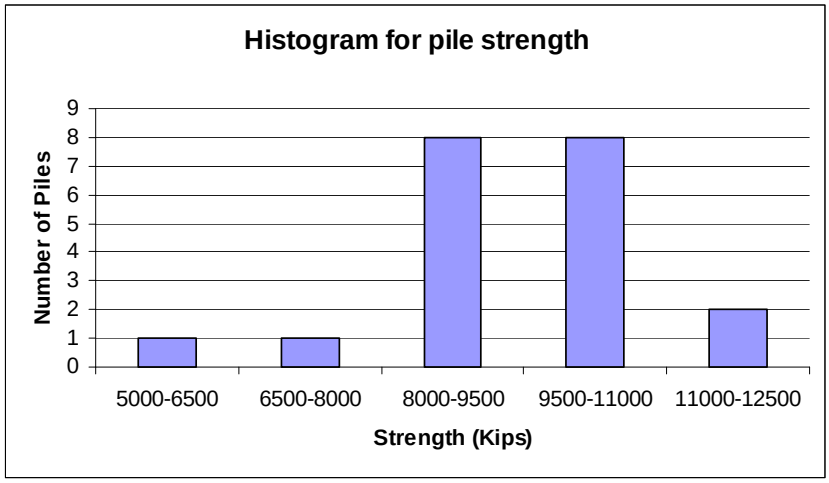


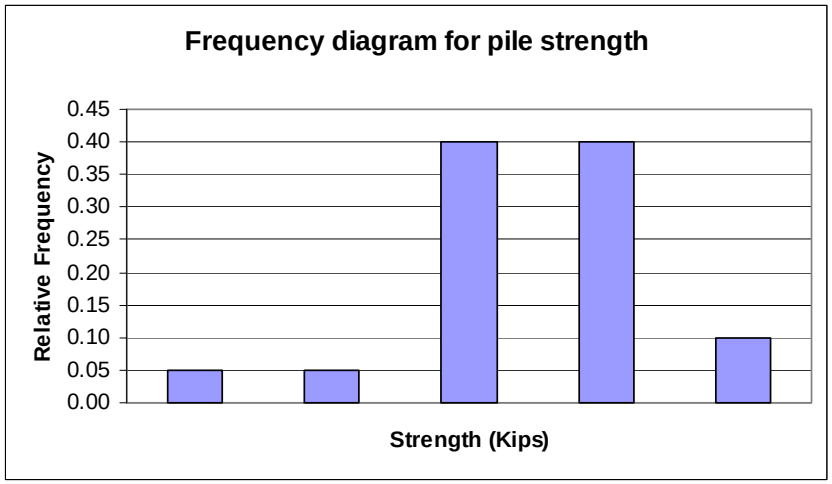
Frequency Diagram:



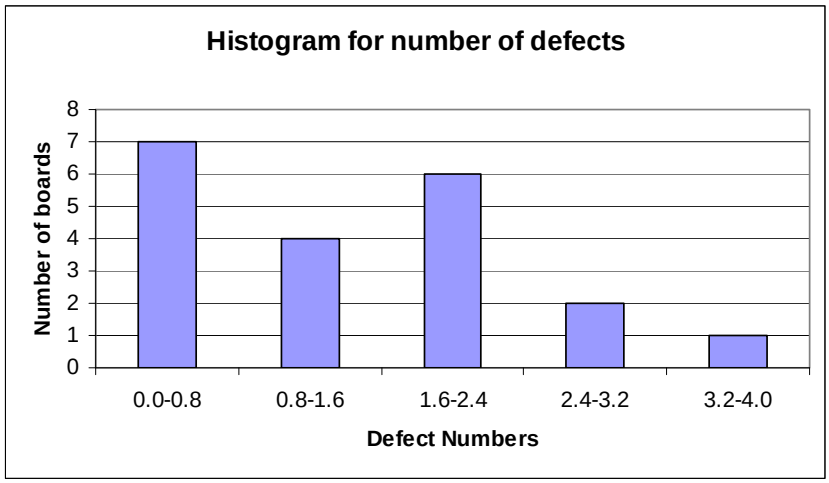

---

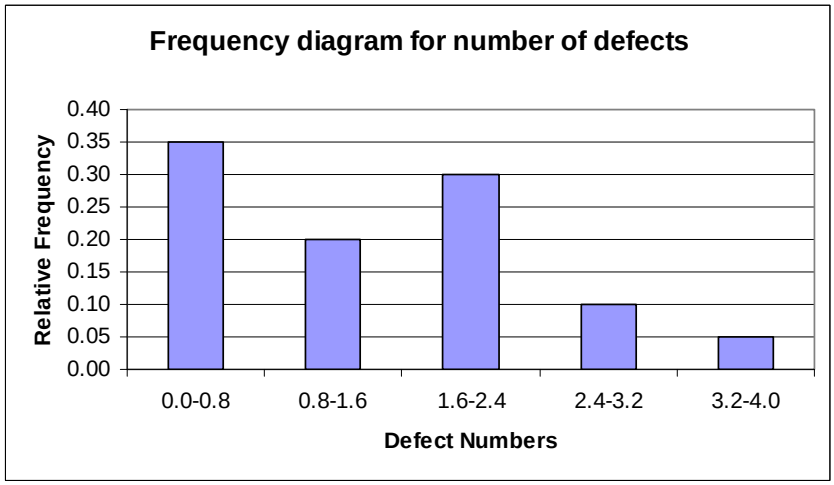
**Problem 2-27.**



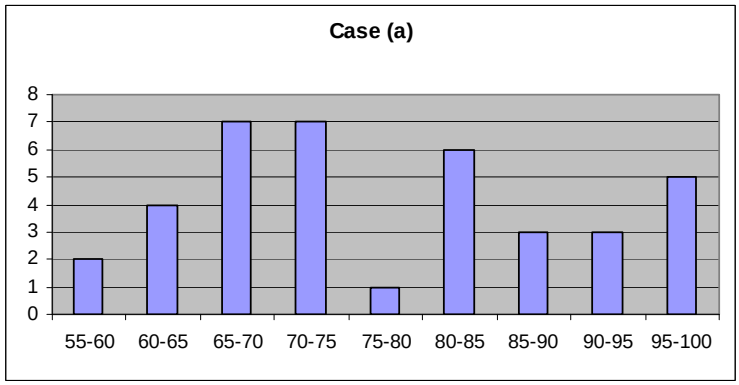



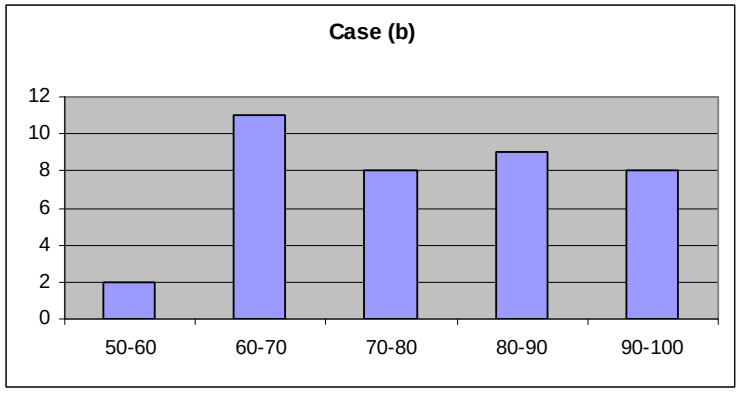
**Problem 2-28.**

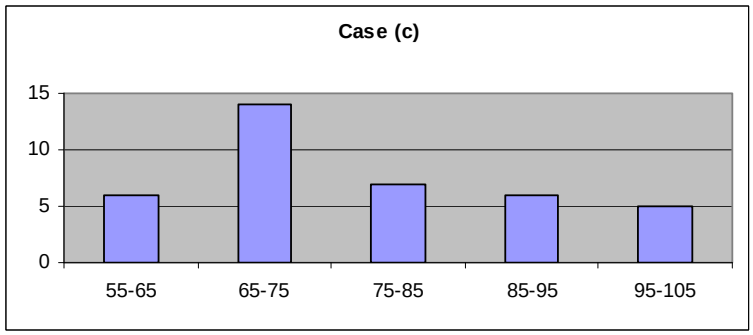



**Problem 2-29.**






While based on the same data, the histograms give different impressions of the grade distribution. Figure (a) indicates a two-peaks distribution, while Figure (b) suggests a uniform distribution and Figure (c) suggests a one-peak distribution.

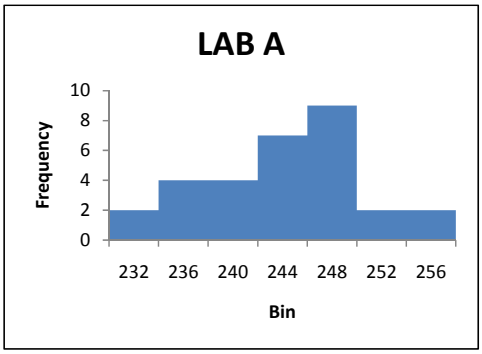
Observations: (1) Histograms based on small samples can be misleading; (2) For small and moderate samples, histograms should be developed for different cell widths and cell bounds before making conclusions about the data.

**Problem 2-30.**

| Sample # | LAB A | LAB B |
|---|---|---|
| 1 | 232 | 241 |
| 2 | 234 | 243 |
| 3 | 236 | 243 |
| 4 | 237 | 244 |
| 5 | 237 | 244 |
| 6 | 239 | 244 |
| 7 | 241 | 246 |
| 8 | 241 | 246 |
| 9 | 243 | 247 |
| 10 | 243 | 247 |
| 11 | 244 | 247 |
| 12 | 246 | 248 |
| 13 | 246 | 248 |
| 14 | 246 | 249 |
| 15 | 246 | 249 |
| 16 | 247 | 249 |
| 17 | 247 | 251 |
| 18 | 248 | 251 |
| 19 | 248 | 251 |
| 20 | 248 | 252 |
| 21 | 249 | 252 |
| 22 | 249 | 253 |
| 23 | 251 | 253 |
| 24 | 251 | 253 |
| 25 | 251 | 254 |

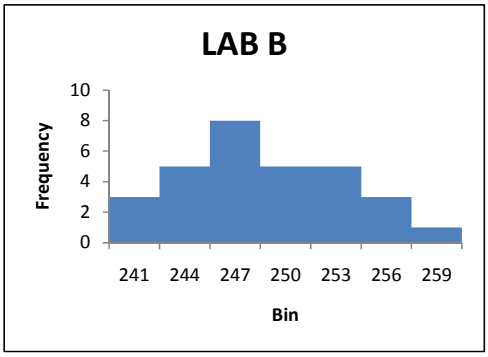
| Mean | Std. Dev. | k | min | max | range | interval |
|---|---|---|---|---|---|---|
| 245.63 | 6.30 | 5.87 | 232.00 | 256.00 | 24.00 | 4.00 |

| Bin | Frequency |
|---|---|
| 232 | 2 |
| 236 | 4 |
| 240 | 4 |
| 244 | 7 |
| 248 | 9 |
| 252 | 2 |
| 256 | 2 |




| Mean | Std. Dev. | k | min | max | range | interval |
|---|---|---|---|---|---|---|
| 249.60 | 4.69 | 5.87 | 241.00 | 259.00 | 18.00 | 3.00 |

| Bin | Frequency |
|---|---|
| 241 | 3 |
| 244 | 5 |
| 247 | 8 |
| 250 | 5 |
| 253 | 5 |




The average from Lab B is closer to the known concentration of 250 ppb than the average from Lab A. Also, the measurements from Lab B are more consistent than the measurements from Lab A because the scatter is smaller – this means that the values deviates to lesser extent from the average. Overall, Lab B presents the best yearly data.

**Problem 2-31.**

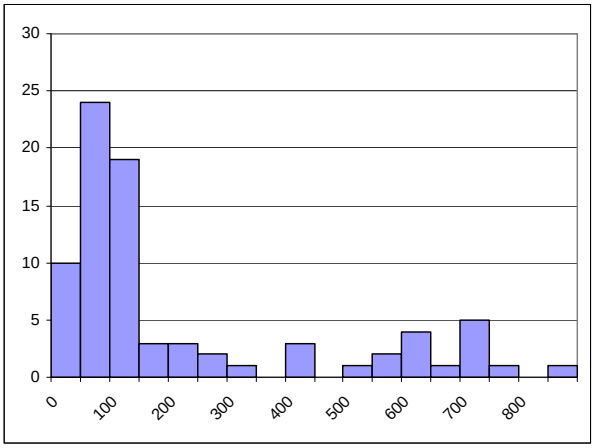
Using the random number generation feature of excel, you could estimate various sample sizes (*n* = 25,50,100, ect.) to find rough boundaries which overestimate and underestimate the population and then iterate to find an appropriate sample size n-ideal. Continue to generate additional values (increasing sample size) and periodically re-compute the ordinates of the relative frequency histogram of the simulated data. Compare each ordinate of the simulated and measured data histograms by computing the absolute value of the difference. When the difference is less than some tolerance, say 0.01%, then assume the sample size of the generated data provides data that represents the measured data. The assumed and simulated data do not agree. The sample size should be increased until the two data sets agree, because and increased sample size will yield

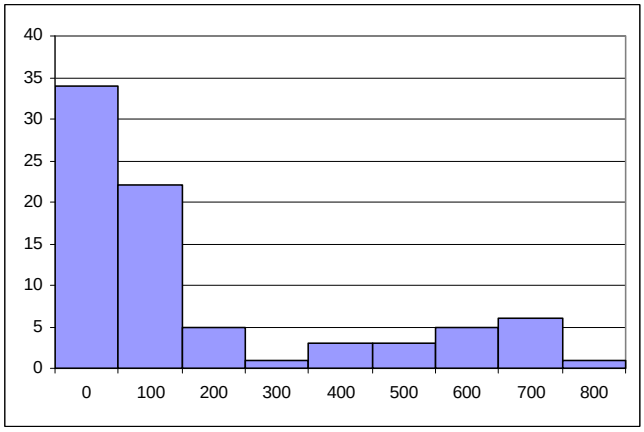
more accurate simulated data. I would increase the sample size until the differences are statistically insignificant.

---

**Problem 2-32.**

| | | |
|---|---|---|
| 0-49 | 10 | |
| 50-99 | 24 | 34 |
| 100-149 | 19 | |
| 150-199 | 3 | 22 |
| 200-249 | 3 | |
| 250-299 | 2 | 5 |
| 300-349 | 1 | |
| 350-399 | 0 | 1 |
| 400-449 | 3 | |
| 450-499 | 0 | 3 |
| 500-549 | 1 | |
| 550-599 | 2 | 3 |
| 600-649 | 4 | |
| 650-699 | 1 | 5 |
| 700-749 | 5 | |
| 750-799 | 1 | 6 |
| 800-849 | 0 | |
| 850-899 | 1 | 1 |



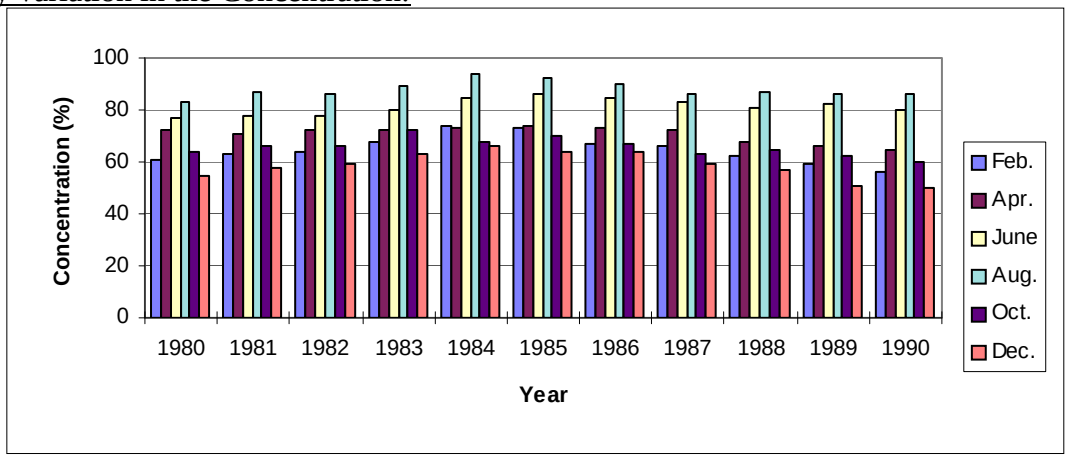




Observations: (1) If the interval is too small, cell ordinates may appear with gaps showing random variation; (2) for samples with most values in a few cells, the shape of the distribution is not decisive, even for moderate samples.

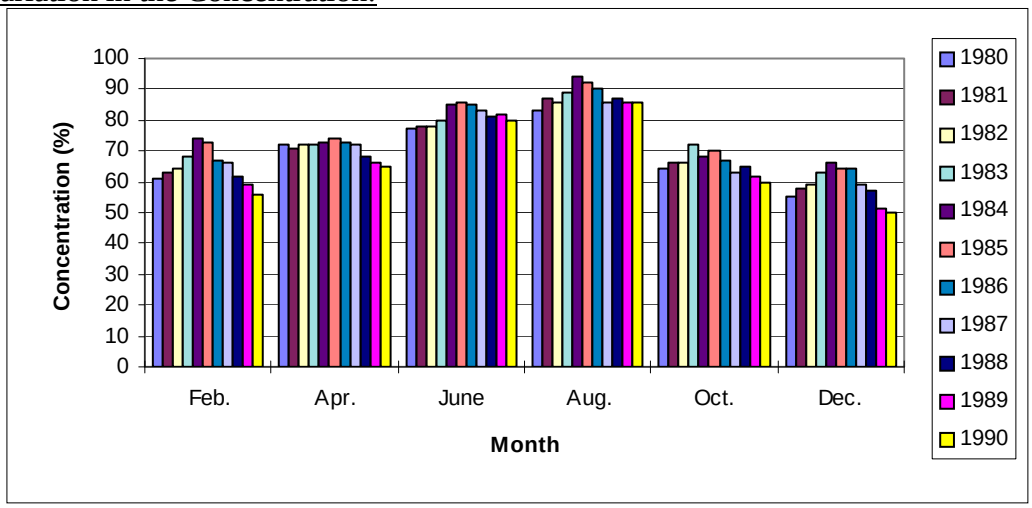
## 2.6. Descriptive Measures

**Problem 2-33.**

Monthly variation in the Concentration:




Annual variation in the Concentration:




Both variables are important. For example, the annual variation is evident for Feb., but less significant for April. The monthly variation, which is expected, is very evident in the first figure.

**Problem 2-34.**

Central tendency measures:

a. Mean = $\frac{\sum_{i=1}^{40} x_i}{40} = 3.45\ ksi$

b. Median = $(x_{20}+x_{21})/2$ = 3.5 ksi

c. Mode = value of highest frequency = 3.5 ksi

**Problem 2-35.**

Central tendency measures:

a. Mean = $\frac{\sum_{i=1}^{20} x_i}{20} = 9564.95\, kips$

b. Median = $(x_{10}+x_{11})/2$ = 9685.5 kips

c. Mode = value of highest frequency = No mode, no value occurred more than once.

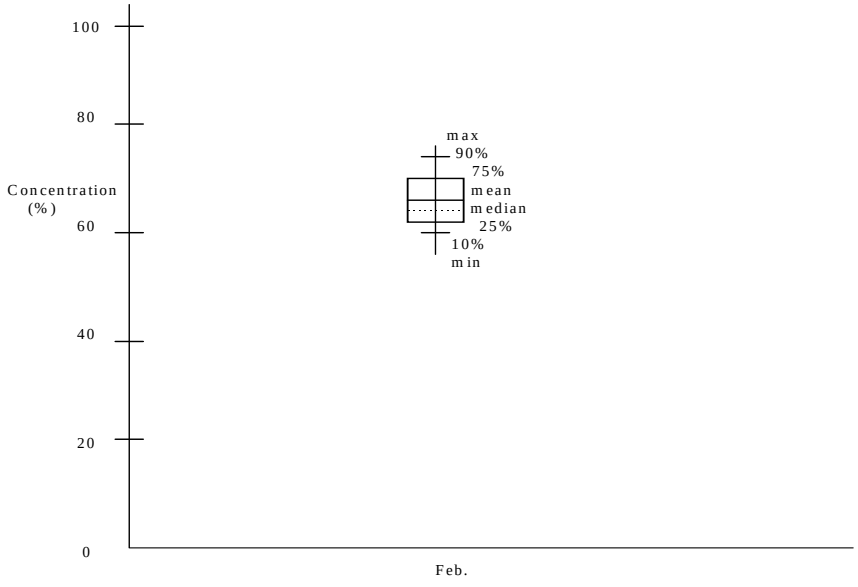
**Problem 2-36.**

Box-and-whisker plot data:

| | Feb. | Apr. | June | Aug. | Oct. | Dec. |
|---|---|---|---|---|---|---|
| Mean | 64.82 | 70.73 | 81.36 | 87.82 | 65.73 | 58.73 |
| Median | 64 | 72 | 81 | 87 | 66 | 59 |
| Min. | 56 | 65 | 77 | 83 | 60 | 50 |
| Max. | 74 | 74 | 86 | 94 | 72 | 66 |
| $x_p$=90 | 73 | 73 | 85 | 92 | 70 | 64 |
| $x_p$=75 | 67.5 | 72.5 | 84 | 89.5 | 67.5 | 63.5 |
| $x_p$=25 | 61.5 | 69.5 | 79 | 86 | 63.5 | 56 |
| $x_p$=10 | 59 | 66 | 78 | 86 | 62 | 51 |

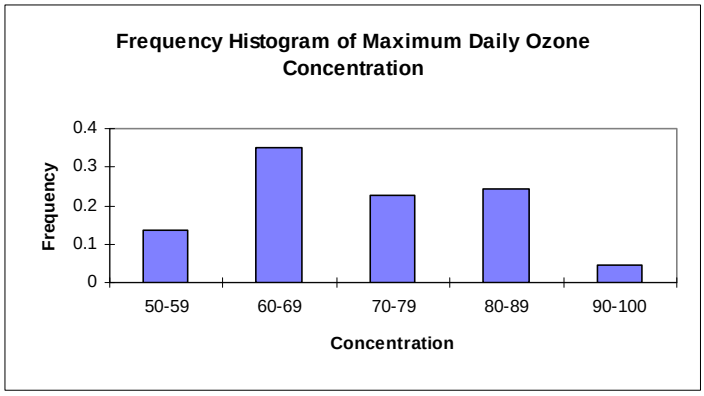
Box-and-whisker plot:

The following is the box-and-whisker plot constructed only for the month of February. For multiple box and whisker plots display, refer to Section 2.5.4 and Figure 2-16 of the textbook.




Frequency Histogram:

| Concentration | 50-59 | 60-69 | 70-79 | 80-89 | 90-100 |
|---|---|---|---|---|---|
| Frequency | 0.136 | 0.348 | 0.227 | 0.242 | 0.045 |




**Problem 2-37.**

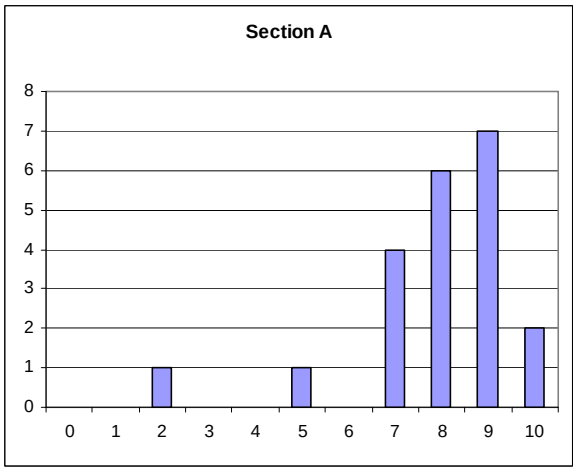
Mean = sum/6 = 165.9/6 = 27.65 mg/l

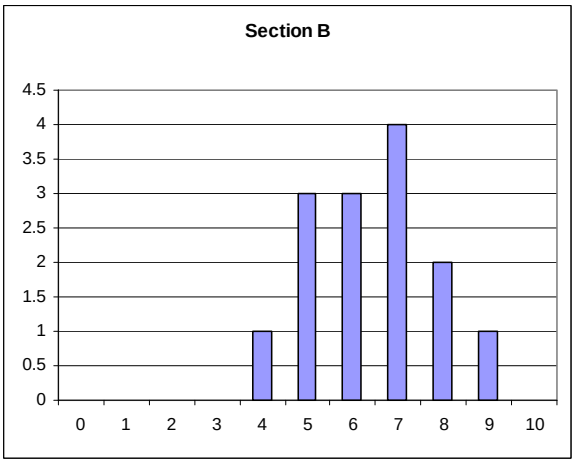
Median = 1.6 mg/l

The extreme value of 157.9 greatly affects the mean but not the median. In general, the median is much less sensitive to highly deviant measurements, which may be due to recording errors or random variation. For the data given, the mean value is 27.65 mg/L, while the median value is 1.6. The median is similar to 5 of the 6 measurements, while the average value is unlike any of the 6 measurements.

**Problem 2-38.**

| | Section A | Section B |
|---|---|---|
| mean | 7.58 | 5.95 |
| median | 8 | 6 |
| mode | 9 | 6 |




The grades in section B are bell-shaped and so the three measures of central tendency are very similar. The grades in section A are skewed towards the lower values so the three measures show a greater difference with the mode much larger than the mean.

**Problem 2-39.**

$\overline{X} = \sum_{i=1}^{k} f_i x_i$ where k is the integer number of scores of $x_i$ and $f_i$ is the frequency of the number of scores $x_i$. The equation provides a weighted sum of the values where $f_i$ are the weights that must add up to one.

**Problem 2-40.**

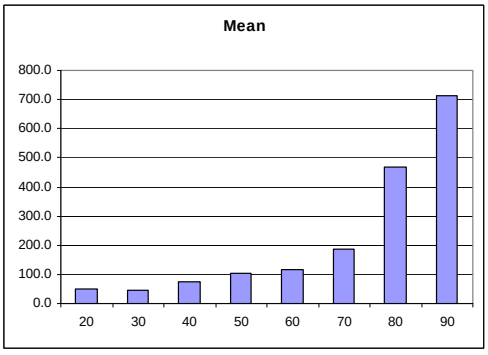
$$\overline{B} = \frac{1}{n}\sum_{i=1}^{n} B_i = \frac{190}{24} = 7.917$$

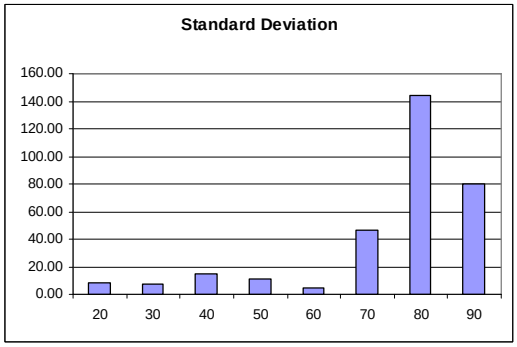
$$S_B = \left[\frac{1}{n-1}\sum (B_i - \overline{B})^2\right]^{0.5} = \left[\frac{1}{23}\sum (B_i - 7.917)^2\right]^{0.5} = 2.669$$

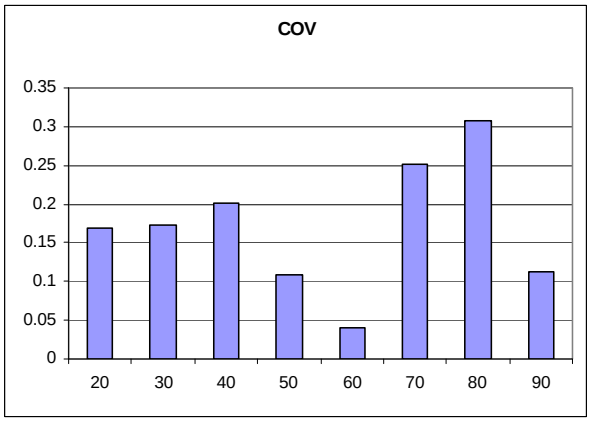
$$COV(B) = S_B / \overline{B} = 0.337$$

**Problem 2-41.**

| Decade | Mean | St. Dev. | COV |
|---|---|---|---|
| 1920-29 | 48.5 | 8.21 | 0.169 |
| 1930-39 | 45.1 | 7.78 | 0.173 |
| 1940-49 | 73.9 | 14.94 | 0.202 |
| 1950-59 | 105.4 | 11.52 | 0.109 |
| 1960-69 | 116.8 | 4.76 | 0.041 |
| 1970-79 | 184.8 | 46.62 | 0.252 |
| 1980-89 | 468.9 | 144.4 | 0.308 |
| 1990-99 | 711.8 | 80.11 | 0.113 |



The mean shows an exponentially increasing trend. Generally, the standard deviation increases near the end. The COV varies randomly over the decades.

---

**Problem 2-42.**

Dispersion measures:

a. Variance = $\frac{\sum_{i=1}^{40}(x_i - mean)^2}{40-1} = 0.124103\ ksi^2$

b. Standard deviation = Square root of variance = 0.35228 ksi ksi

c. Coefficient of variation = Standard deviation/mean = 0.1021

---

**Problem 2-43.**

Dispersion measures:

a. Variance = $\frac{\sum_{i=1}^{20}(x_i - mean)^2}{20-1} = 2270966\ kips^2$

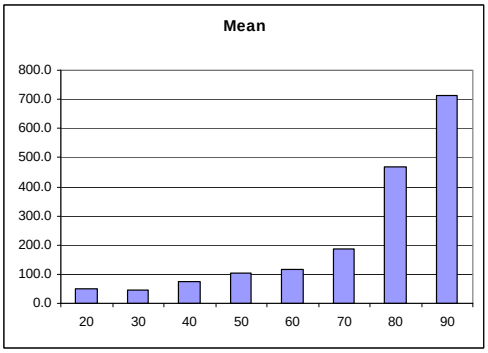
b. Standard deviation = Square root of variance = 1507 kips

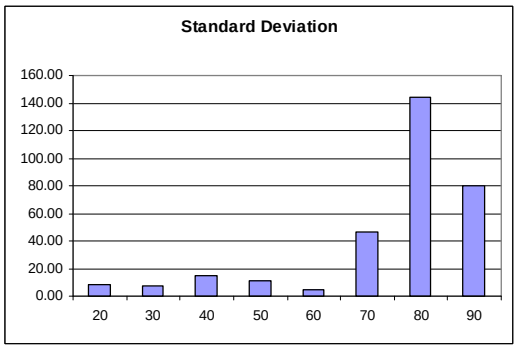
c. Coefficient of variation = Standard deviation/mean = 0.1575

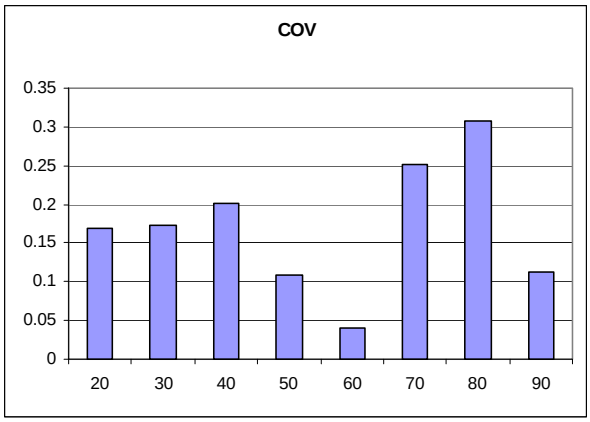
---

**Problem 2-44.**

| Decade | Mean | St. Dev. | COV |
|---|---|---|---|
| 1920-29 | 48.5 | 8.21 | 0.169 |
| 1930-39 | 45.1 | 7.78 | 0.173 |
| 1940-49 | 73.9 | 14.94 | 0.202 |
| 1950-59 | 105.4 | 11.52 | 0.109 |
| 1960-69 | 116.8 | 4.76 | 0.041 |
| 1970-79 | 184.8 | 46.62 | 0.252 |
| 1980-89 | 468.9 | 144.4 | 0.308 |
| 1990-99 | 711.8 | 80.11 | 0.113 |






The mean shows an exponentially increasing trend. Generally, the standard deviation increases near the end. The COV varies randomly over the decades.

---

**Problem 2-45.**

$$S^2 = \frac{1}{n-1}\sum (x-\bar{x})^2 = \frac{1}{n-1}\sum (x^2 - 2\bar{x}x + \bar{x}^2)$$

$$= 1/_{n-1}\left[\sum x^2 - 2\bar{x}\sum x + \bar{x}^2 \sum i\right]$$

$$= \frac{1}{n-1}\left[\sum x^2 - 2n\bar{x}^2 + n\bar{x}^2\right] = \frac{1}{n-1}\left[\sum x^2 - n\bar{x}^2\right]$$

$$= \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$$

---

**Problem 2-46.**

$\overline{Y} = \overline{X} / 3.281$

$S_y = S_x / 3.281$

$S_y^{\,2} = S_x^{\,2} / (3.281)^2$

The general rule is that the units of $\overline{Y}$ equals the units of $\overline{X}$ multiplied by the multiplication constant for transforming *X* to *Y*. The variance of *Y* is the square of the conversion factor times the variance of *X*.

---

**Problem 2-47.**

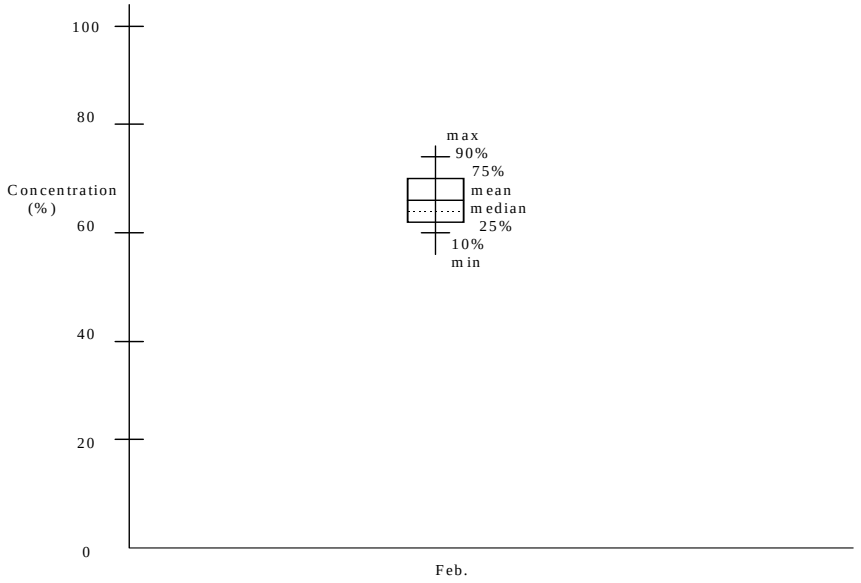
Box-and-whisker plot data:

| | Feb. | Apr. | June | Aug. | Oct. | Dec. |
|---|---|---|---|---|---|---|
| Mean | 64.82 | 70.73 | 81.36 | 87.82 | 65.73 | 58.73 |
| Median | 64 | 72 | 81 | 87 | 66 | 59 |
| Min. | 56 | 65 | 77 | 83 | 60 | 50 |
| Max. | 74 | 74 | 86 | 94 | 72 | 66 |
| $x_p$=90 | 73 | 73 | 85 | 92 | 70 | 64 |
| $x_p$=75 | 67.5 | 72.5 | 84 | 89.5 | 67.5 | 63.5 |
| $x_p$=25 | 61.5 | 69.5 | 79 | 86 | 63.5 | 56 |

| | | | | | | |
|---|---|---|---|---|---|---|
| $x_p$=10 | 59 | 66 | 78 | 86 | 62 | 51 |

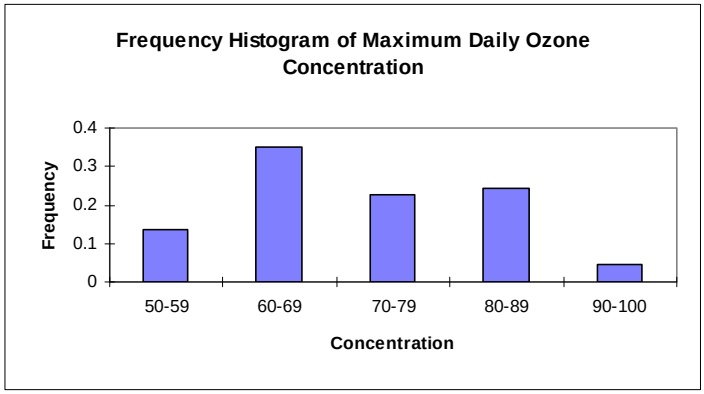
Box-and-whisker plot:

The following is the box-and-whisker plot constructed only for the month of February. For multiple box and whisker plots display, refer to Section 2.5.4 and Figure 2-16 of the textbook.




Frequency Histogram:

| Concentration | 50-59 | 60-69 | 70-79 | 80-89 | 90-100 |
|---|---|---|---|---|---|
| Frequency | 0.136 | 0.348 | 0.227 | 0.242 | 0.045 |




---

**Problem 2-48.**

Ranking the values for the 1920-59 period:

123,120,116,108,108,102,98,96,96,93,92,92,90,83,65,65,64,61,61,60,55,54,54,53,53,52, 52,51,51,50,49,49,47,47,46,39,38,38,30,28.

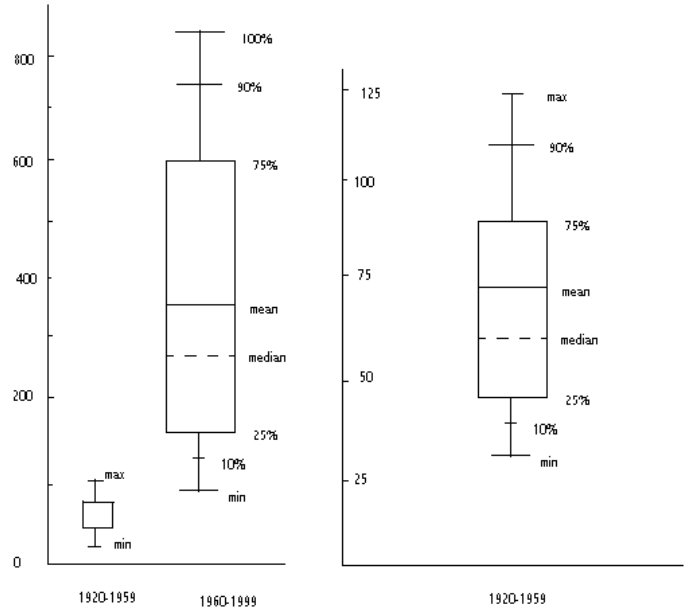
Ranking the values for the 1960-99 period:

870,775,739,736,725,707,700,656,629,619,611,609,581,574,537,426,418,407,317,274,251,229,215,210,187,165,155,145,140,128,123,121,121,120,120,115,114,113,112,109.

Thus the necessary characteristics for the two periods:

| | 1920-59 | 1960-99 |
|---|---|---|
| max | 123.0 | 870.0 |
| 90% | 108.0 | 725.0 |
| 75% | 93.0 | 611.0 |

| | | |
|---|---|---|
| mean | 68.2 | 370.6 |
| median | 57.5 | 262.5 |
| 25% | 49.0 | 123.0 |
| 10% | 38.0 | 114.0 |
| min | 28.0 | 109.0 |




---

**Problem 2-49.**

$\overline{A}$ = 2.042 Standard deviation, SD (A) = 1.681 $\overline{B}$ = 7.917 SD(B) = 2.669

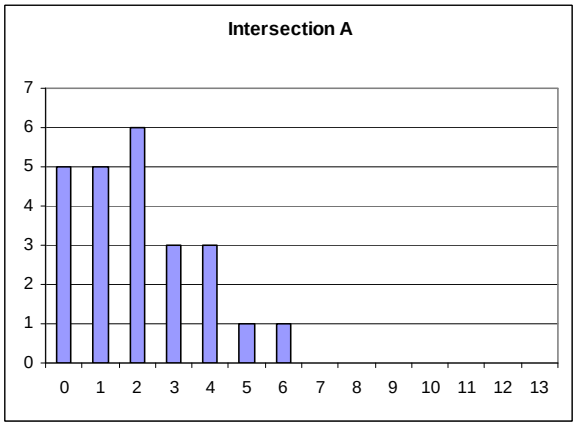
COV(A) = 0.823 COV(B) = 0.337

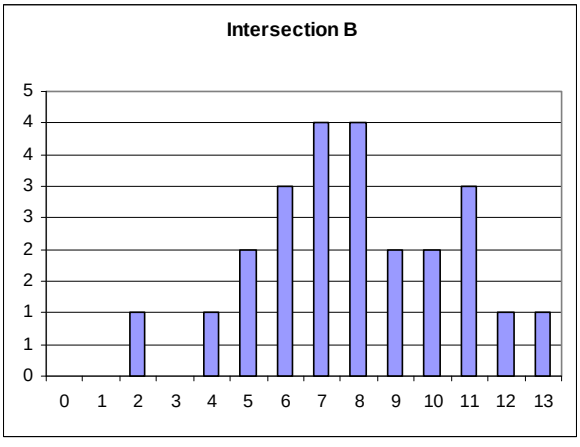
The moments indicate that the traffic control measures at A have reduced the mean number of accidents and the monthly variation in the number of accidents.

B-A = $\{3,2,4,8,7,7,5,6,4,7,7,2,3,12,6,2,6,9,6,6,10,10,1,8\}$

Mean = 5.875 SD(B-A) = 2.864 COV(B-A) = 0.487

The mean of the differences equals the difference of the means. The variation of the differences is larger than the variation of either A or B. The relative variation of the difference is slightly larger than the relative variation of the intersection where controls were not installed.



The bar charts for the number of accidents per month at the two intersections indicate that the accident rate at B has a higher mean and greater spread.