CHAPTER 1

ORGANIZATION OF DATA SETS

When you collect data, it comes to you in more or less a random fashion and unorganized. For example, what if you gave a 35 item test to a class of 50 students and collect the answer sheets. When you score the tests, the order of the scores will be in the same order in which you received the answer sheets. It is now up to you to do some "data organization" so that the pattern and trends of the data will be more obvious. This is the process of imposing some initial order on the data. This section explores some of the more popular and useful procedures for organizing both frequency and trend line data. **NOTE:** From time to time, I will use some **MINITAB** output to illustrate ways in which computers can assist you with the data analysis; however, more detail on the use of Minitab is presented in the **APPENDIX** in the back of the book.

FREQUENCY DATA

<u>Sorting Data</u>

Look at the two sets of data below as they "come off the shelf" so to speak. Assume that, as mentioned above, 50 students have taken a 35 item test; I have called one HardTest since the scores are very low, and the other EasyTest since the scores are much higher. Notice that for both the "HardTest" and the "EasyTest" variables, the data are listed in no particular order. While it is easy to see the high and low scores for the first set, it is not as easy for the second set.

```
HardTest
   6    4    7    2    4    6    5    2    5    4    5
   4    4    6    4   10    5    2    3   10    4    2
   3    2    5    5    6    4    7    6    9    3    4
   3    5    5    6    6    6    7    1    1    5    8
   4    8    4    4    6    4

EasyTest
  20   28   33   24   28   31   20   23   17   23   28
  30   26   30   29   21   20   23   29   23   28   25
  27   25   21   22   14   27   25   24   20   22   31
  30   20   19   27   20   23   26   23   22   30   26
  21   30   28   18   20   24
```

The first thing we could do would be to order each set of scores from high to low, or low to high. After sorting the data, it will be easy to see the lowest and highest values, and the range of the values (distance from top to bottom). If you print the sorted column, you obtain an ordered listing of the scores. Look at the data below which is simply the rearranged values from above.

```
HardSort
   1    1    2    2    2    2    2    3    3    3    3
   4    4    4    4    4    4    4    4    4    4    4
   4    4    5    5    5    5    5    5    5    5    5
   6    6    6    6    6    6    6    6    6    7    7
   7    8    8    9   10   10

EasySort
  14   17   18   19   20   20   20   20   20   20   20
  21   21   21   22   22   22   23   23   23   23   23
  23   24   24   24   25   25   25   26   26   26   27
  27   27   28   28   28   28   28   29   29   30   30
  30   30   30   31   31   33
```

For the HardTest data, the low score is 1 and the high value is 10. For the EasyTest data, the low value is 14 and the high value is 33. HardTest ranges over a 10 point spread while the EasyTest ranges over a 20 point spread. Not only are the HardTest numbers lower values, the variability or spread is smaller.

Frequency Distributions

The next thing that can be done is to make what is called a frequency distribution. A frequency distribution lists the scores from low to high, indicates how many frequencies there are for each score, and then totals up the overall number of frequencies in the total set. Minitab has a **TALLY** command that makes frequency distributions.

HardTest	Count	EasyTest	Count	EasyTest	Count
1	2	14	1	27	3
2	5	17	1	28	5
3	4	18	1	29	2
4	13	19	1	30	5
5	9	20	7	31	2
6	9	21	3	33	1
7	3	22	3	N=	50
8	2	23	6		
9	1	24	3		
10	2	25	3		
N=	50	26	3		

The nice thing about a frequency distribution is that it also gives us an idea of where most of the people are scoring within the range of scores. For example, the HardTest data show that the score of 4 seems to be the most popular whereas for the EasyTest data, relatively high frequencies occurred for several different score values (20, 23, etc.).

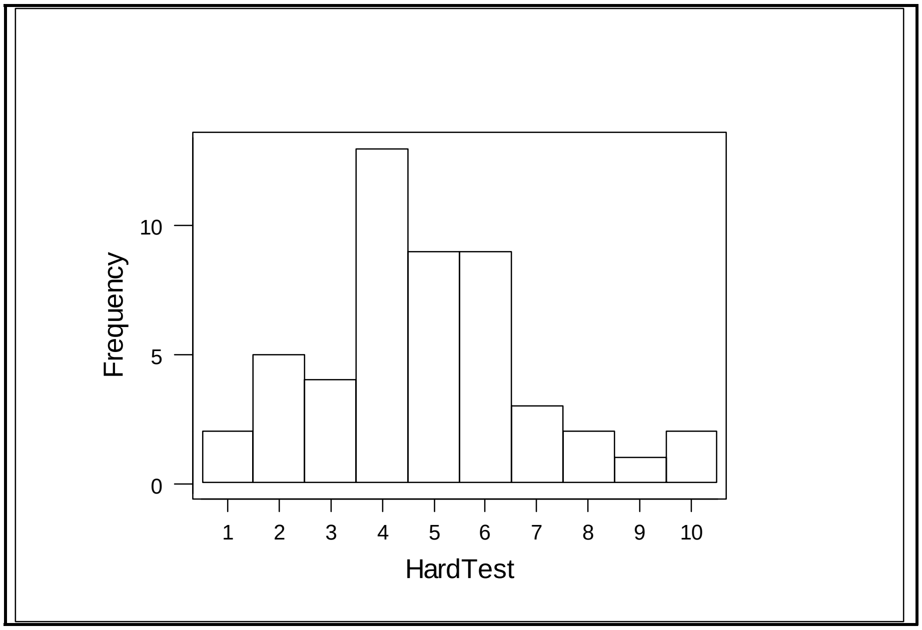
Another way to present the data that are in frequency distributions would be to use a graphical approach. Frequency polygons, referred to as **DOTPLOTS** in Minitab, and **HISTOGRAMS** are commonly used for this purpose. Look at the following dotplots for the HardTest and EasyTest variables.

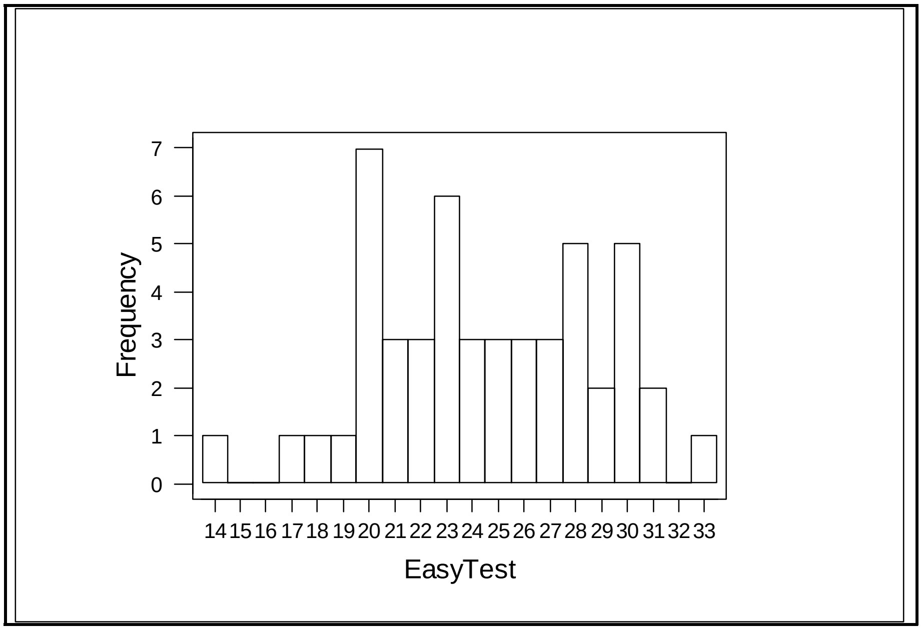
```
                  .
                  :
                  :   .   .
                  :   :   :
          .       :   :   :
          :   :   :   :   :   .
      :   :   :   :   :   :   :   :   .   :
  +---------+---------+---------+---------+-----HardTest
0.0       2.5       5.0       7.5      10.0

                  .
                  :     :         .   .
                  : . . : . . . . :   :
      .     . . . : : : : : : : : : : : :   .
   -----+---------+---------+---------+---------EasyTest
     15.0      20.0      25.0      30.0      35.0
```

Frequency polygons or dotplots place the score values along the X or horizontal axis and the height of the series of dots represents the number of frequencies. Although there is not a frequency scale listed or a formal vertical axis for the Minitab dotplot, you could count the dots to approximate the tally for each score. As was seen from the tally output above, a score of 4 in the HardTest data set seems to be the most popular but again, several scores are "popular" in the EasyTest set.

Another way to graphically present the data is with a histogram. A histogram is a vertical bar chart with the height of the vertical bars representing the frequency. The next series of graphs are histograms for the HardTest and EasyTest variables. Each are presented separately below. For the HardTest, notice that the high point or tallest bar is at 4 or the most frequently occurring score. As the scores get either higher or lower, the frequencies become less and less. For the EasyTest variable, while the highest peak point is for the score of 20, there are several other scores (23, 28, and 30) that also have high peak points. The histogram for the EasyTest data does not show as much a single peak





point as does the HardTest data. The number of dominant peak points will later be referred as modality; ie, how many peak points are there in the distribution. If there is only one, we refer to the distribution as unimodal. If there are two, we would call it bimodal. In referring to a distribution as unimodal or bimodal, it is important to keep in mind that it is meant to be descriptive of the number of dominant high points and not whether there are, technically speaking, several values with high frequencies. For example, what if in the HardTest distribution, the scores of 4, 5, and 6 all had 10 frequencies each. In that case, we would not call it trimodal since 3 values have high frequencies but rather, would call it unimodal since most of the frequencies still concentrate at one general location.

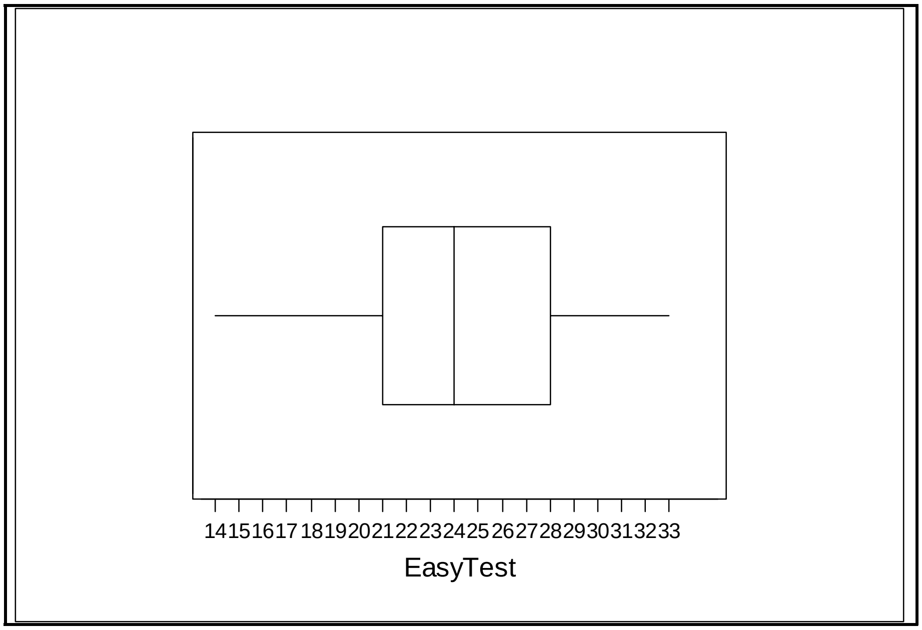
Two other graphical methods sometimes used are the **STEM AND LEAF** diagram, and the **BOXPLOT**. I have shown examples of each of these for the EasyTest variable. First shown is the stem and leaf diagram.

```
 Stem-and-leaf of EasyTest    N  = 50
 Leaf Unit = 1.0

     1    1  4
     2    1  7
     4    1  89
    14    2  0000000111
    23    2  222333333
   (6)    2  444555
    21    2  666777
    15    2  8888899
     8    3  0000011
     1    3  3
```

There are 3 parts to a stem and leaf diagram. On the right side, the first of the two columns (1,1,1,2,2, etc.) is called the "stem". The extreme right segment is called the "leaf". In this case, since the leaf part is a unit of 1, the stem part must be the next higher unit size which would be 10's. So, for the top score in the diagram (actually the lowest score), we combine the stem unit of 10 and the leaf unit of 4 to make a value of 14. If you go back to the frequency distribution for EasyTest, you will see that there was 1 frequency of 14. The next line combines a stem unit of 10 with a leaf unit of 7. Hence, we have a 17. The next line has - first - one 18 and secondly - one 19. What about the fourth line? There are seven 20's and three 21's. This is how we would continue to read the table down to the last value which is one frequency of 33. The far left column is an accumulation of the frequencies up to the line where the middle frequency will be. Since there are 50 frequencies, half of that is 25. So, starting at the bottom, the first line has 1, the second line adds 1 more, the third line adds 2 more (now we are up to 4), the fourth line adds 10 more (now we have 14), the fifth line adds 9 more (now we are up to 23), and finally by the sixth line (which adds 6 more), we would be up to the middle frequency. Minitab puts a () around the value for the line that would contain the middle frequency. You would get to the same place if you started at the lower point of the table and worked your way up. Later, that middle score value will be identified as the median.

Finally, for this review of basic organizational techniques for frequency distributions, I present the boxplot. This is not exactly a graphical version of a frequency distribution but, it is still interesting. See below for a boxplot of the EasyTest data.



Note that the score scale is listed on the baseline. The graph itself looks like what used to be called a "resistor" that has wires extending out at each end of the solid central core part. The end wires are called whiskers and represent, where the lower 25 percent (at the low end) and the upper 25 percent (at the high end) of the distribution fall. About 25% of the 50 scores in the EasyTest variable fall between 14 and 21. The upper 25% of the 50 scores fall between about 28 and 33. Notice that the upper 25% seem to stretch out over a smaller distance than the lower 25%.

The vertical intersections of the whiskers with the central core are called "hinges". The lower hinge is called Q1 or the first quartile. The upper hinge is called Q3 or the third quartile. Between the lower and upper hinges (Q1 to Q3) is the middle 50% of the 50 scores and this middle 50 percent falls between about 21 and 28. The vertical line in the middle core is called Q2 or the median (24). The median is the point where half or 50% of the distribution is below and half or 50% of the distribution is above.

What can a boxplot tell us? First of all, you can estimate the range of the scores, which in this case is about 14 to 33, or 19 points. Second, you are able to see where the middle 50 percent of the distribution lies, in this case, between about 21 and 28. Third, you are able to estimate the median value which is the middle point or about 24 in this case. Finally, if the whiskers are unequal lengths and the median is not in the middle of the central core section, then the distribution is not symmetrical. A symmetrical distribution is one that is evenly balanced on both sides of the middle. More on this point later. What you cannot tell from a boxplot is how many frequencies fall at any one score value.

Shape or Form of Frequency Distributions

It is also important to be able to verbally describe the shape or form of a frequency distribution in only a few words. In doing so, two factors are generally taken into account: symmetry and modality. Symmetrical distributions are ones that are balanced around the middle in that the pattern to the left of the middle is mirrored to the right of the middle. Patterns that are unbalanced and tend to have bunching up of the frequencies at one end or the other, are asymmetrical or skewed. The second factor is modality which represents how many dominant high points there are in the distribution. A distribution with only one peak point is called unimodal, one with two is called bimodal, etc. The examples below represent several different shapes for frequency distributions.

Diagram A

```
    .       .
    :     : :     .   .     . . : .     :
  : : . . : :   : :   :   . : : : : . . : :
  : : : : : : : : : : : : : : : : : : : : :
  : : : : : : : : : : : : : : : : : : : : :
  : : : : : : : : : : : : : : : : : : : : :
  : : : : : : : : : : : : : : : : : : : : :
  : : : : : : : : : : : : : : : : : : : : :
  +---------+---------+---------+---------+----  Rect
0.0       5.0      10.0      15.0      20.0
```

Diagram A shows an approximately equal number of frequencies at each score value and this distribution is called rectangular. The next diagram, Diagram B, shows that a greater number of the frequencies are located at the low end of the score scale with fewer and fewer frequencies as you go to the right. This is skewed to the right, or positively skewed. The trailing off end, to the + or - side of the X axis, determines the skew.

Diagram B

```
               .
        :      :
        :      :
        :      :
        :      :     :
  :     :      :     :     .
  :     :      :     :     :
  :     :      :     :     :     :      .     .
  +---------+---------+---------+---------+----  +Skew
0.0       1.6       3.2       4.8       6.4
```

The next graph, Diagram C, shows the popular distribution that is called the normal distribution. Normal distributions are symmetrical around the middle, with the majority of the frequencies being in the middle, with fewer and fewer being at the extreme sides.

Diagram C

```
                    .  .
                  . : :.::
                  : ::::::
                  :::::::::
             . :::::::::::::::
             :.:::::::::::::::..
          . .::::::::::::::::::::.
   . . ..:::::::::::::::::::::::::..::...
  +---------+---------+---------+---------+----  Normal

0.0       5.0      10.0      15.0      20.0
```

Diagram D below, represents the opposite of Diagram B, which was positively skewed, and therefore is an example of a negatively skewed distribution. In this case, most of the frequencies fall at the upper end of the distribution while fewer and fewer are obtained as you move to the left side of the baseline.

Diagram D

```
                                 .
                                 :
                                 :    :
                            .    :    :
                            :    :    :
                       .    :    :    :    :
                  .    :    :    :    :    :
   .    .    .    :    :    :    :    :    :
  -+---------+---------+---------+---------+---  -Skew
12.0      14.0      16.0      18.0      20.0
```

Diagram E

```
    .                                     :
    :     .                           .   :
    :     :                       .   :   :
    :     :   :                   :   :   :
    :     :   :               :   :   :   :
    :     :   :   .       :   :   :   :   :
    :     :   :   :   :   :   :   :   :   :
  +---------+---------+---------+---------+----  UShape
0.0       5.0      10.0      15.0      20.0
```

Finally, we come to Diagram E. Notice that the majority of the frequencies fall either at the low end of the score scale or the high end of the score scale. If you viewed this distribution as a letter, the letter that would most come to mind would be U. In fact, a symmetrical looking distribution that has peaks at the low and high ends with relatively few frequencies in the middle is referred to as a U shaped distribution.

TREND LINE DATA

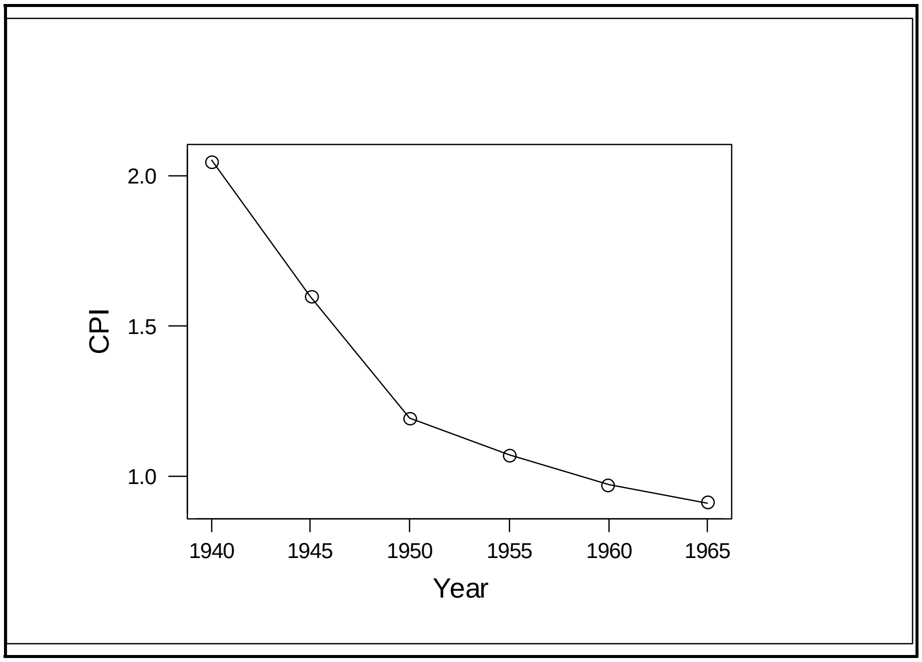
The baseline in a frequency distribution is typically quantitative in nature (test scores, heights, etc.) and the vertical axis is a frequency scale. Although the baseline is numerical, it does not necessarily reflect a time dimension such as years or trials. In many fields of work, another common relationship that is examined is the trend of data over time. For example, in business, it is commonplace to see tables and graphs where the baseline is months or years. Stock market trends use this type of a scale. On the vertical axis, you might see such variables as "number of units sold" or "cost of living". Tables and graphs that have a time dimension on the baseline and some performance indicator on the vertical axis are usually referred to as "time series" or "trend line" data. Here are two examples that fit this pattern.

What if you had data on the consumer price index (CPI) for the years of 1940 to 1965, in increments of 5 years. The data table might look as follows.

Year	CPI
1940	$2.05
1945	$1.60
1950	$1.19
1955	$1.07
1960	$ .97
1965	$ .91

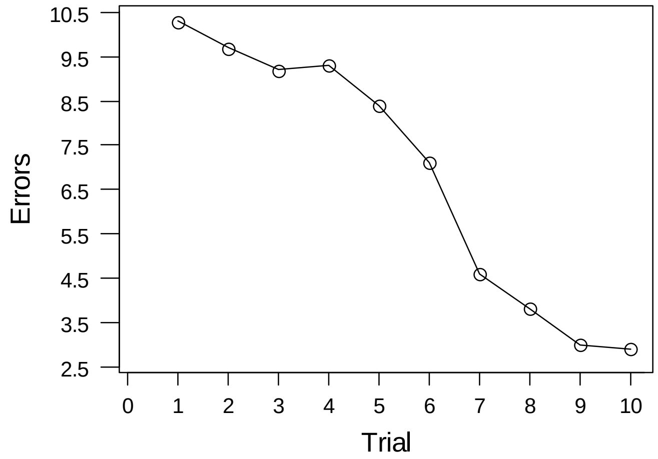
To make a graph of the data, we would put the year on the baseline and the CPI value on the vertical axis. In Minitab, there is a **PLOT** command that allows you to do this. See the first graph below.

If you look at the graph, you will see that the CPI or consumer price index declines over this time period. What this means is that the value or purchasing power of the dollar in 1960 or 1965 is not as great as it was back in 1940 or 1945.



Look at the next example, showing the relationship between the trial that one is on in a learning experiment and the number of errors that happened at that particular trial. See the second graph below. What you can clearly see is that as the experiment goes on, the number of errors that occur decrease; this is what you would expect, practice makes perfect (or almost so!).

Trial	Errors
1	10.3
2	9.7
3	9.2
4	9.3
5	8.4
6	7.1
7	4.6
8	3.8
9	3.0
10	2.9



Summary

This introductory chapter has presented several ways in which to organize data, primarily frequency distribution data but, also some time series or trend line data. Initial organization of the information is a key first step to being able to summarize the information for yourself and, also, being able to effectively communicate the information to others. Using some simple data organizational tools will in many cases answer questions you have about the data, even without using more sophisticated procedures for analyzing the information. However, this is not the end of our look at tools for analyzing data; our trip has just begun!

Useful Minitab Commands

SORT TALLY DOTPLOT HISTOGRAM STEM BOXPLOT PLOT

Practice Problems

For each of the following sets of data, do the following using Minitab or, the comparable procedures in another statistical software package.

Set 1

23	20	24	23	25	20	15	24	20	21	24
18	14	25	22	18	22	22	21	28	23	21
18	15	21	18	19	23	20	22	21	22	21
15	24	11	17	18	23	23	22	23	18	10
21	19	23	16	25	20					

Set 2

29	28	29	30	27	28	29	26	27	22	25
27	27	26	24	28	27	27	27	27	25	28
28	26	27	28	23	28	27	26	28	28	27
24	28	28	26	27	29	28	29	26	28	25
28	26	28	30	26	25	28	28	24	28	26
28	28	26	27	26						

1. Enter the data in Minitab and give names to each variable.
2. Sort the data and print out the sorted values.
3. Make frequency distributions using the **TALLY** command.
4. Make **DOTPLOTS** and **HISTOGRAMS**.
5. Verbally describe each distribution.
6. Make a **STEM AND LEAF** diagram of the first set of data and a **BOXPLOT** of the second set of data. Give some description of each graph.

Make trend line graphs out of the following sets of data.

Set 1

Trial	#Correct
1	6.2
2	7.1
3	6.8
4	10.4
5	12.6
6	13.8
7	21.7
8	22.0
9	28.0

Set 2

Year	Consumption
1960	283
1961	294
1962	304
1963	346
1964	384
1965	333
1966	300
1967	296
1968	264
1969	260
1970	250

7. What happens to the number of correct responses as trials increase?
8. What happens to consumption as the years increase from 1960 to 1970?