Don't Just Teach Exploratory Techniques: Use Them!

Robert W. Hayden
Plymouth State College
Plymouth, NH 03264
hayden@mail.plymouth.edu

Prepared for the

Beyond the Formula Statistics Conference

Rochester, NY
2 August 2001

The tools for data analysis that John Tukey introduced in the 1960s have now appeared in many reform high school mathematics curricula and in many college statistics textbooks. I hope that means that they appear in many statistics courses as well. Even if they do, I wonder how they are used. In some of the statistics textbooks I have seen (Hayden, 2000), they are introduced in the first half of the book, and never used in the second half of the book. There is no point in teaching these techniques unless we show students when and how to use them. The best way to do this is by setting a good example.

The principal places to use data analytic tools are

1. to explore data and
2. to check inference assumptions.

Item 2 implies that data analytic techniques should appear throughout the second half of an introductory statistics course. Yet these techniques often disappear as soon as they are "covered". (In this case, "covered" seems to mean "buried".)

If you consider an inference topic such as comparing the means of two populations, you can see why someone might not make data displays. Whether you use the traditional procedure with a "pooled variance estimate" that is a bear to calculate, or the more appropriate procedure with fractional degrees of freedom, also a bear to calculate, you are in for a lot of arithmetic. Do we want also to insist that the students make displays of the data as well? Indeed, there is a temptation to "help" them by providing summary statistics *instead* of data. This eases the arithmetic burden, but now they *can't* look at the data. The solution lies in two corollaries to actually using data analysis tools:

Corollary 1 We have to have data in order to look at the data. You cannot evaluate the assumptions underlying inference without looking at the data.

Corollary 2 We need to use technology to ease the computational and artistic burden on the students. Their focus should be on the data, not on the arithmetic.

What I would like to do in this paper is illustrate by example how I think data analytical tools can and should be used throughout an introductory statistics course. I'll try to follow Corollary 2 and use statistical software, Versions 8 and 11 of Minitab, to illustrate my points. I will start with some examples of categorical data, partly because textbooks almost never present raw categorical data, and partly because my students keep telling me you can't have outliers in categorical data. My first example is modeled after a real consulting experience. I had been asked to review and comment on a puzzling output from a statistical software program. I usually consult for free, so that I can easily say "no" and so I can set the rules. One of them is that I will not look at a computer printout unless accompanied by the raw data. My client grumbled but complied. I did not keep this client's data, but I reproduce a similar set of numbers below. The subjects were athletes, male and female.

```
MTB > print c1

Sex
     1     0     1     1     1     1     1     1     0     2     0     1
     0     0     1     1     1     0     1     1     1     1     1     1
     0     2     0     0     1     1     1     0     1     1     1     1
     1     0     1     1     1     1     1     0     1     0     1     1
     1     1     0     0     1     0     0     0     0     1     1     0
     1     0     1     0     1     0     1     1     1     0     1     0
     0     1     1     0     0     1     1     1     1     0     1     1
     0     0     1     1     0     0     1     0     1     1     1     1
     0     0     1     1     1     1     1     1     1     1     0     0
     1     1     1     0     0     1     0     1     1     0     1     0
     1     1     1     1     1     1     1     1     1     0     1     0
     0     1     1     0     0     1     0     1     1     1     1     1
     1     1     1     0     1     1     1     2     1     1     0     1
     0     0     1     1     0     0     1     1     0     0     0     1
     0     1     1     1     1     1     1     0     1     0     0     0
     1     0     1     0     0     1     0     1     1     0     1     0
     1     0     0     0     0     0     1     1
```

Do you see any problems with this data? I tried to re-create it so a problem could be spotted from the data, but perhaps not instantly by everyone. A simple tally is more revealing.

```
MTB > tally c1

     Sex  Count
       0     74
       1    123
       2      3
      N=    200

Saving worksheet in file: I:\MINITAB\ELEVEN\DATA\3SEXES.MTW
```

At this point there was no need to look at the client's computer printout. I simply returned the entire package with the query, "Which three sexes were studied?"

I will take a brief digression here to elaborate on this example because the actual problem was much more serious than the tally above might suggest. One might suppose that perhaps three 2s were accidentally entered. In that case all we would need to do is replace them with zeros or ones, or missing value symbols if we do not know the actual sex of the subject. What actually happened to my client was much worse, and is related to the fact that the data were collected long ago and stored on punched cards. Readers my age or older can probably imagine what went wrong. For the younger of my readers, I'll present a simplified version with just 30 athletes. Suppose the data were stored like this:

Group 1	Sex	1 0 0 1 0 0 0 1 1 0 1 0 1 1 0
	Teams	1 2 0 0 1 1 0 0 2 0 1 0 2 0 1
Group 2	Sex	0 0 1 0 1 0 1 1 0 1 0 0 0 1 0
	Teams	1 0 0 1 1 2 1 1 1 0 1 1 0 1 0

Here I have added a hypothetical variable: the number of teams an athlete played on. The computer read the data as if the numbers for the Sex variable were in Rows 1 and 2 rather than 1 and 3. As a result the data on Sex and Teams was all mixed together and summary statistics for both variables were wrong. To put it another way, it was not just the three 2s that were errors, but half the data analyzed. 50% outliers! And, this affected not just these variables, but every variable in the study! My curt note was just the prod my client needed to look more closely at the data and discover the problem. When it was corrected and run through the computer again, no special help was needed with the output, which now seemed quite sensible.

One more study saved by ***looking at the data***!

Here is another example, this one based on a common typing error I have seen in my students and occasionally made myself. Would you (or your students) see any problem with the following hypothesis test comparing two proportions? (What if no one had aroused their suspicions?)

```
MTB > twosample t on c21 and c22

           N      Mean     StDev   SE Mean
males     25     0.360     0.490     0.098
females   24     0.75      2.03      0.41

95% CI for mu C21 - mu C22: ( -1.266,  0.49)
T-Test mu C21 = mu C22 (vs not =): T= -0.92  P=0.37  DF=  25
```

We are comparing proportions of male and female faculty favoring a particular candidate for Dean of the College. It seems a bit odd that there would be as large a difference as between 36% and 75%, and odder still that such a large difference is not significant. Yet odder than any of these things is the standard deviation of 2.03 for the data on the women, and a confidence interval for a difference in proportions that says a 126.6% difference would be compatible with the data. Such data are usually coded in the computer as zeros and ones, in this case, “1=favor” and “0=does not favor”. The mean of such data must be between 0 and 1, so when you calculate

the standard deviation, the residuals will all be between 0 and 1. Since the standard deviation is a typical value for a residual, it too must be between 0 and 1.

Since few textbooks show you raw categorical data, here it is for the males

```
Retrieving worksheet from file: D:\WP\SPRINT\MYPAPERS\BEYOND\TEN.MTW

MTB > print c21

C21
   0   1   0   1   0   0   0   0   0   1   1   0   1   0   0   0   1   0
   0   1   1   0   1   0   0
```

and for the females.

```
MTB > print c22

C22
    0    1    0    1    0    0    0    0    0    1   10    1    0    0    0
    1    0    0    1    1    0    1    0    0
```

In the unlikely event that you missed the problem, here are two data displays that should make it obvious.

```
MTB > histogram c21

Histogram of C21   N = 25

Midpoint   Count
       0      16  ****************
       1       9  *********

MTB > histogram c22

Histogram of C22   N = 24

Midpoint   Count
       0      15  ***************
       1       8  ********
       2       0
       3       0
       4       0
       5       0
       6       0
       7       0
       8       0
       9       0
      10       1  *
```

If you go back and look at the data, you can see that the only difference between C21 (Minitab-speak for the data in Column 21) and C22 is that in C22 the space was left out between a 1 and a 0, thereby creating a 10. This greatly inflated both the proportion favoring this candidate and the standard deviation.

These first two examples illustrate that you can indeed have outliers in categorical data and that they can lead to gross errors. The best way to avoid them is to

look at the data.

Next we move on to measurement data. I recently reviewed a textbook manuscript that defined a normal distribution as any distribution that was perfectly symmetric and mound-shaped. Does the following data satisfy that definition? It's the results of a chemistry experiment conducted by two classes. The answer they were supposed to get was 0.110. (This data comes with most versions of Minitab.)

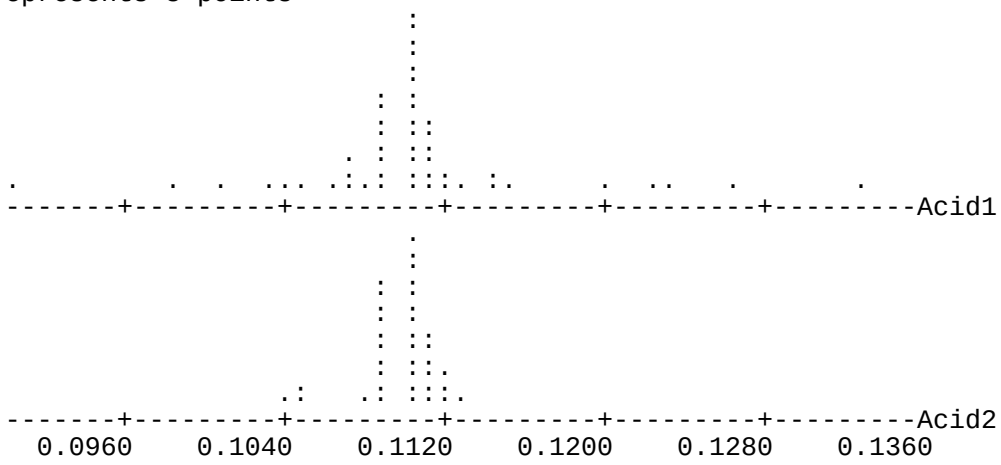
```
MTB > Retrieve  'I:\MINITAB\ELEVEN\DATA\ACID.MTW'.
Retrieving worksheet from file: I:\MINITAB\ELEVEN\DATA\ACID.MTW

MTB > stem c1 c2

Stem-and-leaf of Acid1     N  = 124
Leaf Unit = 0.0010


    1    9 0
    2    9 8
    8   10 113444
   45   10 567777777888999999999999999999999
  (72)  11 000000000000000000000000000000000000000000011111111111111111122222+
    7   11 5
    6   12 0233
    2   12 6
    1   13 3


Stem-and-leaf of Acid2     N  = 37
Leaf Unit = 0.00010


    1  104 0
    3  105 00
    3  106
    3  107
    4  108 0
   14  109 0000000000
  (13) 110 0000000000000
   10  111 000000
    4  112 000
    1  113 0
```

The “+” at the far right of the stem and leaf plot for Acid1 means it ran off the edge of the paper. Here is a histogram.

```
MTB > histogram c1

Histogram of Acid1   N = 124
Each * represents 2 observations

Midpoint   Count
   0.090       1  *
   0.095       0
   0.100       3  **
   0.105      14  *******
   0.110      92  **********************************************
   0.115       8  ****
   0.120       2  *
   0.125       3  **
   0.130       0
   0.135       1  *
```

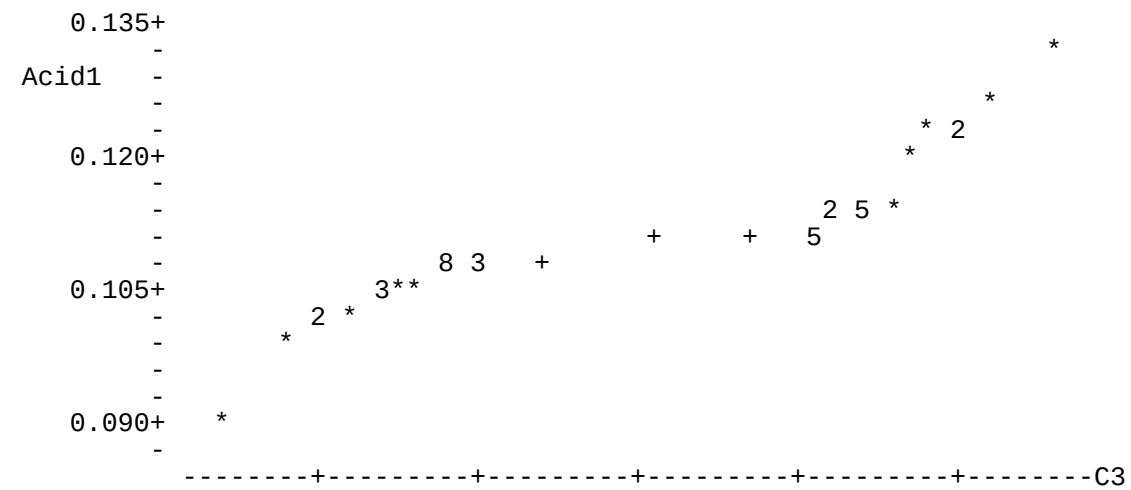
Acid1 looks more symmetric than Acid2, but perhaps the fact that the plots are on different scales is misleading us.

```
MTB > dotplot c1 c2;
SUBC> same.

Each dot represents 3 points
```

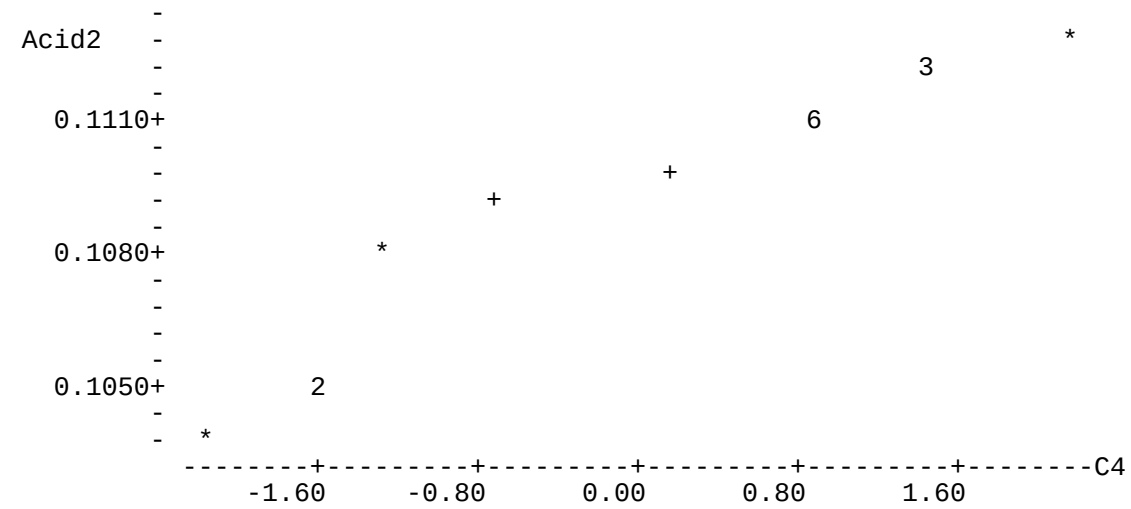


Now they both look fairly mound-shaped, but Acid1 has much longer tails. For this we need a normal probability plot, not a new (i.e., it’s more than 50 years old) technique, but the one that best makes heavy tails stand out.

```
MTB > nscores c1 in c3
MTB > nscores c2 in c4
MTB > plot c1 c3
```



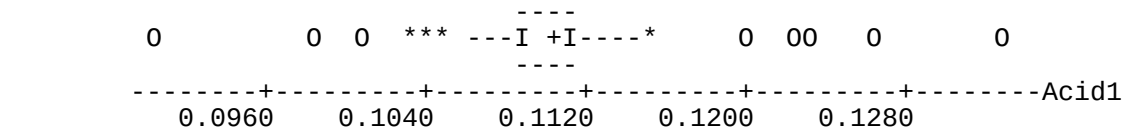

```
MTB > plot c2 c4
```



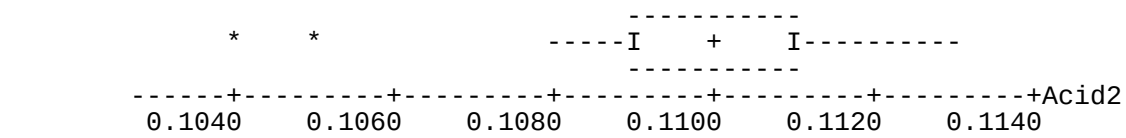
If you are not familiar with these plots, my advice on interpretation is to draw or imagine a straight line through the middle third of the data. If most of the data are close to that line, the data are close to normal. If the data show an S-shape, there are problems in the tails. Acid1 has heavy tails; Acid2 probably has three low outliers.

If you don't want to use normal probability plots, boxplots can also do the job, if you know how to interpret them.

```
MTB > boxplot c1
```



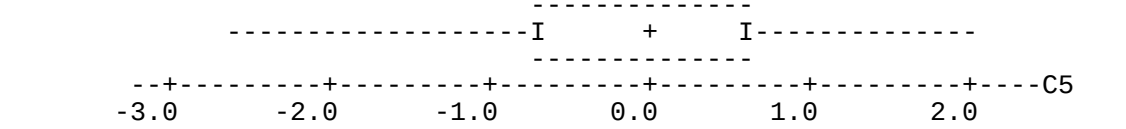

```
MTB > boxplot c2
```



The plot of Acid2 is unexceptional except for the moderate low outliers. The plot of Acid1, on the other hand, looks like it is mostly outliers. They appear on both sides and some appear extreme (marked as O rather than *). They all represent a problem with heavy tails rather than true outliers. (Similarly, a boxplot of a highly skewed distribution will often show "outliers" in the direction of the skew. These represent the skewness rather than an outlier problem.)

Incidentally, another reason for using data displays throughout an introductory statistics course is that this gives students lots of practice in interpreting such displays. As the course proceeds, they learn of more things to look for, and get better at finding old things. Technology can help them to get a lot of practice in a very short time. For example, to guide myself in evaluating the boxplot of Acid1, I generated six samples of the same size from a normal distribution and made boxplots of these.

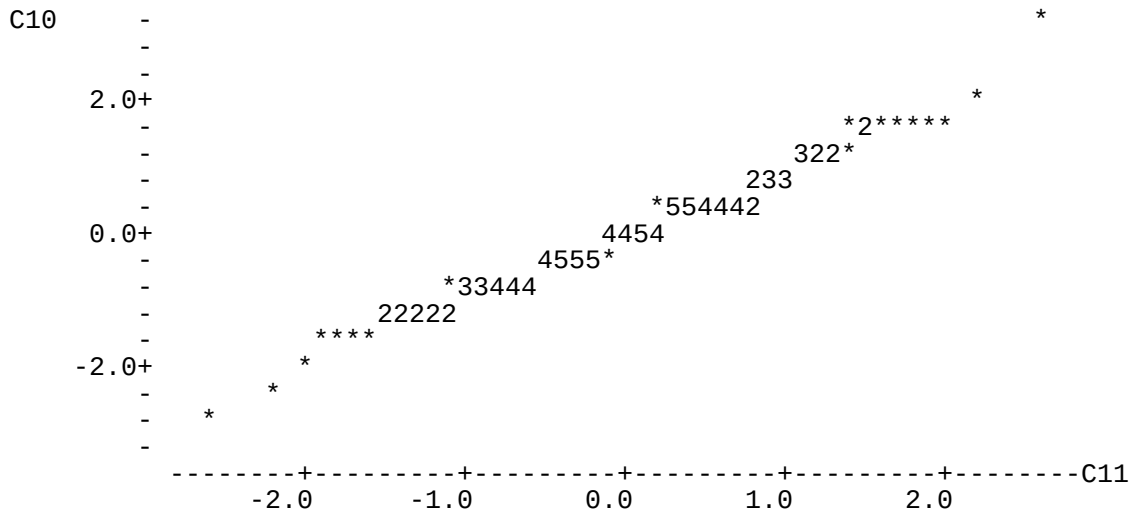
```
MTB > random 124 in c5-c10
MTB > boxplot c5
```



```
MTB > boxplot c6

                                  ------------
             *   -----------------I     +    I-----------------
                                  ------------
          +---------+---------+---------+---------+---------+------C6
        -3.0      -2.0      -1.0       0.0       1.0       2.0
MTB > boxplot c7

                                 --------------
               -----------------I      +     I----------------
                                 --------------
          --------+---------+---------+---------+---------+--------C7
               -2.0      -1.0       0.0       1.0       2.0

MTB > boxplot c8

                               -----------
               ---------------I    +    I-------------        *
                               -----------
          ----+---------+---------+---------+---------+---------+--C8
           -2.4      -1.2       0.0       1.2       2.4       3.6

MTB > boxplot c9

                                 -------------
                 ---------------I     +     I-----------------
                                 -------------
          --------+---------+---------+---------+---------+--------C9
               -2.0      -1.0       0.0       1.0       2.0

MTB > boxplot c10

                               ------------
               *  -------------I    +     I-----------        *
                               ------------
          ------+---------+---------+---------+---------+---------+C10
             -2.4      -1.2       0.0       1.2       2.4       3.6
```

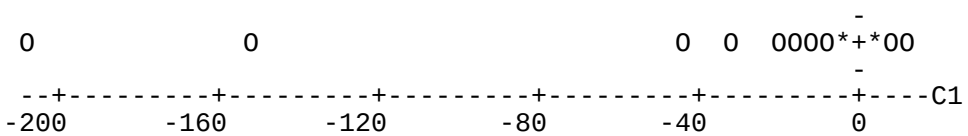
They average about one modest outlier per sample, a far cry from what we saw for Acid1, which is approximately symmetric and mound-shaped, but nowhere near normal. I'll just give one histogram of 124 observations sampled from a true normal distribution just to show that a normal distribution is much more "triangular," and much less "pointy", than Acid1, and one normal probability plot of a similar sample, to show what that might look like for such a sample.

```
MTB > histogram c5

Histogram of C5   N = 124

Midpoint   Count
    -2.5       1  *
    -2.0       3  ***
    -1.5       8  ********
    -1.0      19  *******************
    -0.5      17  *****************
     0.0      26  **************************
     0.5      25  *************************
     1.0      16  ****************
     1.5       6  ******
     2.0       3  ***
     2.5       0
     3.0       0
     3.5       0

MTB > nscores c10 in c11
MTB > plot c10 c11
```

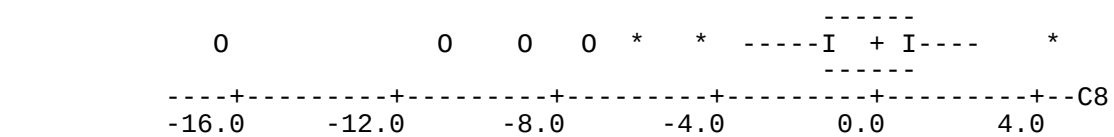


For comparison, here are boxplots of *t*-distributions with one (Cauchy), two and three degrees of freedom.

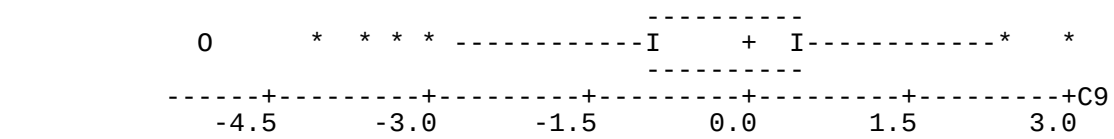
```
MTB > random 124 in c1;
SUBC> t 1 df.
MTB > boxplot c1
```



```
MTB > random 124 in c8;
SUBC> t 2.
MTB > boxplot c8
```




```
MTB > random 124 in c9;
SUBC> t 3.
MTB > boxplot c9
```




```
Saving worksheet in file: I:\MINITAB\ELEVEN\DATA\CAUCHY.MTW
```

For the past decade, I have used introductory statistics textbooks by Siegel (1998) or by Siegel and Morgan (1996). These are outstanding in their use of data displays throughout the book, and for their choice of datasets to illustrate important issues. One data set from pages 50-52 of Siegel concerns golf prize monies. Let's start with a stem and leaf plot.

```
Retrieving worksheet from file: E:\STATS\MINITAB8\STATS1A\MISC\GOLF79.MTW

 MTB > stem c1

 Stem-and-leaf of Prize-k$  N  = 65
 Leaf Unit = 1.0

     3    1 112
    23    1 55555555555555569
    29    2 222222
    29    2
    30    3 3
    (3)   3 667
    32    4
    32    4 55555555559
    21    5 004444444444444
     6    5
     6    6 0033
     2    6
     2    7 22
```

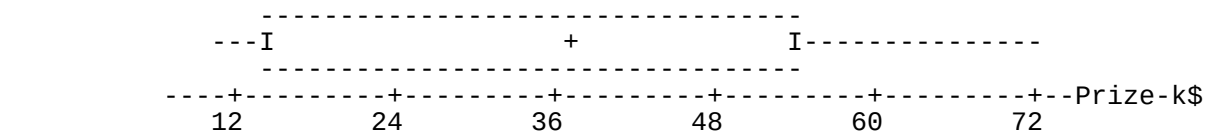
The distribution has two peaks. Why? If you check the *World Almanac* you will find there are tournaments for men and tournaments for women.

```
Men
      3 3
      3 66
      4
      4 55555555559
      5 004444444444444
      5
      6 0033
      6
      7 22

Women
      1 112
      1 55555555555555555569
      2 222222
      2
      3
      3 7
```

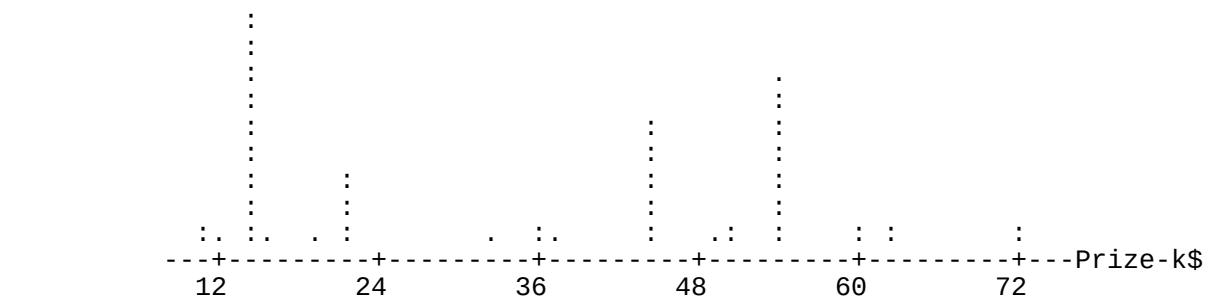
Here we have measurement data on one variable, prize winnings, but there is data for two different groups, men and women. Thus, we actually have a two variable problem. The prize monies are a measurement variable and the sex of the golfer is a categorical variable. In the case of the original display showing both sexes lumped together, sex is a lurking variable that accounts for the bimodality of the distribution. To keep the sexes straight, I coded 0=male and 1=female in C2. Then the second time I used the **stem** command I also used the **by** subcommand. This gives a separate stem and leaf diagram for each value in the column specified after the **by** subcommand. Note here that Minitab labeled these displays "Sex = 0" and "Sex = 1". Can you guess which is men and which is women?

In using data analytic tools, it is important to know their strengths and weaknesses. For example, the boxplot below of the pooled golf prize data does not show the bimodality evident in both the initial stem and leaf plot and in the dotplot (on the same scale) shown directly beneath the boxplot for comparison.

```
MTB > boxplot c1
```

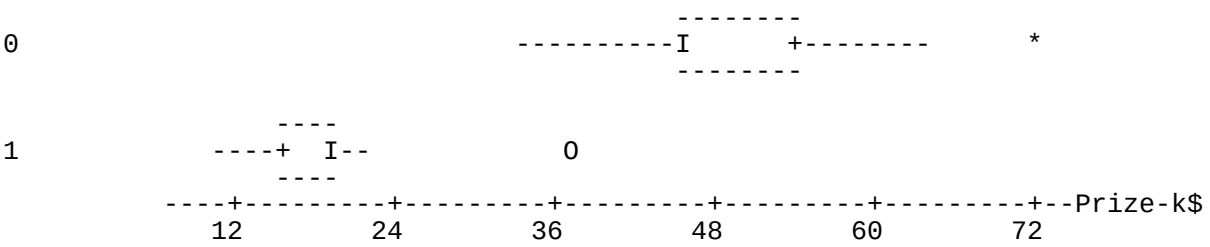



```
MTB > dotplot c1
```



And, although boxplots are noted for detecting outliers, pooling the data here also causes the boxplot to miss the outliers, especially the big one among the data for women's tournaments. They do show up if you separate the men from the women.

```
MTB > boxplot c1;
SUBC> by c2.

Sex
```



To see bimodality or other details of shape, we can use a dotplot, stem and leaf plot, or a traditional histogram. The dotplot (sometimes called a "line plot" or even "number line plot" in K-12 materials) is a simple display that can be (and is) taught in the elementary grades. The stem and leaf is a bit more sophisticated. The fact that we can extract individual observations from a stem and leaf is one of its advantages over a conventional histogram. The golf data was extracted from Siegel in just this way. The traditional histogram has an advantage for very large data sets, where marking every point would be too distracting or run off the edge of the

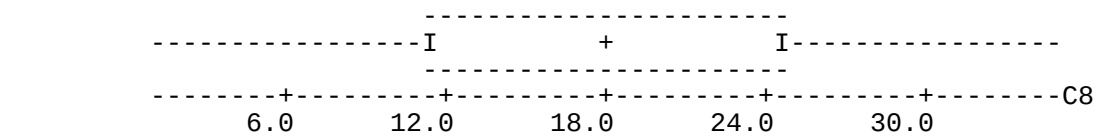
page. It is also more familiar to an audience not trained (or not trained recently) in statistics.

Here is another example that shows features of several displays. First a histogram.

```
MTB > histogram c8

Histogram of C8   N = 201
Each * represents 2 observations

Midpoint   Count
       0       1  *
       4       2  *
       8      17  *********
      12      80  ****************************************
      16       0
      20       1  *
      24      80  ****************************************
      28      17  *********
      32       2  *
      36       1  *
```

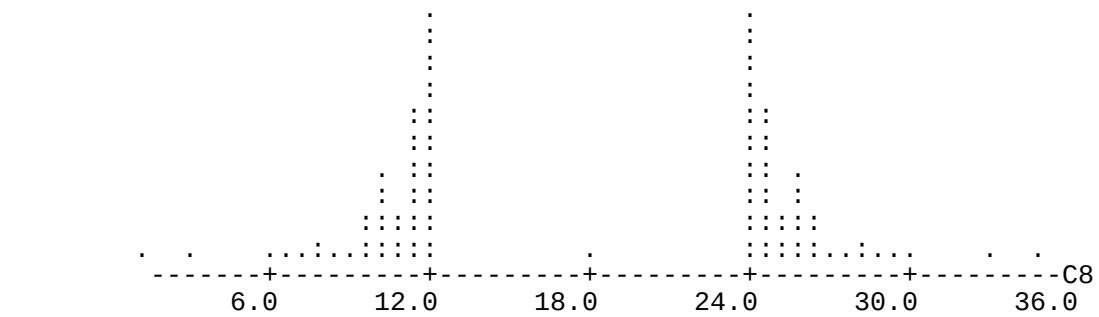
This looks nearly symmetric and bimodal. As expected, a boxplot shows the first but not the second of these attributes.

```
MTB > boxplot c8
```



My favorite is a dotplot.

```
MTB > dotplot c8

Each dot represents 2 points
```



I would say we have two groups and an outlier right in the middle that does not appear to belong to either group. Because of situations like this, I prefer to define an "outlier" as an observation that does not fit into the pattern of the rest of the data, rather than an observation that is high or low.

The histogram failed to detect the outlier due to an unfortunate choice of class boundaries on the part of Minitab. This illustrates this danger. We can override Minitab's default choices and make the outlier reappear.

```
MTB > histogram c8;
SUBC> increment 2;
SUBC> start at 0.

Histogram of C8   N = 201
Each * represents 2 observations

Midpoint   Count
    0.00       0
    2.00       1  *
    4.00       1  *
    6.00       3  **
    8.00       6  ***
   10.00      26  *************
   12.00      63  *******************************
   14.00       0
   16.00       0
   18.00       1  *
   20.00       0
   22.00       0
   24.00      63  *******************************
   26.00      26  *************
   28.00       6  ***
   30.00       3  **
   32.00       1  *
   34.00       1  *

Saving worksheet in file: I:\MINITAB\ELEVEN\DATA\TWOPEAKS.MTW
```

While we have a variety of displays for measurement data for a single group, multiple groups restrict our options considerably. Back-to-back stem and leaf plots can help for only two groups. For more we want a boxplot for each group, all on the same scale. Indeed, such comparisons are what boxplots are for. If you only have one group, a more detailed display is usually preferable.

Here are populations for each county for the New England states and New York. (As this is being written, Kentucky Fried Chicken is claiming New York *is* a New England state, and since I am giving this paper in the state of New York, I did not want to leave them out.) The states are coded as

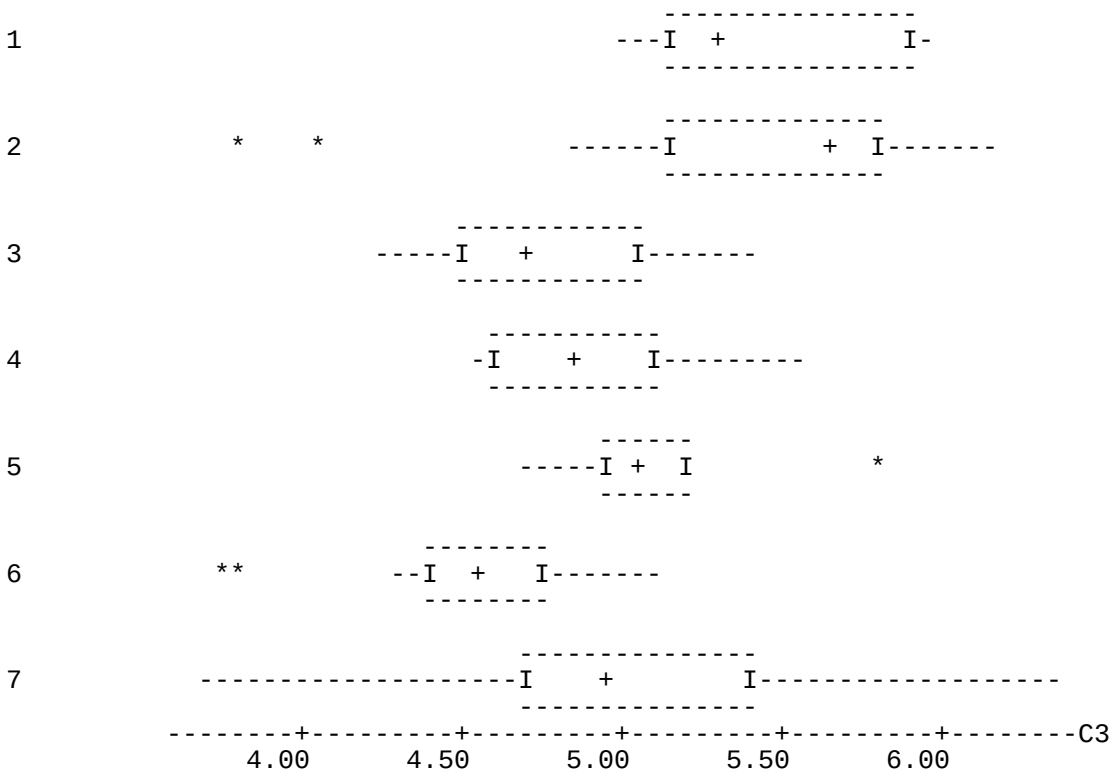
1=CT
2=MA
3=ME
4=RI
5=NH
6=VT
7=NY

```
MTB > boxplot c2;
SUBC> by c1.

State

              --------------
1            -I+            I-
              --------------

              -----------
2          ---I     +   I----------------
              -----------

            --
3          -+I- *
            --

            --
4           +I  * *
            --

             --
5           -+I        0
              --

           --
6          I+ *
           --

            -----
7          -I+  I----    *  0 0    0 0   0        0      0
            -----
           +---------+---------+---------+---------+---------+------Pop.
           0    500000   1000000   1500000   2000000   2500000
```

The usual inference technique for comparing means across several groups is Analysis of Variance. However, it requires that the variabilities be about the same from group to group, and that the data in each group be more or less normally distributed. We seem to have differing variabilities and data that is often skewed toward high values. A transformation is one way to deal with these problems. (If we ask students to check assumptions, we should give them some clues as to what

to do if assumptions are violated. Siegel (1988)has the best elementary treatment of transformations, Siegel and Morgan (1996) the second best.)

```
MTB > let c3=logten(c2)
MTB > boxplot c3;
SUBC> by c1.
```




```
Worksheet saved into file: d:\wp\sprint\mypapers\neandny.MTW
```

While not perfect, the logarithms are much closer to meeting the ANOVA assumptions than the original data were.

Let me close with another example from Siegel and Morgan. It's such a masterpiece that it makes a fitting finale. The data are on the average heights of girls from pages 554-557, and illustrate why we want to calculate and plot residuals.

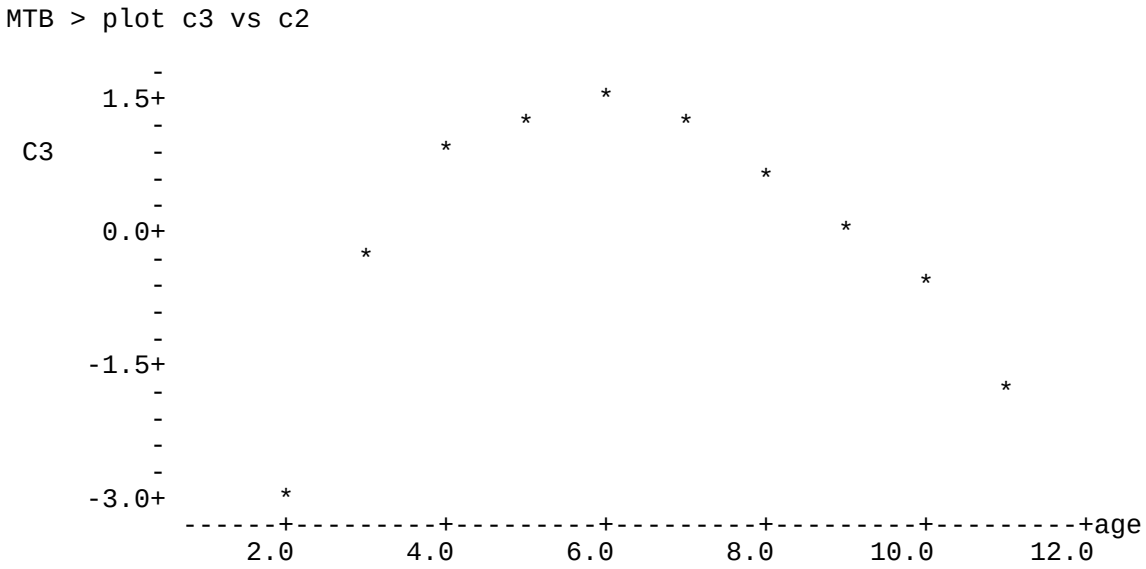
```
MTB > info

COLUMN    NAME      COUNT
C1        height       10
C2        age          10
```

```
MTB > plot 'height' versus 'age'

          -
          -                                                    *
          -
      140+                                                *
          -
 height   -                                           *
          -                                      *
          -                                 *
      120+
          -                            *
          -
          -                       *
          -                  *
      100+
          -             *
          -
          -       *
          -
            ------+---------+---------+---------+---------+---------+---------+age
                2.0       4.0       6.0       8.0      10.0      12.0
MTB > correlation c1 c2

Correlation of height and age = 0.997
```

Both the scatter plot and the correlation suggest the data are very close to a straight line.

```
MTB > regress height in c1 vs. 1 ind. var. age in c2;
SUBC>residuals in c3.

The regression equation is
height = 76.6 + 6.37 age

Predictor       Coef       Stdev    t-ratio        p
Constant      76.641       1.188      64.52    0.000
age           6.3661      0.1672      38.08    0.000

s = 1.518       R-sq = 99.5%     R-sq(adj) = 99.4%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      3343.5      3343.5   1450.45    0.000
Error         8        18.4         2.3
Total         9      3361.9

Obs.     age    height       Fit Stdev.Fit  Residual   St.Resid
  1      2.0    86.500    89.373     0.892    -2.873     -2.34R

R denotes an obs. with a large st. resid.
```

Looks about as linear as you can get, until you see the residuals.



The residual plot shows some curvature that was hidden in the original scatter plot. In general, residual plots magnify any lack of fit between the data and the model. An ideal residual plot shows no patterns at all – only random scatter. In this case we could get an even better fit with some sort of curve. (Once again, if assumptions are not met, it is good to give students a clue as to what might be done about it, even if we have to tell them they will need to take another statistics course if they want the details.) Of course, you may wish to try a quadratic as homework. If you do, don't forget to plot the residuals from ***that*** model. I promise they will give you something to think about!

References

Hayden, Robert W., "Advice to Mathematics Teachers on Evaluating Introductory Statistics Textbooks" in *Resources for Undergraduate Instructors Teaching Statistics*, Thomas L. Moore, ed., 2000, Mathematical Association of America, Washington, DC.

Siegel, Andrew F., *Statistics and Data Analysis: An Introduction*, 1988, John Wiley and Sons, New York.

Siegel, Andrew F., and Charles J. Morgan, *Statistics and Data Analysis: An Introduction*, 2nd. edition, 1996, John Wiley and Sons, New York.