Transforming Univariate Data

Many statistical procedures are appropriate only if certain assumptions are met. One technique that is often used to remedy problems with assumptions in to transform the data. At a more modest level, we may need to transform the data before we can even make a useful display. That is illustrated here with data from *The World Almanac* on the areas of major islands in the Atlantic Ocean. Because there are 27 data points, this would be a ***lot*** of work to transform this data by hand! We will use Minitab to help us. As always, the first step is to ***look at the data***!

```
MTB > print c1

A
    3066       34      902     5380       20      785       10     2808
    3981     1750      540     4700        7   840000    39769     1396
     307    15528       91      108       46    42030     2184       47
    1450    18800       40
MTB > histogram c1

Histogram of A   N = 27

Midpoint   Count
       0      26  **************************
  100000       0
  200000       0
  300000       0
  400000       0
  500000       0
  600000       0
  700000       0
  800000       1  *
```

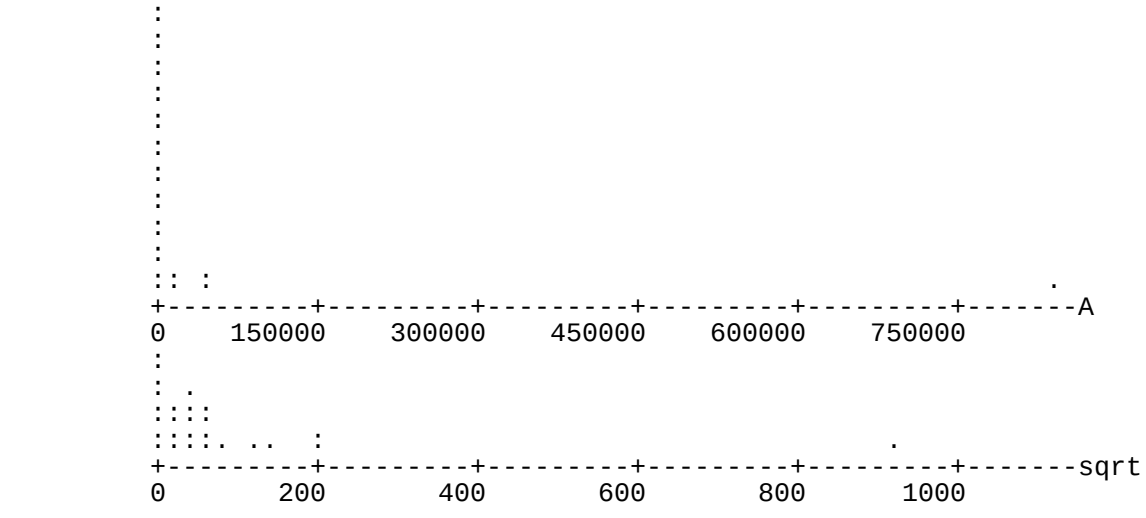
This does not look good! At first we might think that the extreme outlier in our data is an error, but if you look at a map you will see that Greenland really is a very large island! We will use Minitab to carry out some common transformations. Between the commands themselves and the names given to the columns, you should be able to follow this.

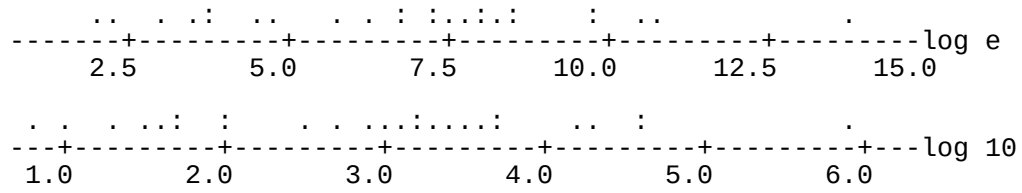
```
MTB > let c2 = sqrt(c1)
MTB > let c3 = loge(c1)
MTB > let c4 = logten(c1)
MTB > name c2 'sqrt' c3 'log e' c4 'log 10'

MTB > print c1-c4

 ROW        A      sqrt     log e    log 10

   1     3066    55.371    8.0281   3.48657
```

```
 2       34     5.831    3.5264   1.53148
 3      902    30.033    6.8046   2.95521
 4     5380    73.348    8.5904   3.73078
 5       20     4.472    2.9957   1.30103
 6      785    28.018    6.6657   2.89487
 7       10     3.162    2.3026   1.00000
 8     2808    52.991    7.9402   3.44840
 9     3981    63.095    8.2893   3.59999
10     1750    41.833    7.4674   3.24304
11      540    23.238    6.2916   2.73239
12     4700    68.557    8.4553   3.67210
13        7     2.646    1.9459   0.84510
14   840000   916.515   13.6412   5.92428
15    39769   199.422   10.5908   4.59955
16     1396    37.363    7.2414   3.14489
17      307    17.521    5.7268   2.48714
18    15528   124.611    9.6504   4.19112
19       91     9.539    4.5109   1.95904
20      108    10.392    4.6821   2.03342
21       46     6.782    3.8286   1.66276
22    42030   205.012   10.6461   4.62356
23     2184    46.733    7.6889   3.33925
24       47     6.856    3.8501   1.67210
25     1450    38.079    7.2793   3.16137
26    18800   137.113    9.8416   4.27416
27       40     6.325    3.6889   1.60206
```

Whenever you have Minitab fill columns up with numbers, it is a good idea to look and see if the results are what you intended. If you are using a Mac or Windows, you can see the results in the data window. (You may need to scroll around.) Otherwise, you can use the **print** command. Even with the Mac and Windows, **print** serves the useful purpose of putting the results into the session window where they can become part of a record of your work. Another type of check is to calculate one row of the table by hand, which takes some work, but a lot less than doing all 27 rows by hand! Once we are sure we have the right numbers, we should make some sort of display to see the effects of our transformations.

```
MTB > dotplot c1-c4
```





Remember that the purposes for transformations are to

1. get the data spread out in a dispaly so we can see individual items, rather than having them all clumped up at one end,

2. get the data to look more symmetric, and

3. get the data distribution to look more like a normal distribution.

The dotplots show extreme clumping and skewness for A, somewhat less for the square root of A, and even less for the logarithms of A. We probably have too much detail in this summary. Let's try a histogram or stem and leaf. (Note that now we are evaluating the type of display to use rather than the transformation. We want to make sure we have a good display of the transformed data before we draw any conclusions about the value of the transformations.)

```
MTB > histogram c1-c4

Histogram of A   N = 27

Midpoint   Count
       0      26  **************************
  100000       0
  200000       0
  300000       0
  400000       0
  500000       0
  600000       0
  700000       0
  800000       1  *

Histogram of sqrt   N = 27

Midpoint   Count
       0      17  *****************
     100       7  *******
     200       2  **
     300       0
     400       0
     500       0
     600       0
     700       0
     800       0
     900       1  *
```

```
Histogram of log e   N = 27

Midpoint   Count
       2       2  **
       3       1  *
       4       4  ****
       5       2  **
       6       2  **
       7       5  *****
       8       5  *****
       9       1  *
      10       2  **
      11       2  **
      12       0
      13       0
      14       1  *

Histogram of log 10   N = 27

Midpoint   Count
     1.0       2  **
     1.5       5  *****
     2.0       2  **
     2.5       2  **
     3.0       5  *****
     3.5       6  ******
     4.0       1  *
     4.5       3  ***
     5.0       0
     5.5       0
     6.0       1  *

MTB > stem c1-c4

Stem-and-leaf of A         N  = 27
Leaf Unit = 10000

  (26)   0 00000000000000000000001134
    1    1
    1    2
    1    3
    1    4
    1    5
    1    6
    1    7
    1    8 4

Stem-and-leaf of sqrt      N  = 27
Leaf Unit = 10

  (22)   0 00000001122333445667
    5    1 239
    2    2 0
    1    3
    1    4
    1    5
    1    6
    1    7
    1    8
    1    9 1
```

```
Stem-and-leaf of log e     N  = 27
Leaf Unit = 0.10

    1    1 9
    3    2 39
    7    3 5688
    9    4 56
   10    5 7
   13    6 268
   (5)   7 22469
    9    8 0245
    5    9 68
    3   10 56
    1   11
    1   12
    1   13 6
Stem-and-leaf of log 10    N  = 27
Leaf Unit = 0.10

    1    0 8
    3    1 03
    8    1 56669
   10    2 04
   13    2 789
   (6)   3 112344
    8    3 567
    5    4 12
    3    4 56
    1    5
    1    5 9
```

Here, the plots of the logs do not look as ragged. In fact, they look pretty good!

Boxplots provide an even briefer summary. Typically, boxplots are used to compare the centers and variabilities of several groups of numbers. In that situation, we put all the boxplots on the same scale. Here we are using boxplots to look at the shape of the original data and the results of each transformation, so we make a separate boxplot for each.

```
MTB > boxplot c1

          -
          +O O                                                    O
          -
          +---------+---------+---------+---------+---------+------A
          0    150000    300000    450000    600000    750000
MTB > boxplot c2

          ----
          I +I----  *                                  O
          ----
          +---------+---------+---------+---------+---------+------sqrt
          0       200       400       600       800      1000
```

```
MTB > boxplot c3

                      -----------------
                ---------I           +   I----------------------
                      -----------------
           --------+---------+---------+---------+---------+--------log e
                 2.5       5.0       7.5      10.0      12.5
MTB > boxplot c4

                     -------------------
             ----------I            +    I-----------------------
                     -------------------
           ----+---------+---------+---------+---------+---------+--log 10
             1.0       2.0       3.0       4.0       5.0       6.0
```

Note: * and O represent outliers or extreme values.

The shortest summary we could use here is that provided by the **describe** command.

```
MTB > describe c1-c4

                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
A              27    36510     1396     5831   160965    30978
sqrt           27     82.0     37.4     51.8    175.9     33.8
log e          27    6.747    7.241    6.663    2.864    0.551
log 10         27    2.930    3.145    2.894    1.244    0.239

               MIN      MAX       Q1       Q3
A                7   840000       47     4700
sqrt           2.6    916.5      6.9     68.6
log e        1.946   13.641    3.850    8.455
log 10       0.845    5.924    1.672    3.672
```

One way to use this is to compare means and medians. For a symmetric distribution, they should be about the same. For the original data, the mean is about 25 times bigger than the median! For the square roots, it is only about twice as big. For the logs the median is larger, but only by about 7%. That is about as good as we can expect to get.

Comparing means with medians gives us some idea of how ***symmetric*** the distribution is. We might also look at how ***variable*** the different distributions are. The measures of variability we have studied so far are not appropriate, since the numbers in the six columns to be compared differ greatly in size. A useful measure of ***relative*** variability is the **coefficient of variation**, defined as the (absolute value of) the ratio of the standard deviation to the mean. For our data, the c.v.'s are

A	4.4
sqrt	2.1
log e	0.4
log 10	0.4

A large c.v. indicates a large amount of variability, possibly due to long, fat tails or to outliers. For “nice” distributions, the c.v. is considerably less than 1. Here, only the logs appear “nice”.

There are two main lessons to be drawn from our analysis of the island areas data. First, most indicators suggest that a logarithmic transformation would be appropriate for this data. A more general lesson has to do with how well the ***lengths*** of the various summaries suited the purpose at hand. The six dotplots showed too much detail for our purpose of comparing the shapes of the distributions. The stem and leaf and histogram displays showed about the right amount of detail. The boxplots showed enough detail to enable us to select the logarithm as the right transformation, but would not show us some other possible problems (such as bimodality) very well. We also cannot see individual islands. The one-number summaries were generally too short. Because we are interested in the ***shapes*** of the distributions, it is better to use a display that shows the shape rather than a single number summary. Here, stem and leaf plots or histograms, each on its own scale, give the most useful information about shape.