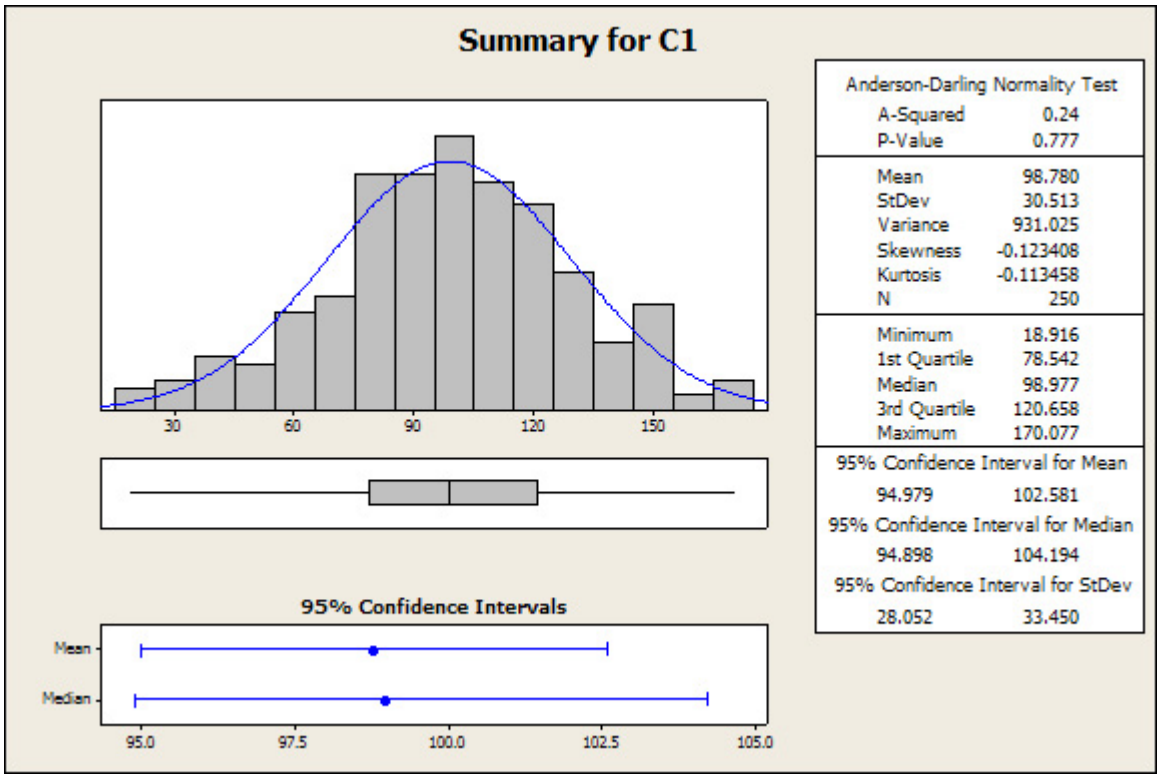
some without indicating nonnormality. If a sample of 30 observations contains 4 outliers, two of which are extreme, would it be reasonable to assume the population from which the data were collected has a normal frequency curve? Probably not.



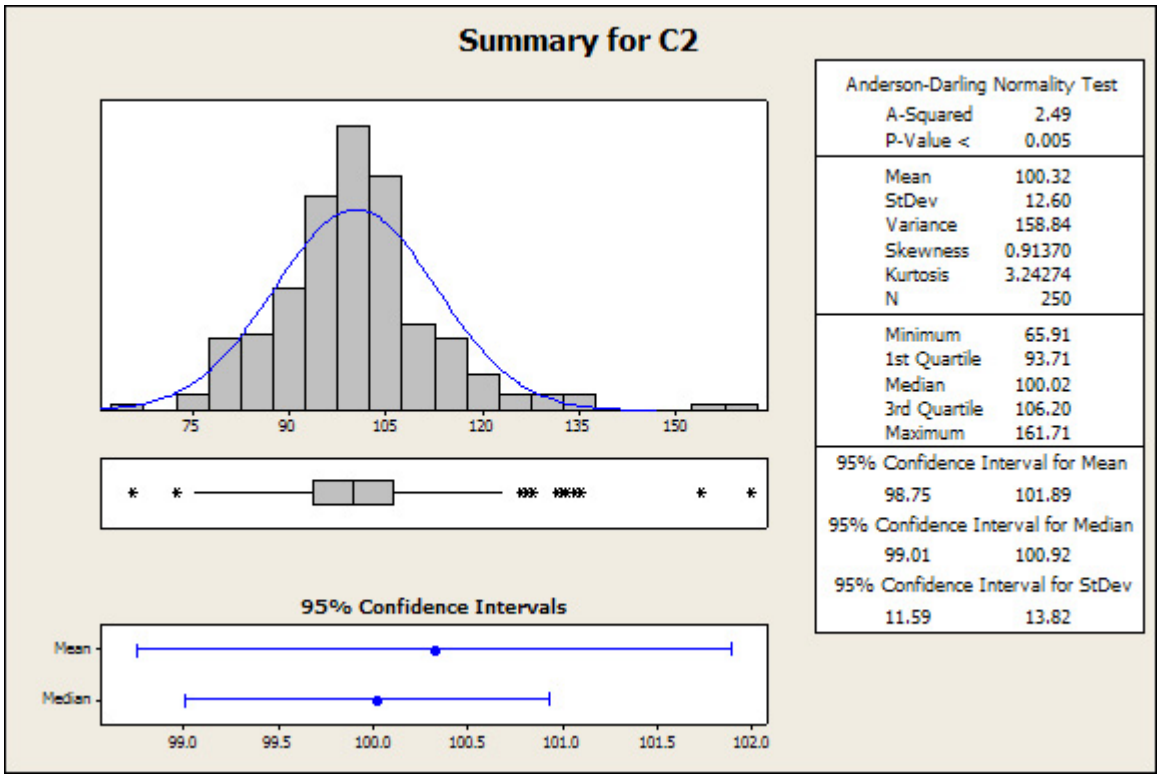

```
Stem-and-Leaf Display: C1

Stem-and-leaf of C1   N   = 250
Leaf Unit = 1.0

  1     1    8
  5     2    1378
  9     3    3379
  17    4    11223567
  25    5    13455789
  38    6    2222444458899
  65    7    11222223345555567777888889
  98    8    000011112233445555666666678888889
(32)    9    11112223344455555555667888899999
  120   10   000123334444444455566667788889
  90    11   001111122233344455566689 9
  64    12   00011111122234444555556 89
  39    13   001112344466779
  24    14   011366677778
  12    15   001133
  6     16   04669
  1     17   0
```

The boxplot is better at highlighting outliers than are other displays. The histogram and stem and leaf displays below appear to have the same basic shape as a normal curve (unimodal, symmetric). However, the boxplot shows that we have a dozen outliers in a sample of 250 observations.

We would only expect about two outliers in 250 observations when sampling from a population with a normal frequency curve. The frequency curve is best described as unimodal, symmetric, and **heavy-tailed**.

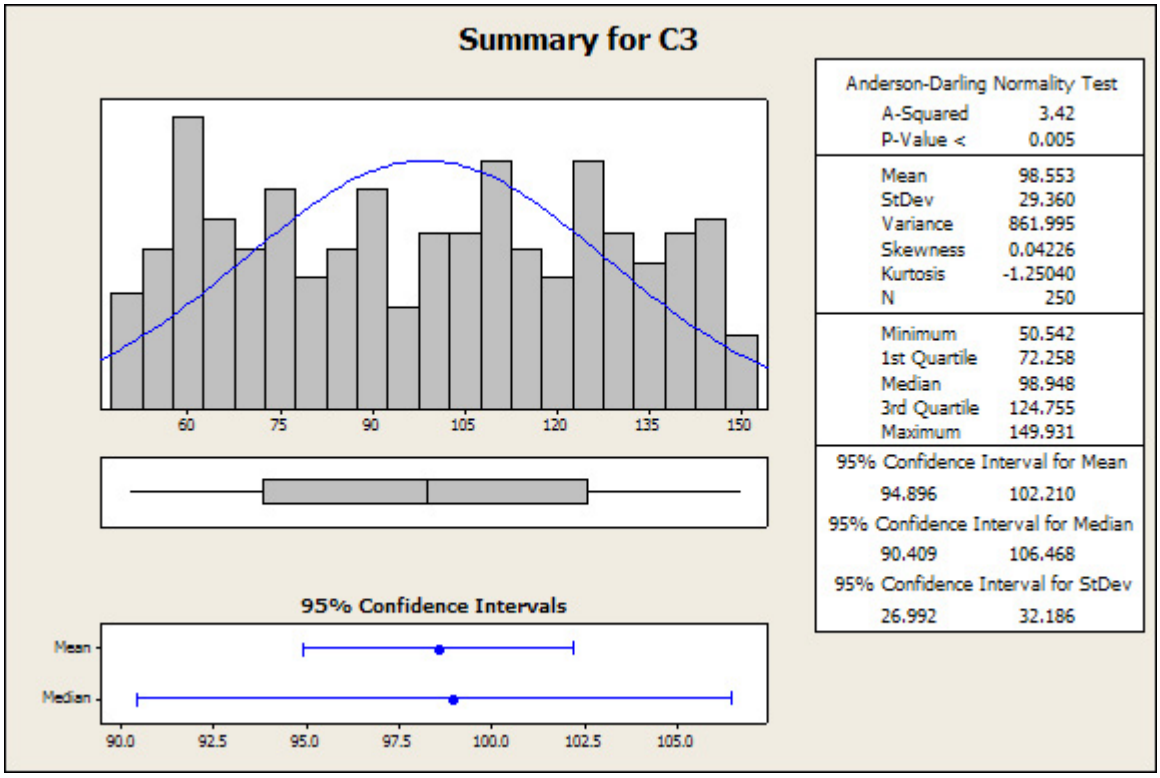



```
Stem-and-Leaf Display: C2

Stem-and-leaf of C2  N  = 250
Leaf Unit = 1.0

 1    6   5
 11   7   2578899999
 45   8   00011222333333344567777778888899999
 124  9   000000111222222333333344444455555555566666666666677777777888888889+
(84)  10  000000000000000000111111111112222223333333334444444444455555566666+
 42   11  00000011122222333345556689
 16   12  000113567
 7    13  12345
 2    14
 2    15  3
 1    16  1
```

Not all symmetric distributions are mound-shaped, as the display below suggests. The boxplot shows symmetry, but the tails of the distribution are shorter (lighter) than in the normal distribution. Note that the distance between quartiles is roughly constant here.



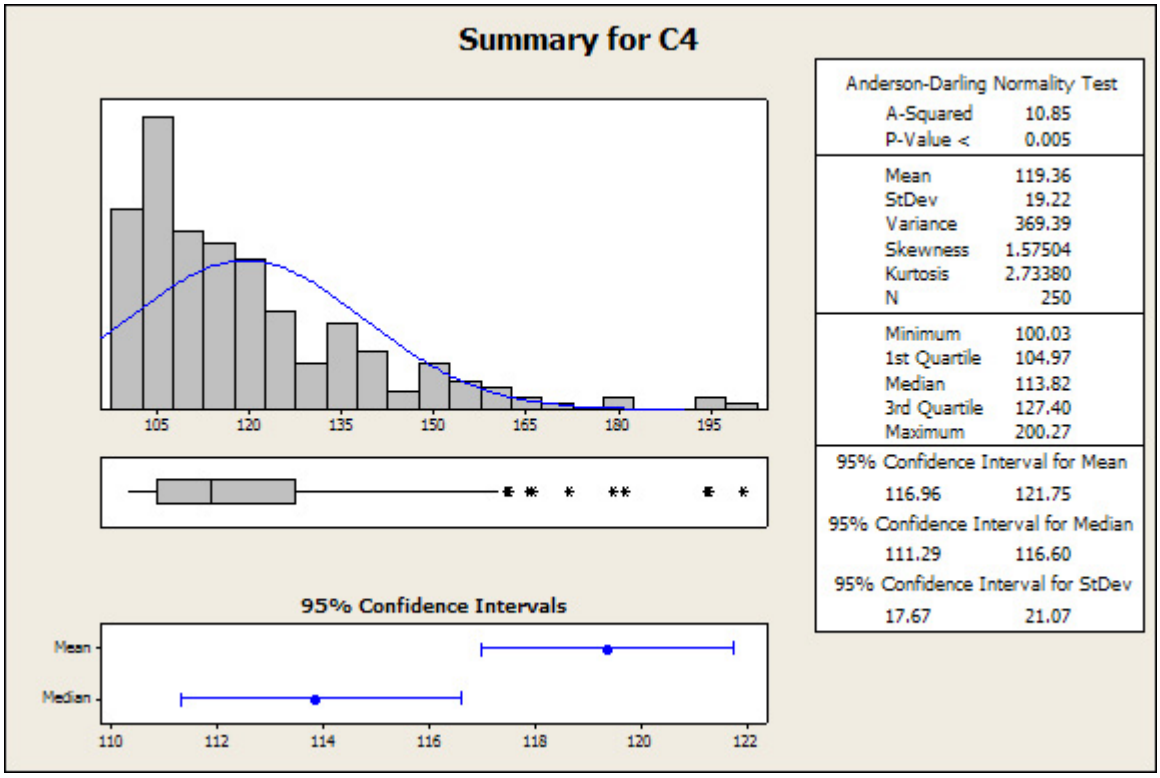

```
Stem-and-Leaf Display: C3

Stem-and-leaf of C3  N  = 250
Leaf Unit = 1.0

 29    5   00111122334555666677777899999
 56    6   000001111223334566666777889
 82    7   000111233344455556667889
 108   8   111222223445566677788888889
(18)   9   001113334466788889
 124   10  000001223455566667788899999
 97    11  00000111223334566668899
 73    12  00113334444445555666778899
 48    13  000001112333444567778889999
 22    14  0001244455566666777999
```

The mean and median are identical in a population with a (exact) symmetric frequency curve. The histogram and stem and leaf displays for a sample selected from a symmetric population will tend to be fairly symmetric. Further, the sample means and medians will likely be close.

The distribution below is unimodal, and asymmetric or **skewed**. The distribution is said to be **skewed to the right**, or upper end, because the right tail is much longer than the left tail. The boxplot also shows the skewness - the region between the minimum observation and the median contains half the data in less than 1/5 the range of values. In addition, the upper tail contains several outliers.

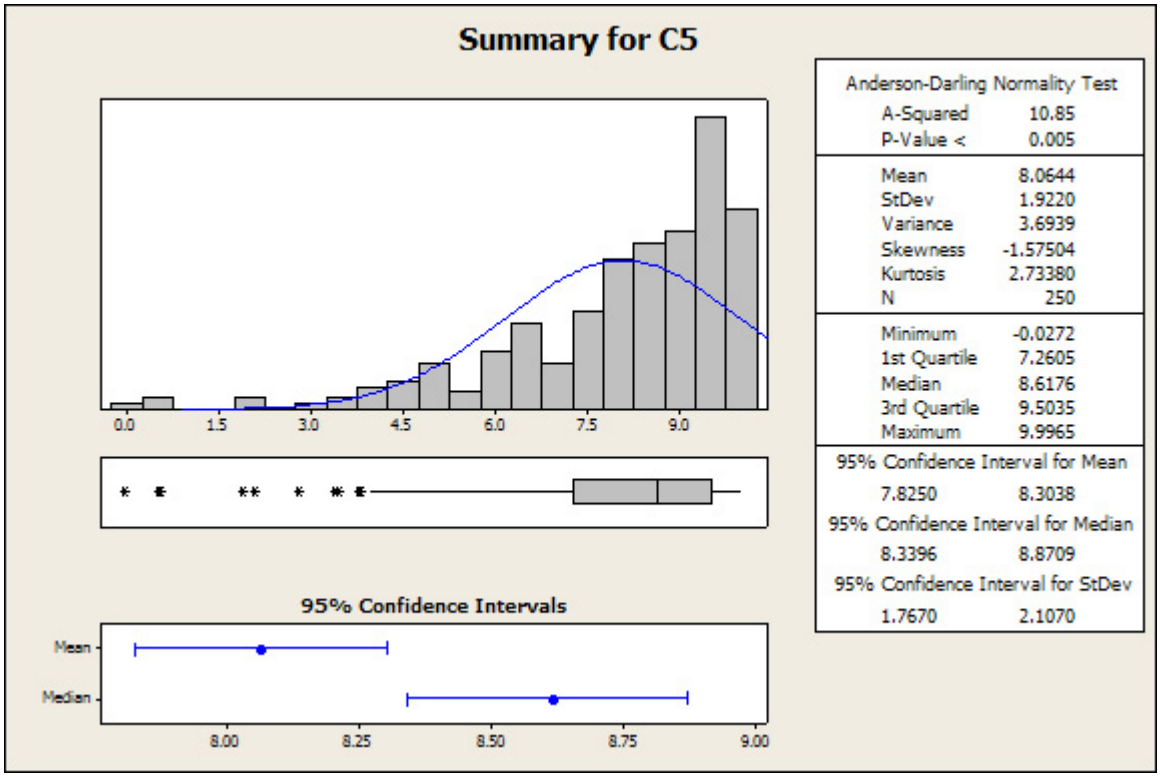



```
Stem-and-Leaf Display: C4

Stem-and-leaf of C4  N  = 250
Leaf Unit = 1.0

 105   10  000000000000000000000001111111111222222222333333333333334444444555+
(57)   11  000111112222222333333344444444455666667777777788888899999
 88    12  000000011223344445556666677788
 57    13  0112333445555666678889
 34    14  00112234489
 23    15  00111235556
 12    16  001256
 6     17  19
 4     18  1
 3     19  44
 1     20  0
```

The distribution below is unimodal and **skewed to the left**. The two examples show that extremely skewed distributions often contain outliers in the longer tail of the distribution.

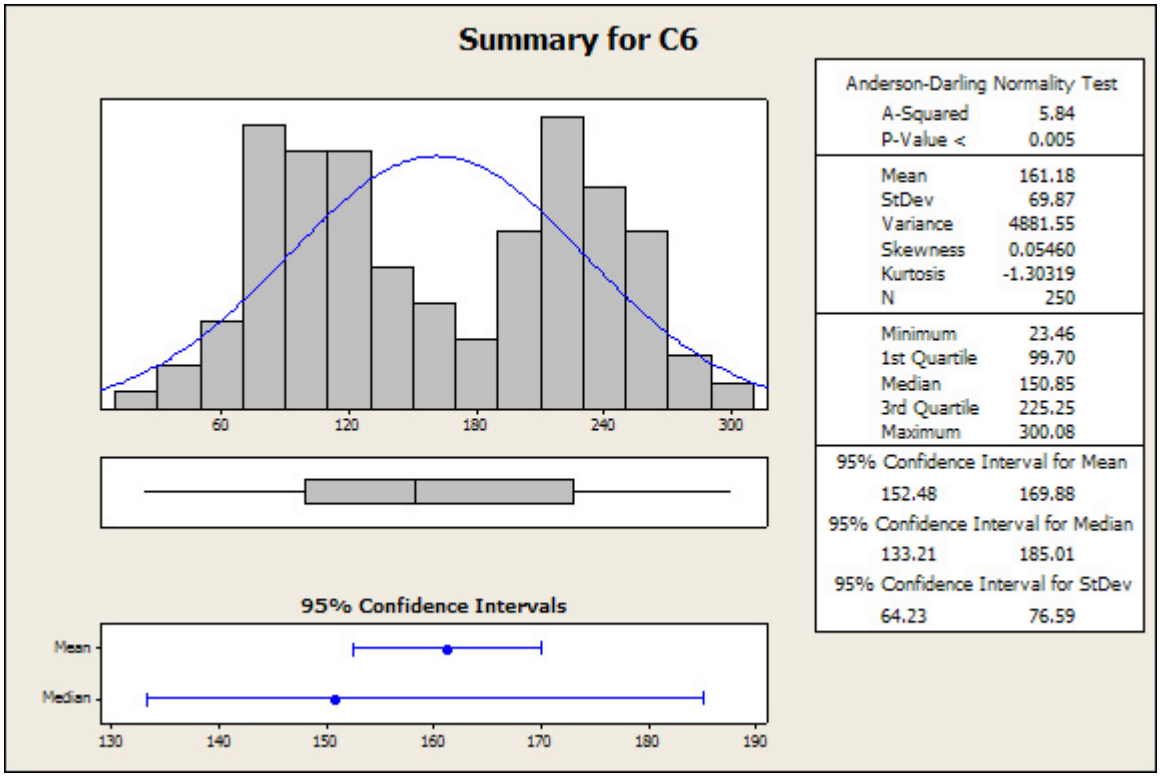



```
Stem-and-Leaf Display: C5

Stem-and-leaf of C5  N  = 250
Leaf Unit = 0.10

 3    0   055
 4    1   8
 6    2   08
 12   3   347899
 23   4   34446788899
 34   5   01556778899
 57   6   011122333344445566678889
 88   7   11222233333444555566778899999999
(57)  8   000001111112222222233333445555555556666666777777788888999
 105  9   00000000111111111122222223333333444444444445555555666666666666677+
 0    10
```

Not all distributions are unimodal. The distribution below has two modes or peaks, and is said to be **bimodal**. Distributions with three or more peaks are called **multi-modal**.




```
Stem-and-Leaf Display: C6

Stem-and-leaf of C6  N  = 250
Leaf Unit = 10

  4     0  2233
 12     0  44455555
 32     0  66666777777777777777
 64     0  88888888888888888999999999999999
 95     1  00000000000001111111111111111
 115    1  22222222222333333333
(15)    1  444444445555555
 120    1  6666677777
 110    1  88899999999
 99     2  000000000001111111111111111
 71     2  22222222222222222333333333333333
 38     2  4444444445555555555555
```

The boxplot and histogram or stem and leaf display (or dotplot) are used **together** to describe the distribution. The boxplot does not provide information about modality - it only tells you about skewness and the presence of outliers.

As noted earlier, many statistical methods assume the population frequency curve is normal. Small deviations from normality usually do not dramatically influence the operating characteristics of these methods. We worry most when the deviations from normality are severe, such as extreme skewness or heavy tails containing multiple outliers.

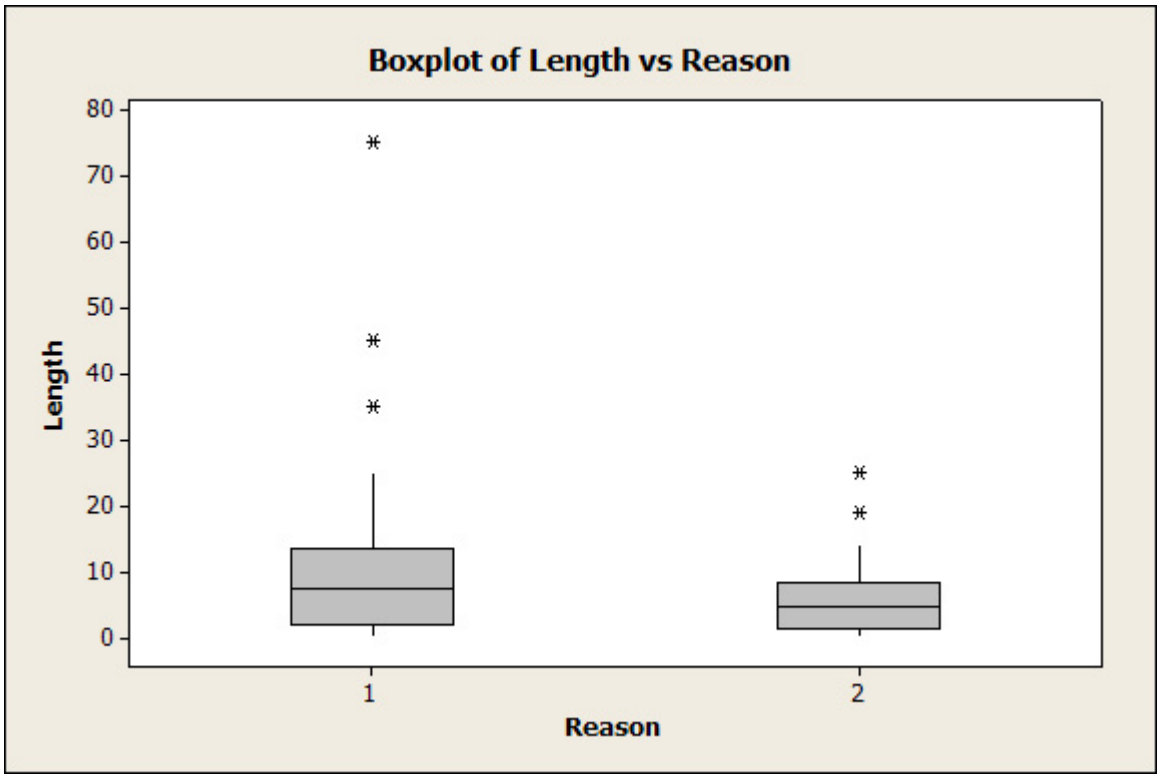
Interpretations for Examples

The **MAO** samples are fairly symmetric, unimodal (?), and have no outliers. The distributions do not deviate substantially from normality. The various measures of central location ($\overline{Y}$, M) are fairly close, which is common with reasonably symmetric distributions containing no outliers.

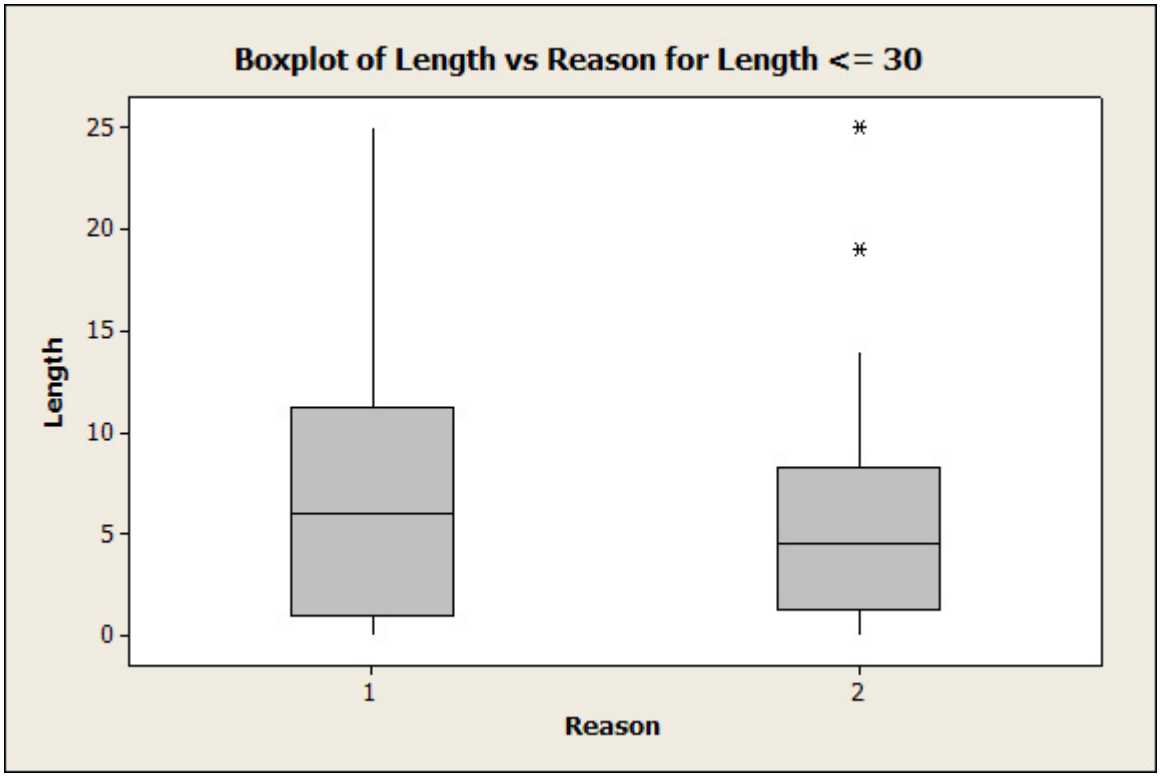
The **SIDS** sample is unimodal, and skewed to the right due to the presence of four outliers in the upper tail. Although not given, we expect the mean to be noticeably higher than the median (Why?). A normality assumption here is unrealistic.

Example: Length of Stay in a Psychiatric Unit

Data on all 58 persons committed voluntarily to the acute psychiatric unit of a health care canter in Wisconsin during the first six months of a year are stored in the worksheet HCC that installs with Minitab. Two of the variables are Length (of stay, in number of days), and Reason (for discharge, 1=normal, 2=other). It is of interest to see if the length of stay differs for the two types of discharge. The main parts of the boxplots comparing the groups are rather compressed (and not very useful) because outliers are using up all the scale.



The solution in a case like this is to zoom in using the Data Options in the boxplot display. In this case let us exclude rows where Length > 30. Now we have a little more basis for comparison.



Does it look like there is a really large difference between the groups? What would you say about the shape of the distributions? Does it look like these are normally distributed values?

Examine the descriptive statistics. What is a reasonable summary here and what probably is pretty distorted? What is your summary of the data based upon the boxplots and numerical summaries?

Descriptive Statistics: Length

Variable	Reason	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Length	1	42	0	11.55	2.25	14.60	0.00	1.75	7.50	13.50
	2	16	0	6.44	1.78	7.11	0.00	1.25	4.50	8.25

Variable	Reason	Maximum
Length	1	75.00
	2	25.00