Numerical Measures for Grouped Data

Learning Objectives

- sample mean
- sample variance
- sample standard deviation
- sample median
- empirical rule
- box plots
- skewness and kurtosis

When we encounter situations where the data are grouped in the form of a frequency table, we no longer have individual data values.

The following formulas will give approximate values for x and s^2. Let the grouped data have l classes, with m_i being the midpoint and f_i being the frequency of class i, $i = 1, 2, \ldots, l$.

Let $n = \sum_{i=1}^{l} f_i$

Definition 1. *The* **mean** *for a sample of size n,*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{l} f_i m_i,$$

where m_i is the midpoint of the class i and f_i is the frequency of the class i.

Similarly the **sample variance**,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} f_i \left(m_i - \overline{x}\right)^2 = \frac{\sum m_i^2 f_i - \frac{\left(\sum_i f_i m_i\right)^2}{n}}{n-1}.$$

Example 1.

The grouped data in Table 1. represent the number of children from birth through the end of the teenage years in a large apartment complex. Find the mean, variance, and standard deviation for these data:

Table 1. Number of Children and Their Age Group

Class	0–3	4–7	8–11	12–15	16–19
Frequency	7	4	19	12	8

Solution

For simplicity of calculation we create Table 2.

Table 2.

Class	f_i	m_i	$m_i f_i$	$m_i^2 f_i$
0–3	7	1.5	10.5	15.75
4–7	4	5.5	22	121
8–11	19	9.5	180.5	1714.75
12–15	12	13.5	162	2187
16–19	8	17.5	140	2450
	$n = 50$		$\sum m_i f_i = 515$	$\sum m_i^2 f_i = 6488.5$

The sample mean is

$$\bar{x} = \frac{1}{n}\sum_i f_i m_i = \frac{515}{50} = 10.30.$$

The sample variance is

$$s^2 = \frac{\sum m_i^2 f_i - \frac{\left(\sum_i f_i m_i\right)^2}{n}}{n-1} = \frac{6488.5 - \frac{(515)^2}{50}}{49} = 24.16.$$

The sample standard deviation is $s = \sqrt{s^2} = \sqrt{24.16} = 4.92.$

Using the following calculations, we can also find the *median for grouped data*. We only know that the median occurs in a particular class interval, but we do not know the exact location of the median. We will assume that the measures are spread evenly throughout this interval. Let

L = lower class limit of the interval that contains the median
n = total frequency
F_b = cumulative frequencies for all classes before the median class
f_m = frequency of the class interval containing the median
w = interval width of the interval that contains the median

Then the median for the grouped data is given by

$$M = L + \frac{w}{f_m}(0.5n - F_b).$$

Example 2

For the data of Example 1, find the median.

Solution

First develop Table 3.

Table 3.

Class	f_i	Cumulative f_i	Cumulative f_i/n
0–3	7	7	0.14
4–7	4	11	0.22
8–11	19	30	0.6
12–15	12	42	0.84
16–19	8	50	1.00

The first interval for which the cumulative relative frequency exceeds 0.5 is the interval that contains the median. Hence the interval 8 to 11 contains the median. Therefore, $L = 8$, $f_m = 19$, $n = 50$,$w = 3$, and $F_b = 11$. Then, the median is

$$M = L + \frac{w}{f_m}(0.5n - F_b) = 8 + \frac{3}{19}((0.5)(50) - 11) = 10.211.$$

It is important to note that all the numerical measures we calculate for grouped data are only approximations to the actual values of the ungrouped data if they are available.

One of the uses of the sample standard deviation will be clear from the following result, which is based on data following a bell-shaped curve. Such an indication can be obtained from the histogram.

EMPIRICAL RULE

When the histogram of a data set is “bell shaped” or “mound shaped,” and symmetric, the empirical rule states:

1. Approximately 68% of the data are in the interval $(x - s, x + s)$.
2. Approximately 95% of the data are in the interval $(x - 2s, x + 2s)$.
3. Approximately 99.7% of the data are in the interval $(x - 3s, x + 3s)$.

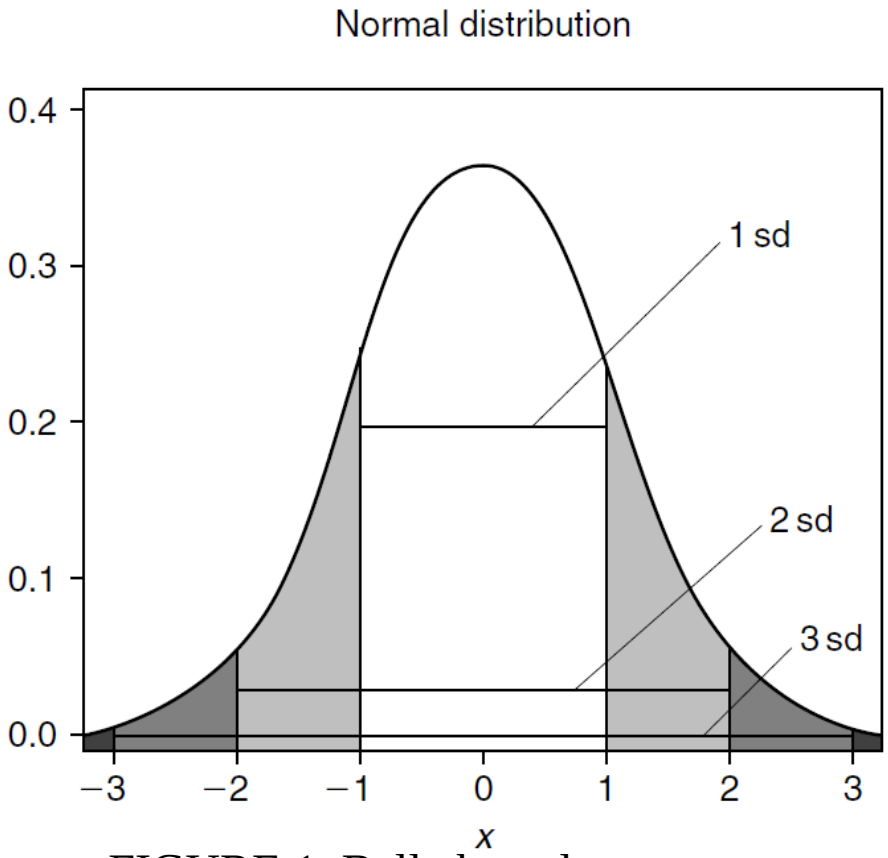


FIGURE 1. Bell-shaped curve.

The bell-shaped curve is called a normal curve and is discussed later. A typical symmetric bell-shaped curve is given in Figure 1.

Example 3

The scores for all high school seniors taking the verbal section of the Scholastic Aptitude Test (SAT) in a particular year had a mean of 490 and a standard deviation of 100. The distribution of SAT scores is bell-shaped.

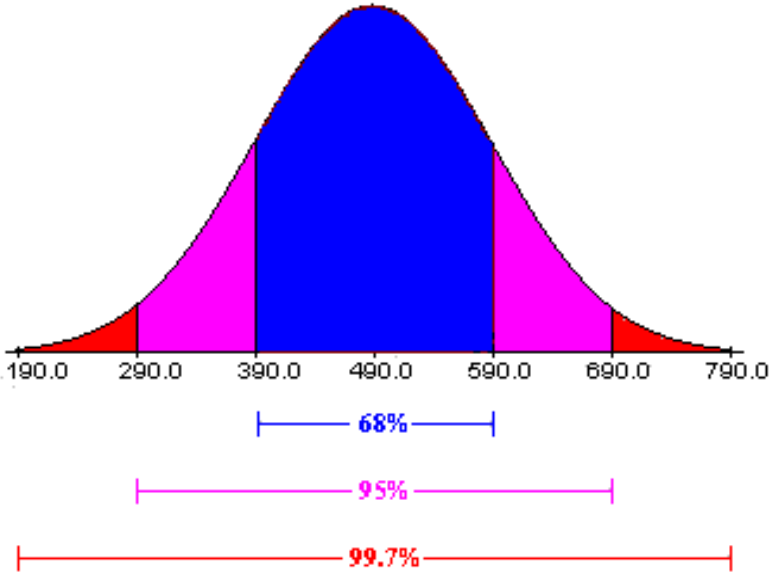
A. What percentage of seniors scored between 390 and 590 on this SAT test?

B. One student scored 795 on this test. How did this student do compared to the rest of the scores?

C. A rather exclusive university only admits students who were among the highest 16% of the scores on this test. What score would a student need on this test to be qualified for admittance to this university?

Solution

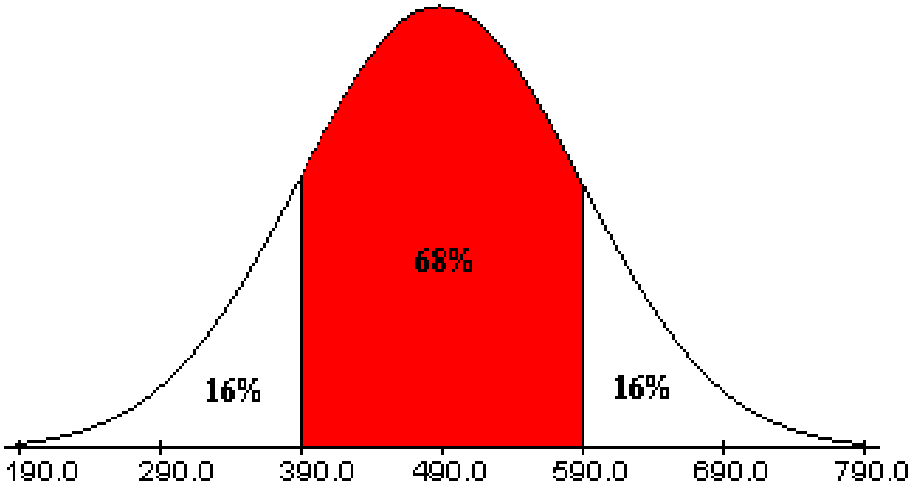
The data being described are the verbal SAT scores for all seniors taking the test one year. Since this is describing a population, we will denote the mean and standard deviation as ***m = 490*** *and* ***s = 100****, respectively. A bell shaped curve summarizing the percentages given by the empirical rule is below.*



A. From the figure above, about 68% of seniors scored between 390 and 590 on this SAT test.

B. Since about 99.7% of the scores are between 190 and 790, a score of 795 is excellent. This is one of the highest scores on this test.

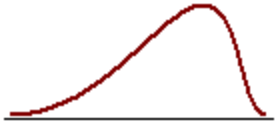
C. Since about 68% of the scores are between 390 and 590, this leaves 32% of the scores outside this interval. Since a bell-shaped curve is symmetric, one-half of the scores, or 16%, are on each end of the distribution. The figure below shows these percentages.



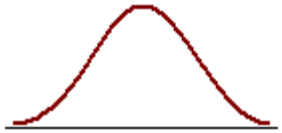
Since about 16% of the students scored above 590 on this SAT test, to be qualified for admittance to this university, a student would need to score 590 or above on this test.

Skewness and Kurtosis

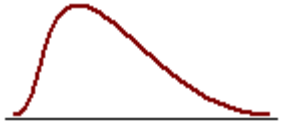
*The coefficient of **Skewness** is a measure for the degree of symmetry in the variable distribution.*



Negatively skewed distribution
or Skewed to the left
Skewness <0

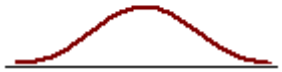


Normal distribution
Symmetrical
Skewness = 0



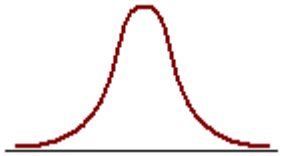
Positively skewed distribution
or Skewed to the right
Skewness > 0

*The coefficient of **Kurtosis** is a measure for the degree of peakedness/flatness in the variable distribution.*



Platykurtic distribution
Low degree of peakedness
Kurtosis <0

Normal distribution
Mesokurtic distribution
Kurtosis = 0



Leptokurtic distribution
High degree of peakedness
Kurtosis > 0

Box Plots

The sample mean or the sample standard deviation focuses on a single aspect of the data set, whereas histograms and stem-and-leaf displays express rather general ideas about data. A pictorial summary called a *box plot* (also called *box-and-whisker plots*) can be used to describe several prominent features of a data set such as the center, the spread, the extent and nature of any departure from symmetry, and identification of outliers. Box plots are a simple diagrammatic representation of the five number summary: minimum, lower quartile, median, upper quartile, maximum.

PROCEDURE TO CONSTRUCT A BOX PLOT

1. Draw a vertical measurement axis and mark Q1, Q2 (median), and Q3 on this axis as shown in Figure 2.
2. Construct a rectangular box whose bottom edge lies at the lower quartile, Q1 and whose upper edge lies at the upper quartile, Q3.
3. Draw a horizontal line segment inside the box through the median.
4. Extend the lines from each end of the box out to the farthest observation that is still within 1.5(IQR) of the corresponding edge. These lines are called whiskers.
5. Draw an open circle (or asterisks *) to identify each observation that falls between 1.5(IQR) and 3(IQR) from the edge to which it is closest; these are called mild outliers.
6. Draw a solid circle to identify each observation that falls more than 3(IQR) from the closest edge; these are called extreme outliers.

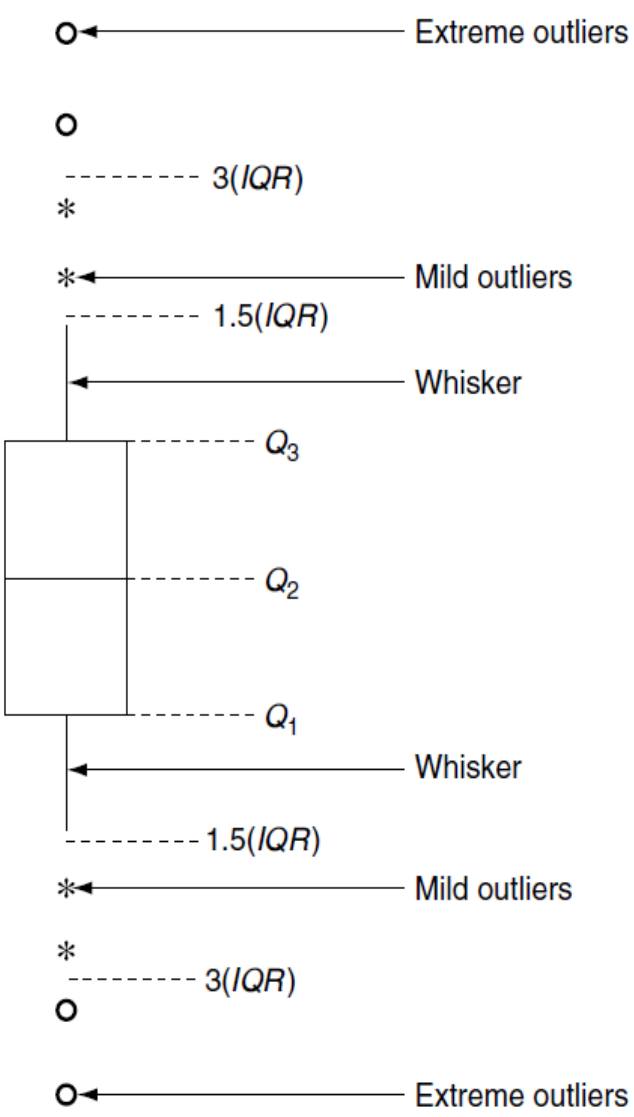


FIGURE 2. A typical box-and-whiskers plot.

Example 4.

The following data identify the time in months from hire to promotion to chief pharmacist for a random sample of 25 employees from a certain group of employees in a large corporation of drugstores.

5	7	229	453	12	14	18	14	14	483
22	21	25	23	24	34	37	34	49	64
47	67	69	192	125					

Construct a box plot. Do the data appear to be symmetrically distributed along the measurement axis?

Solution

We find that the median, $Q_2 = 34$.

The lower quartile is $Q_1 = \frac{14+18}{2} = 16$.

The upper quartile is $Q_3 = \frac{67+69}{2} = 68$.

The interquartile range is $IQR = 68 - 16 = 52$.

To find the outliers, compute

$$Q_1 - 1.5(IQR) = 16 - 1.5(52) = -62$$

and

$$Q_3 + 1.5(IQR) = 68 + 1.5(52) = 146.$$

*Using these numbers, we follow the procedure outlined earlier to construct the box plot in Figure 3. The * in the box plot represents an outlier. The first horizontal line is the first quartile, the second is the median, and the third is the third quartile.*

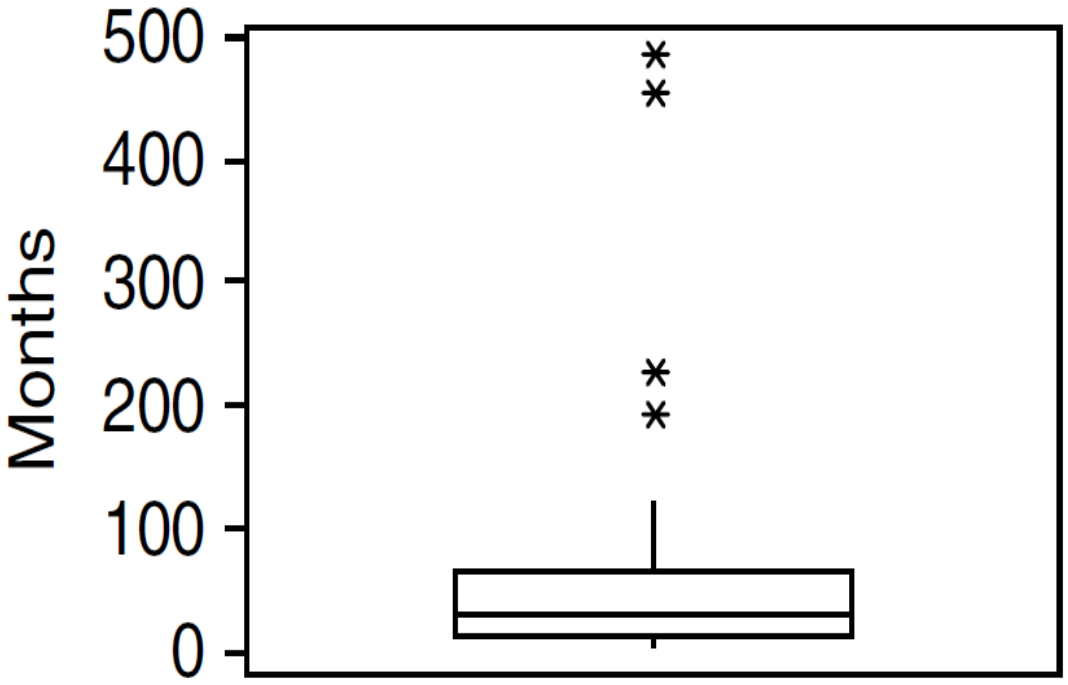
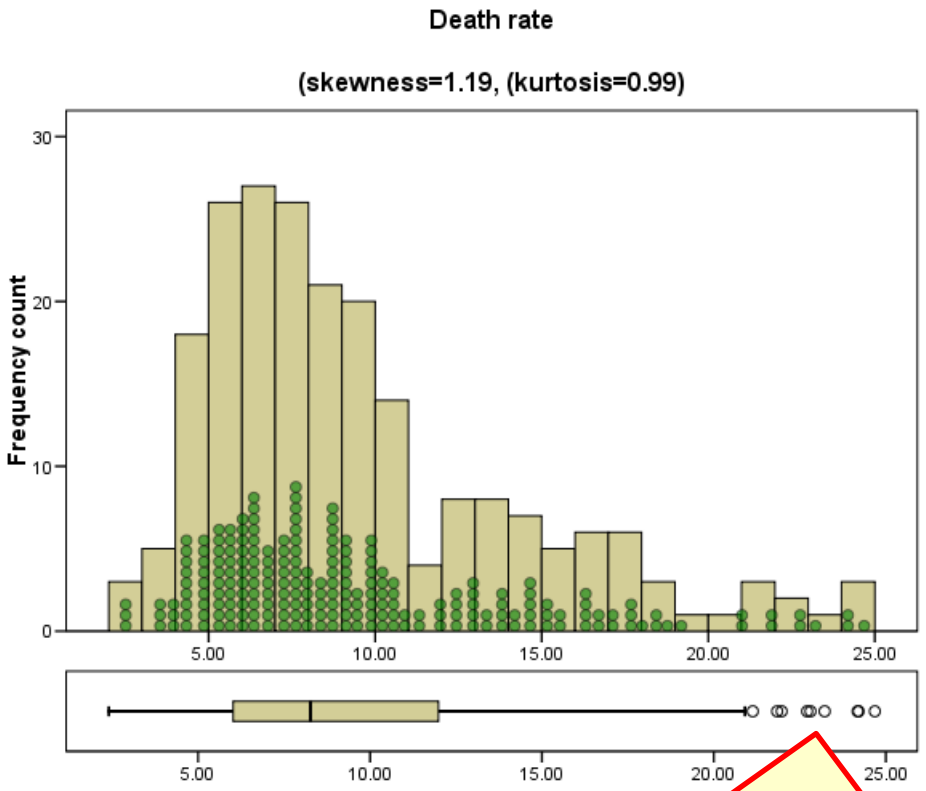


FIGURE 4. Box plot for months to promotion.

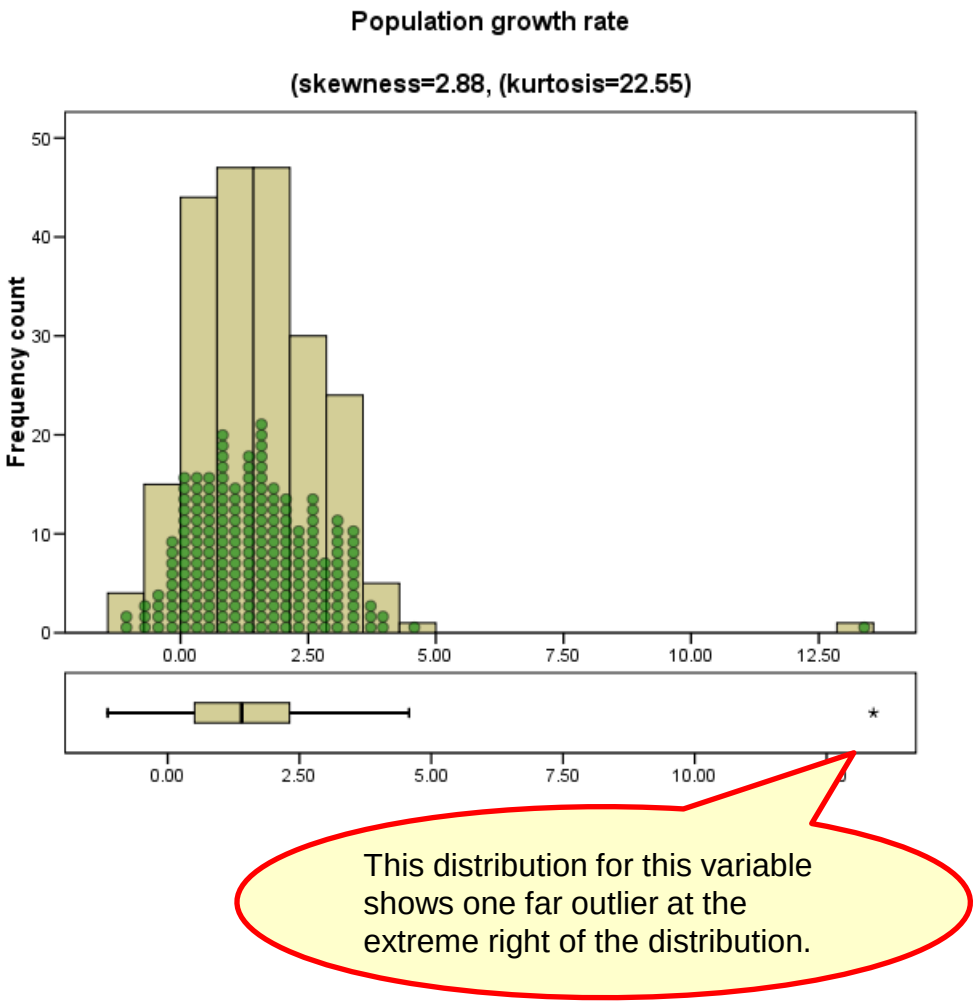
By examining the relative position of the median line (the middle line in Figure 3.), we can test the symmetry of the data. For example, in Figure 3., the median line is closer to the lower quartile than the upper line, which suggests that the distribution is slightly nonsymmetric. Also, a look at this box plot shows the presence of two mild outliers and two extreme outliers.

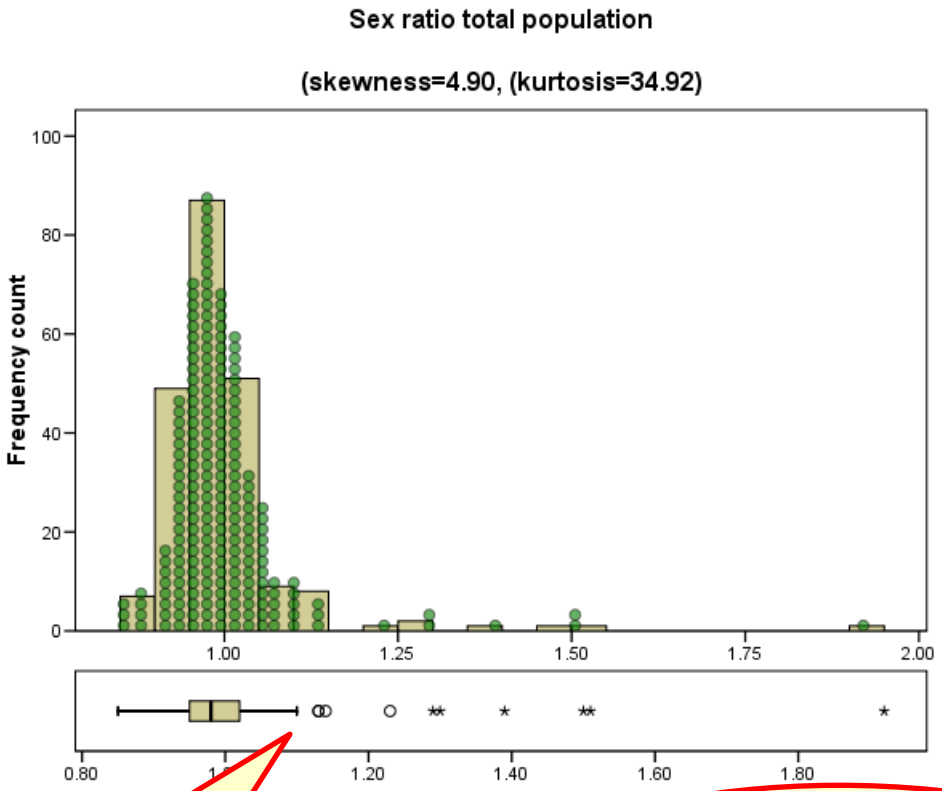
Box Plots and Histogram



The script for this week positions the boxplot under the histogram. In this chart, we see a number of circles at the right end of the distribution. These are outliers, and there are no far outliers in this distribution. As we would expect, this distribution has a skewness problem (skewness=1.19) in the subtitle to the chart.

NOTE: the horizontal axis for the boxplot approximates the axis for the histogram, but is not exact.

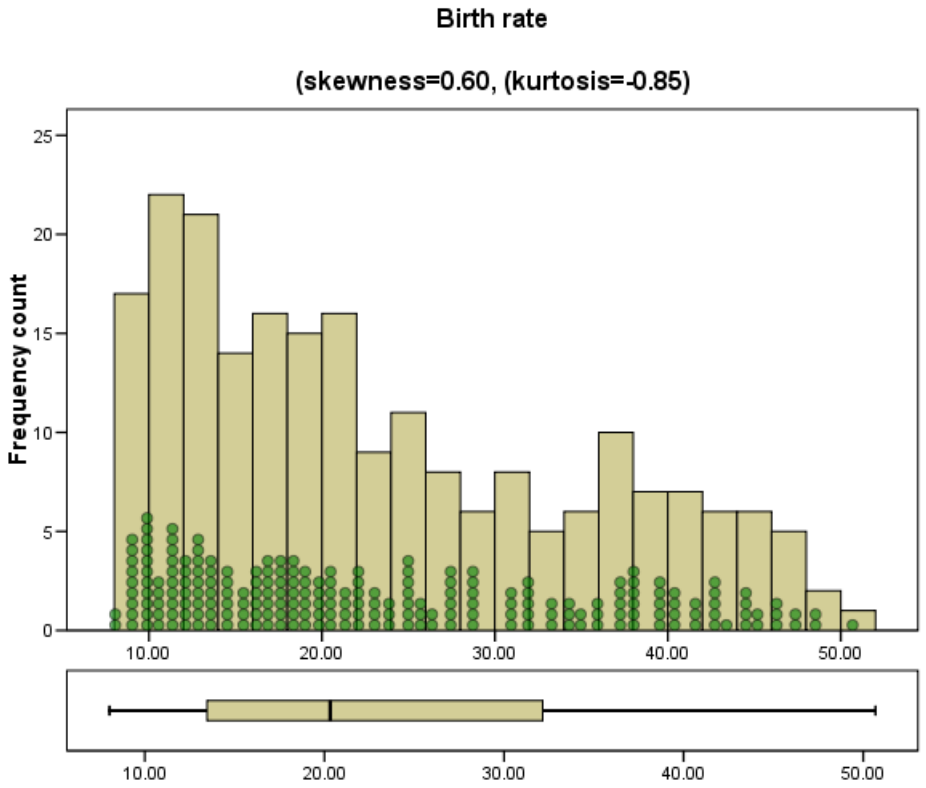

Population growth rate
(skewness=2.88, (kurtosis=22.55)
Frequency count
50
40
30
20
10
0
0.00
2.50
5.00
7.50
10.00
12.50
*
0.00
2.50
5.00
7.50
10.00
This distribution for this variable shows one far outlier at the extreme right of the distribution.



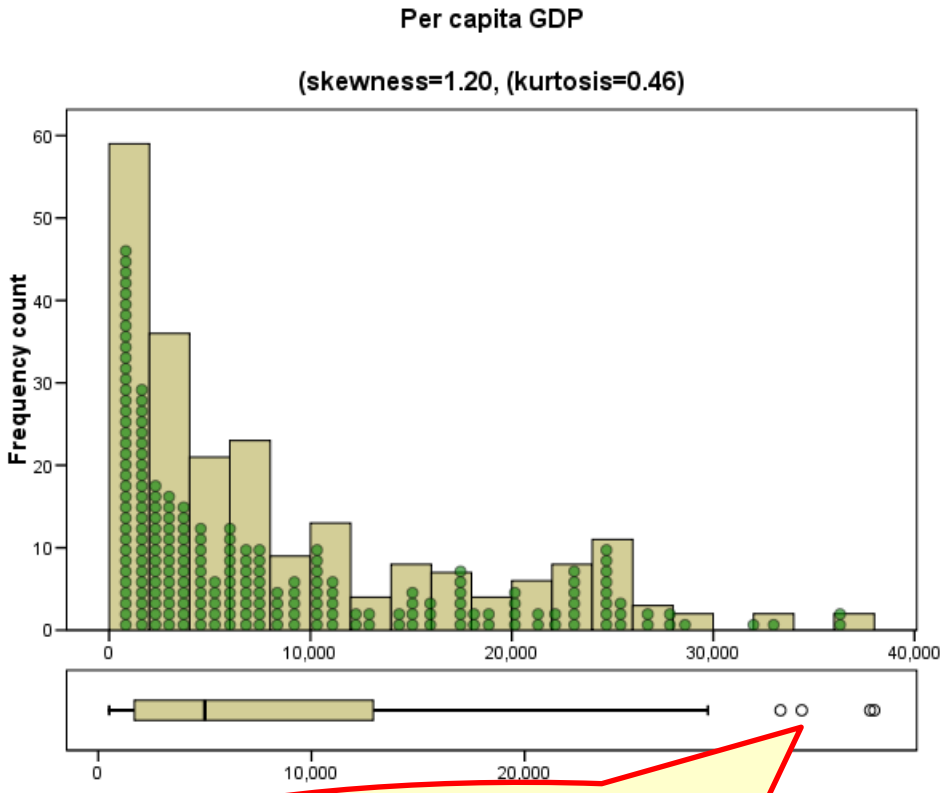
Some distributions will show both outliers and far outliers.

Our problems will state the number of outliers, and the number of far outliers as a subset of the total number of outliers.

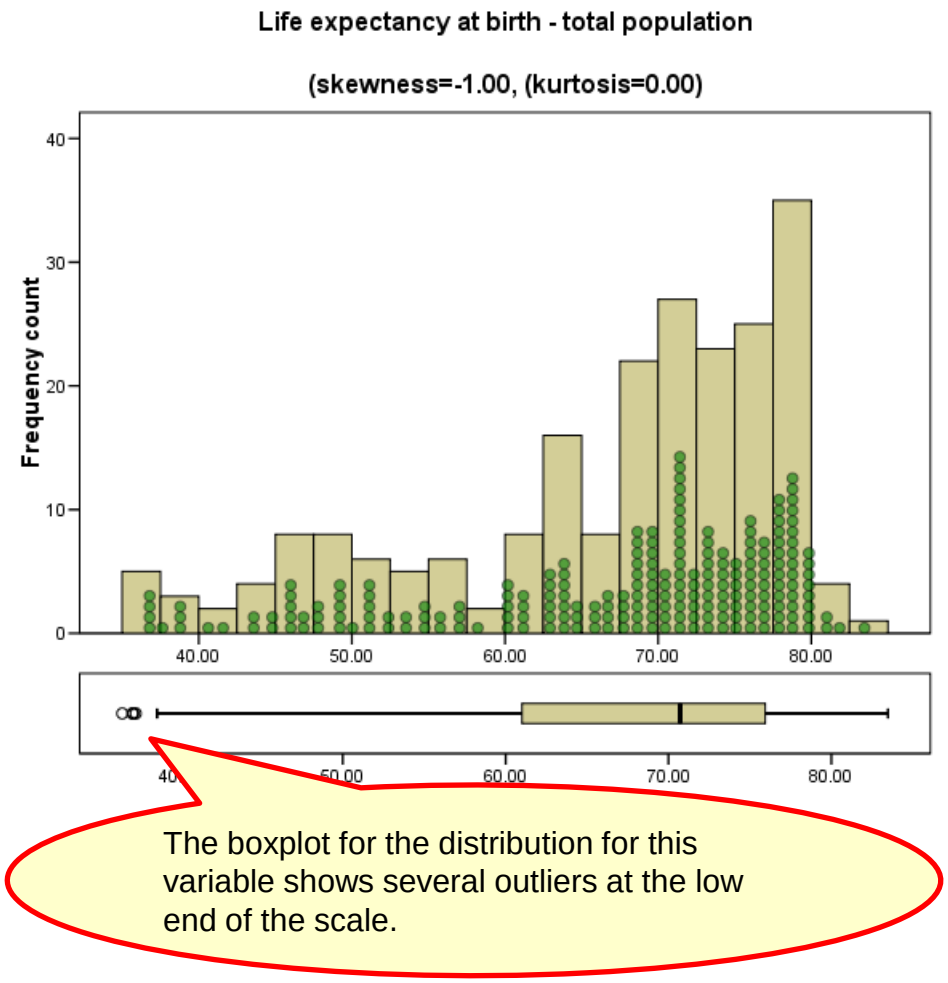
Note that the chart shows the presence or absence of outliers, but does not necessarily provide an exact count since the outlier symbol might represent more than one case with the score.

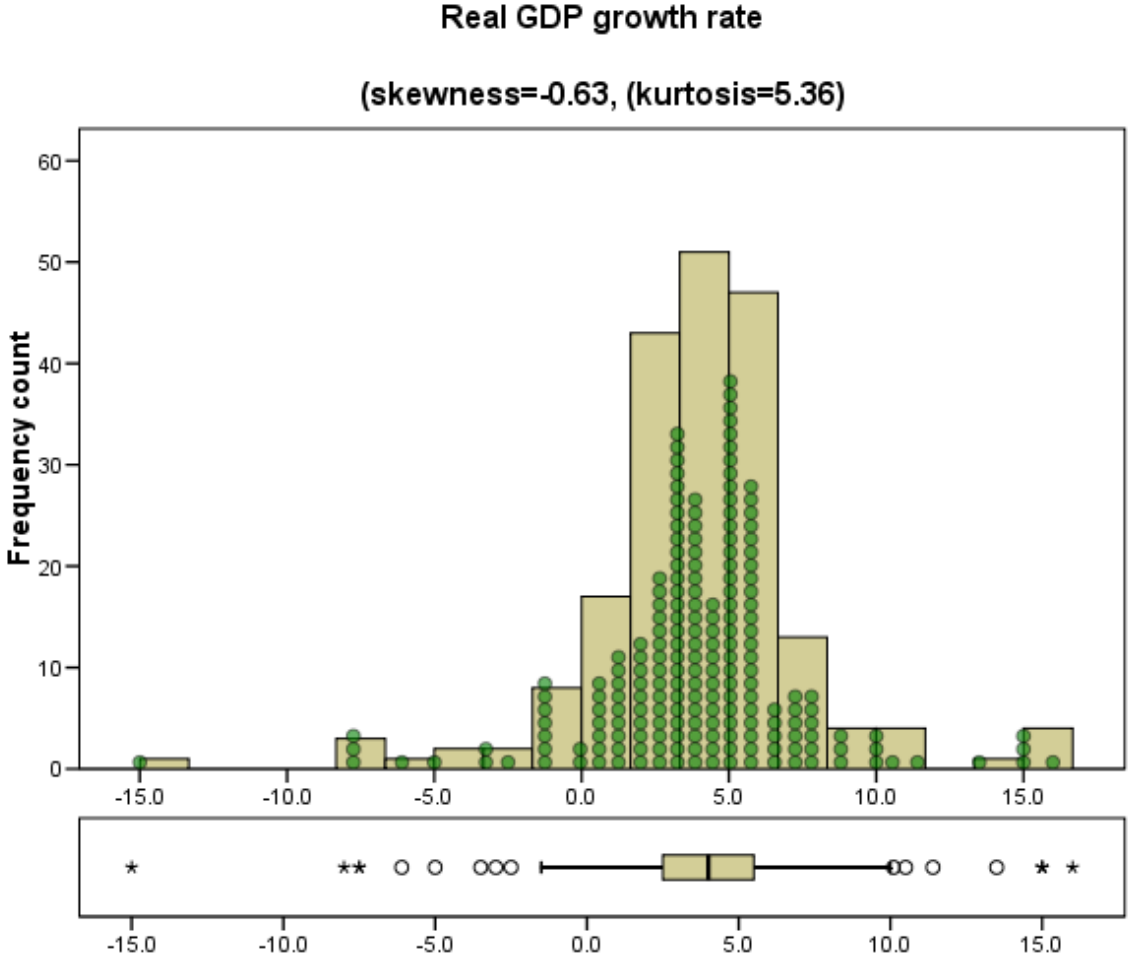


The boxplots for some distributions will indicate that there are no outliers.



The boxplot for the distribution for this variable shows several outliers at the right end of the distribution.


Life expectancy at birth - total population
(skewness=-1.00, (kurtosis=0.00)
Frequency count
40
30
20
10
0
40.00
50.00
60.00
70.00
80.00
40
50.00
60.00
70.00
80.00
The boxplot for the distribution for this variable shows several outliers at the low end of the scale.


Real GDP growth rate
(skewness=-0.63, (kurtosis=5.36)
Frequency count
60
50
40
30
20
10
0
-15.0
-10.0
-5.0
0.0
5.0
10.0
15.0