CHAPTER TWO

Descriptive Statistics

2.1 Central Tendency

Central tendency is an attempt to devise a statistical method that yields a single value that would tell us something about the *typical* value(s) of a distribution. The three most common central tendency statistics are the arithmetic mean, the median, and the mode.

The **mean** or arithmetic mean or average is the sum of all values divided by the number of values. When the mean is determined from a subset of data of a population, it is often denoted by the symbol $\overline{X}$ or ***M***. When the mean is of the entire population, it is designated by the symbol, μ. All the scores of a dataset contribute to the calculation of the arithmetic mean. The number of scores or individuals in a sample dataset is called the *sample size* and is denoted by the symbol ***n*** (***N*** for the population size).

The **mean** (arithmetic mean) of a dataset is the sum of the scores divided by the number of scores.
The *population* mean is

$\mu = \frac{\Sigma X}{N}$, where ΣX is all scores in population; N is size of population

The *sample* mean is

sample mean $= M = \overline{X} = \frac{\Sigma X}{n}$, where ΣX is all scores in sample; n is sample size

Problem 2.1.1: Calculate the mean for a population of data {1, 3, 5, 9, 12}.

The mean is

$$\mu = \frac{\Sigma X}{N} = \frac{30}{5} = 6$$

The **median** is a central tendency statistics that attempts to find the exact center of or mid-point of the data or scores. The median is the value that separates the upper half of the date set or distribution from the lower half; often this is called the 50^{th} percentile.

The **median** is the score that divides a distribution of data scores exactly in half. Exactly 50% of the individuals in a distribution have scores at or below the median.

To calculate the median, order or rank the dataset (distribution of dataset) from smallest to largest. When *N* is an **odd number**, the median is the middle score. When *N* is an **even number**, the median is the average of the two middle scores.

Median - Method 1: **Odd *N***

Find the median for {2, 4, 5, 3, 8, 10, 7}

Step 1: Order data from smallest to largest: {2, 3, 4, **5**, 7, 8, 10}

Step 2: Locate middle score: So median is **5**

Median – Method 2: **Even *N***

Find the median of {7, 3, 5, 8, 12, 9}

Step 1: Order data from smallest to largest: {3, 5, **7**, **8**, 9, 12}

Step 2: Locate the two middle scores: {7, 8}

Step 3: Add two middle scores and divide by 2 to get the median:

$$\text{median} = \frac{7+8}{2} = \frac{15}{2} = 7.5$$

The **mode** is the statistics that shows which score(s) is the most frequent. The mode is often found by selecting the score(s) that has the highest frequency from the *simple frequency* distribution table. A bimodal distribution has two modes.

The **mode** is the score(s) that occurs most frequently. In a frequency distribution, the mode is the score or category that has the highest frequency.

The best method to find the mode, especially of a large dataset, is to identify the score(s) that is most frequent from a frequency distribution. In Figure 2.1.1 below, the modes are 4 and 7 (this is a bimodal distribution).

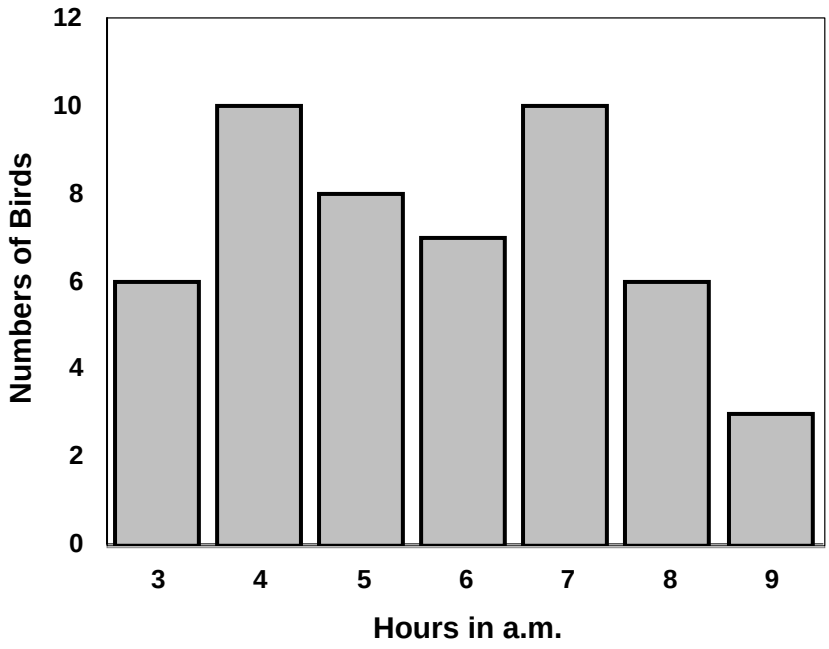


Figure 2.1. 1 Number of Birds Observed in the Morning

Tony and Central Tendency

Problem 2.1.2: From the following frequency distribution table, compute the mean, median and mode:

X	7	6	5	4	3	2	1	0
Frequency, f	1	0	3	2	3	5	4	2

Tony: "It seems that central tendency determination can be easily done by observations and minimal calculations with small dataset; but how do perform such calculations with large sets of data?"

Rose: "For small dataset, I would recommend using MS Excel or a similar spreadsheet to calculate or assess these statistics (mean, median, and mode)." "Often a frequency distribution, as shown in the problem frequency table above, comes from a much larger dataset and if the values of *X* are the exact scores (not midpoints or other class width calculations – see Chapter 2.2 on Frequency Distribution) one can reconstruction the original dataset: {7, 5, 5, 5, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 1, 1, 1, 1, 0, 0}." "Often, only from a frequency distribution can you tell the mode of a distribution, especially if there are more than one mode; the mode for this problem is **2**." From an Excel spreadsheet or statistics program, the mean is **2.65** and the median is **2** (see Figure below)."

Problem 2.1.3: Use SPSS or Stats4U to generate the descriptive statistics showing the mean, median and mode for the *pass4th* variable from the ODE data table.

Table 2.1.1 *Central Tendency Measures of 9th Grade (ODE)*

Statistics	**Values**
Mean	65.86
Median	67.00
Mode	68, 73

*Multiple modes exist. The smallest value is shown in the frequency statistical summary.

Central Tendency

n =	20				
Mean =	**2.65**	**Median =**	**2**	**Mode =**	**2**
					(only one mode shown)

Enter Data Below
7
5
5
5
4
4
3
3
3
2
2
2
2
2
1
1
1
1
0
0
End of Data Entry

Figure 2.1.2 Excel Output: Central Tendency

The Excel program for central tendency is designed so that you can enter data insert or delete rows of data from between the points: “Enter Data” to “End of Data Entry” and the program would still do the correct calculation; this will be true for most of the excel programs. You can also delete rows of data entries without any problem in calculations. Note the data does not have to be ranked to do the calculation in Excel.

The Stats4U illustration of central tendency and all other statistics calculations are shown in the Stats4U tutorials.

Figure below shows the SPSS Method: Load ODE Data; Analyze -> Frequencies -> Statistics, Select Mean, Median, and Mode -> Select Display Frequency, Continue.

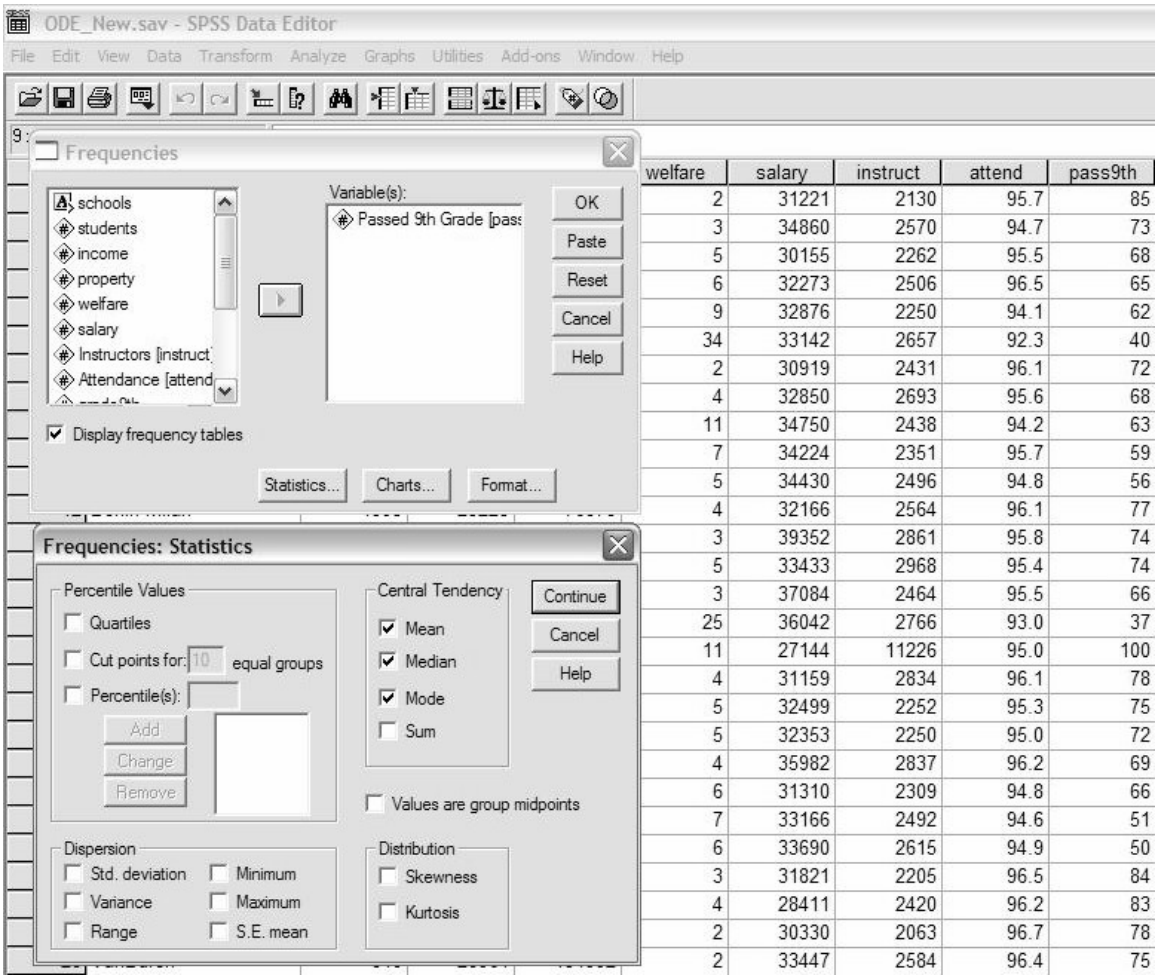


Figure 2.1.3 Central tendency measures of pass9th variable from ODE.

Tony: "How does one know which measure of central tendency is appropriate to describe a given distribution of scores?"

Rose: "Consider the mean as the 'balance point' of a distribution because the distance above the mean must have the same total as the distance below the mean." "There must be at least one score above and below the mean."

"The median, however, identifies the middle of the distribution in terms of the scores of the dataset." "The median is always located so that 50% of the scores are at or below it."

"The mode is often referred to as the '*most popular score'*." "It is the most common score among a group of score."

"In addition to deciding on which measure of central tendency to use based on the definitions or calculations of these three statistics, one can also exam the variable type or additional information about the data to make a judgment about which measure is most appropriate." "The decision matrix table below can be used as a guide for such decisions."

Table 2.1.2 *Selecting a Measure of Central Tendency*

Mean (preferred)	Median	Mode
1. When must include every score in distribution	1. When distribution has a few extreme values (outliers)	1. For scores that are nominal values (frequency)
2. Closely related to variance and standard deviation	2. When the distribution is skewed (not normal)	2. For discrete variables (e.g. number of cars per household)
3. When making inference about center of data	3. When there are scores that are undetermined or scores > or < scores in dataset	3. Locating the peak of a distribution
	4. For open-ended distributions (no lower or upper boundaries)	4. Easiest of central tendency measurement to calculate
	5. For scores that are ordinal values	

Distribution Shape and Central Tendency

The position of the mean, mode, and median of a distribution is related to the shape of the distribution. When mode = median = mode, we say that the distribution is symmetrical or normally distributed. If we have a number of distributions with the same mean, we might assume that these distributions are symmetrically distributed about that mean value if we don't have enough data to examine the actual data distributions; how well we make such assumption is based on other statistics that measures the variation of the data about the mean. Figure 2.1.4 shows a symmetrical distribution.

The skewness of a distribution is the extent to which it departs from symmetry. When most of the distribution of data is to the right or the mode > median > mode, we say that the data is negatively skewed (Figure 2.1.5b). When most of the distribution of data is to the left or the mean > median > mode, we say that the data is positively skewed (Figure 2.1.5a).

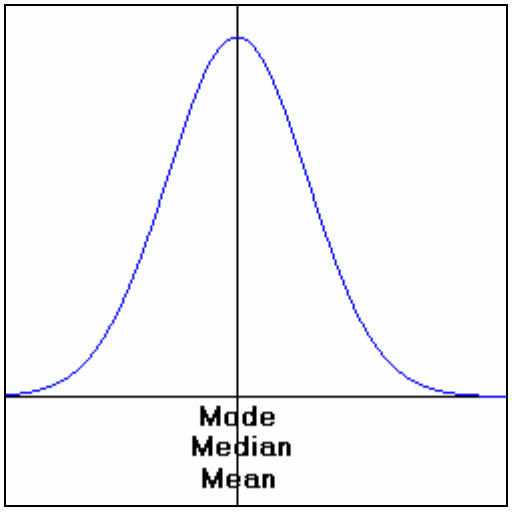
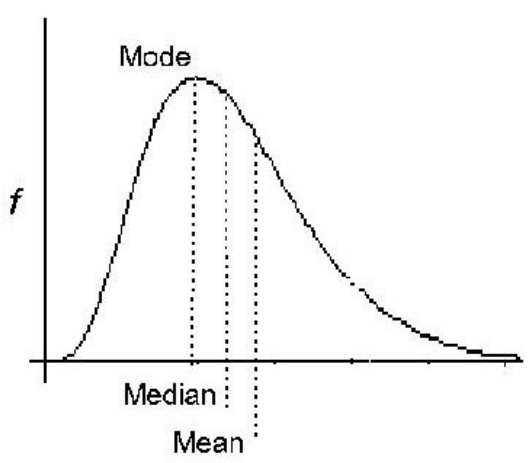
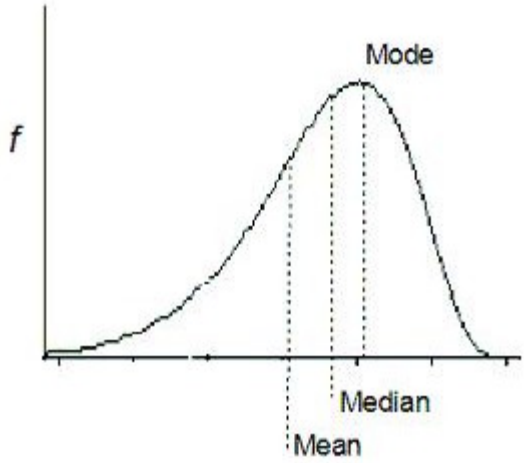


Figure 2.1. 4 Symmetrical Distribution



(a) Positively Skewed (tail to right)



(b) Negatively Skewed (tail to left)

Figure 2.1.5 Skewed Distributions