MODULE # 2

Statistics for the Social & Behavioral Sciences

Performing and Interpreting Basic Descriptive Statistics

A survey was conducted using a sample of 30 (thirty) students. Each student was asked to report the number of sex partners he/she had in the past month. The information collected is detailed below.

1	5	5	4	5	4	0	1	1	0	0	1	0	4	2	0	0	1	1	4	5	5	0	0	1	5	5	15	5	0

NOTE: In this exercise we will summarize the distribution of these data by calculating, interpreting, and reporting the statistical measures mentioned below. When we summarize, we provide the essentials. We will also represent the distribution using graphical formats (line graph, bar chart, and histogram).

By doing this, we have both a numerical and a pictorial representation of the data, and together this allows for greater clarity and more effective communication of the results. The numerical measures alone do not tell the entire story of the data, therefore, do not forget the graph (or picture of the data). We all know how effective a picture can be in conveying information.

Hence we will focus on the following:

- Measures of central tendency: Mean, Median, and Mode (CENTRE)
- Measures of Dispersion: Standard Deviation, Variance, and Range (SPREAD)
- Line Graph, Bar Chart, and Histogram (SHAPE)

Recall from the notes that in order to adequately describe a distribution we must comment on its CENTER, SHAPE, and SPREAD.

As you read these notes and other relevant material, you will observe that there are symbols that are used to represent various statistical measures. These differ for the sample (statistics) and the population (parameters).

Sample measures are often assigned Roman letters (e.g. M, s, $\bar{x}$), whereas the equivalent unknown values (parameters) in the population are represented by Greek letters (e.g. μ and σ) – see below.

NOTE: The Greek letter Σ (upper case S) indicates summation (or total) in statistics.

Measures	Sample (statistic)	Population (parameter)
Mean	M, $\bar{x}$	μ (mu)
Standard Deviation	S, SD	σ (sigma)

MPORTANT: The steps in the calculations are detailed and explained below. PLEASE PERFORM ALL CALCULATIONS TO TWO DECIMAL PLACES.

Column # 1	Column # 2	Column #3	Column # 4	Column # 5	Column # 6
These are the different values or observations (from the 30 listed above) – ranked from lowest to highest. Ordering the values is important especially for obtaining the MEDIAN (which is addressed below).	f represents the frequency or how often each x value (from column 1) occurred. For example, 9 students reported 0 sex partner. We are grouping the data so that it is better and meaningfully organized. Therefore, we do not have to list each of the 30 values separately.	This is the product of f and x, that is, multiplying each x value (from column 1) by the corresponding f value (from column 2). This helps us to get the sum of the 30 values in steps, which is more methodical and organized. This is the same as adding up the 30 values.	$\bar{x}$ (pronounced x bar) represents the mean of the sample , which as calculated below is 2.67 for these data. This column represents the deviation (or difference) of each x value (column1) from the mean ($\bar{x}$). For example , 0 – 2.67 = -2.67 hence the first entry here is: -2.67	This is the square of each value in column 4. This is done so as to get rid of the negative signs and allow for meaningful summation of the values. Hence the first entry here is: -2.67 x -2.67 = 7.13 (recall that the square of a negative number is a positive value).	In this column, each value from column # 5 is multiplied by its corresponding f value from column # 2. This is done so as to account for all the values in the distribution. Hence the first entry is 9 x 7.163 = 64.17

x	f	fx	$x - \bar{x}$ $(x - 2.67)$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
0 ←	9	0	-2.67	7.13	64.17
1	7	7	-1.67	2.79	19.53
2	1	2	-.67	.45	.45
4	4	16	1.33	1.77	7.08
5	8	40	2.33	5.43	43.44
15	1	15	12.33	152.03	152.03
	$\Sigma f = 30$	$\Sigma fx = 80$			$\Sigma = 286.70$

The arrow in the table above indicates the mode or modal value (which is explained below).

Mean ($\bar{x}$) = $\frac{\sum fx}{\sum f}$ = $\frac{80}{30}$ = 2.67 (approximately 3 partners)

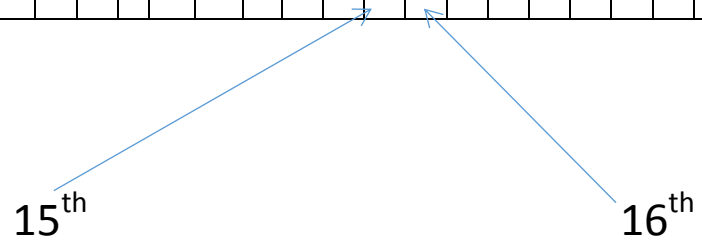
In general, this would be interpreted as follows: The typical number of sex partners reported by students in the sample is 2.67 (approximately 3, as these are discrete data)

Note: Although the mean value is correct, it does not represent the data. In fact, no one reported 3 sex partners. This resulted from the value 15 (an outlier) which is skewing the distribution to the right. Recall, that for skewed distributions, the preferred and most reliable measure of central tendency is the MEDIAN. Remember that because the mean is calculated using all the values in the distribution, it is sensitive to or affected by outliers. This is not the case for the MEDIAN.

The MEDIAN is derived as follows. Keep in mind that median refers to the middle of the distribution of the values, after they are placed in order. There are 30 values, hence there is not a single middle value or balance point, as 30 is an even number. Hence we have to find the two middle values and take their average (mean).

The values are ordered below. The two middle values are the 15th and 16th. To identify those count either from the left or right. From the left, the values are shown below.

Values (x)	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	4	4	4	4	5	5	5	5	5	5	5	5	15



The 15th value = 1

The 16th value = 1

They just happen to be the same for these data. This is not always the case.

Median = $\frac{1+1}{2} = \frac{2}{2} = 1$

A median of 1 indicates that at least 50% of the students reported 1 or no sex partner.

It can also be interpreted as:

At least 50% of the students reported 1 or more sex partners.

Indeed, in this case, the median (compared to the mean) better represents this distribution.

Note if the distribution consists of an odd number of values such as 29, then there will be an exact middle value. For 29 values, the median will be the 15th value in the distribution, after the values are placed in order.

The MODE or Modal Value is the value with the highest frequency, in other words, the most frequently occurring value.

First, organize your data as in the table above (see page 3).

Next, identify the highest frequency from column # 2 (see arrow in table)

Finally, as indicated by the arrow, find the corresponding x value, which is 0 in this case.

The mode or modal value is 0.

MEASURES OF DISPERSION (SPREAD, VARIABILITY OR DISTANCE)

We will now calculate measures of dispersion beginning with the standard deviation (SD). More specifically, we will calculate the standard deviation of the mean.

Practically speaking, the standard deviation of the mean is the average deviation of the values in the distribution from the mean. It indicates (on average) how the values vary or spread around the mean.

$$SD = \sqrt{\frac{\sum f\left(x - \bar{X}\right)^2}{n-1}} = \sqrt{\frac{286.70}{30-1}} = \sqrt{\frac{286.70}{29}} = \sqrt{9.88} = 3.14 \text{ (sex partners)}$$

Note, this value (9.88) is the Variance.

Hence, on average, the values in this distribution varied by 3.14 (approximately 3 sex partners) from the mean. Indeed, numbers must be interpreted in context, and given the values in this distribution, and what is generally known about the number of sex partners for this group, it seems reasonable to say that this value (3.14) represents a relatively high level of variability or dispersion.

Also, whenever, the SD is close to or higher than the mean, this generally suggests "high" variability (dispersion or spread) within the data.

This is to be expected here, especially given the presence of the value 15 (an outlier; that is, an extreme value that does not fit the body of the distribution)

VARIANCE is the square of the standard deviation, and is indicated above by the arrow (see page 7). In this case, the variance is 9.88. Because it is the square of the SD, the variance is always larger than the SD for the same set of data.

The variance is not a practical outcome measure; it has more theoretical and mathematical value. Also, given that the variance can be large, it can be difficult to interpret practically.

And strictly speaking, the variance is measured in squared units.

For example: If the SD for the distribution of heights = 5inches (ins)

Then the Variance = the square of SD = 5ins x 5ins = $25\ \text{ins}^2$

However, it is neither easy nor customary to think about height in terms of squared inches, hence variances are not easily and readily interpretable.

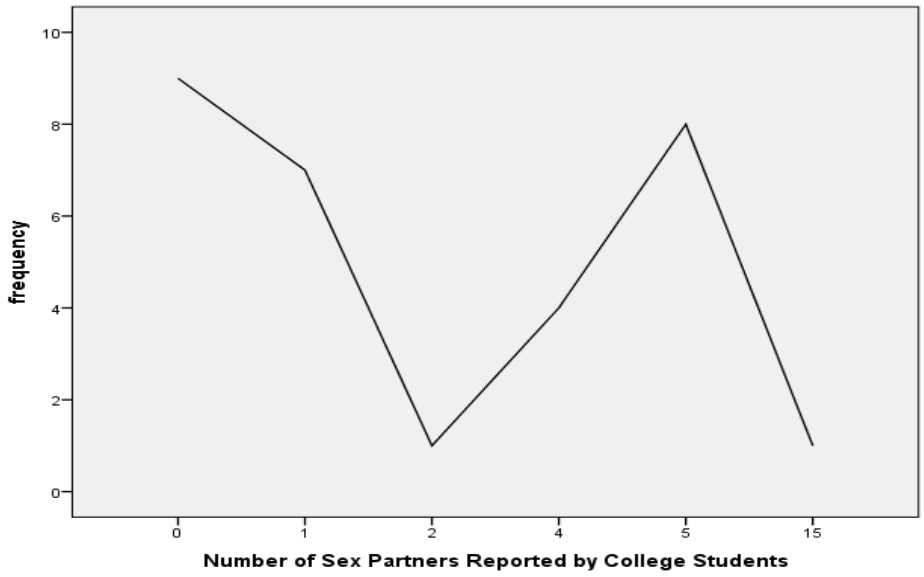
For this reason, the SD (standard deviation) is preferred as a measure of dispersion (or spread), and is generally required and reported.

RANGE – This is represents the difference between the highest and lowest values in the distribution. These values come from column #1 in the table above (see page 3).

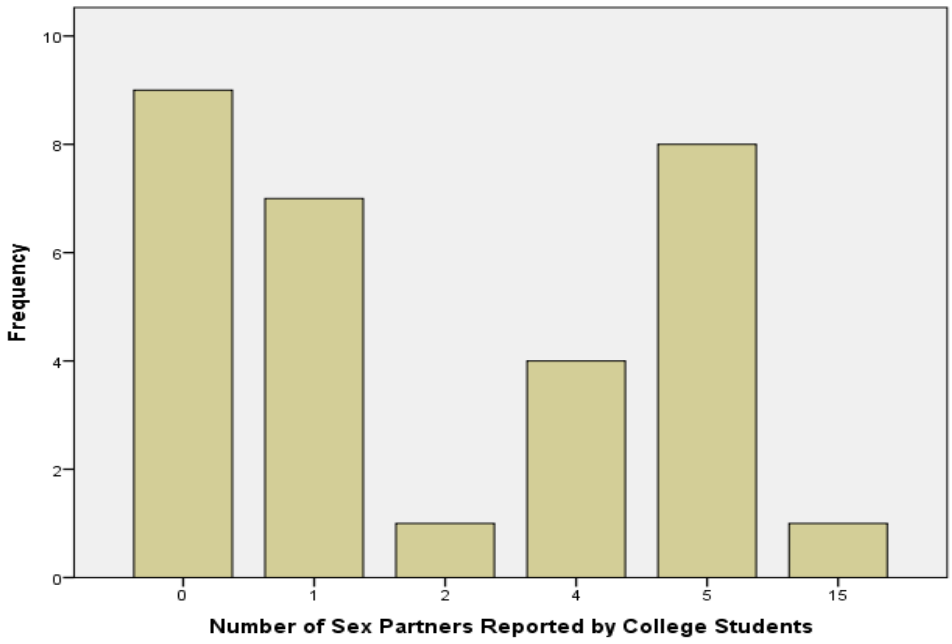
Range = 15 (highest value) – 0 (lowest value) = 15

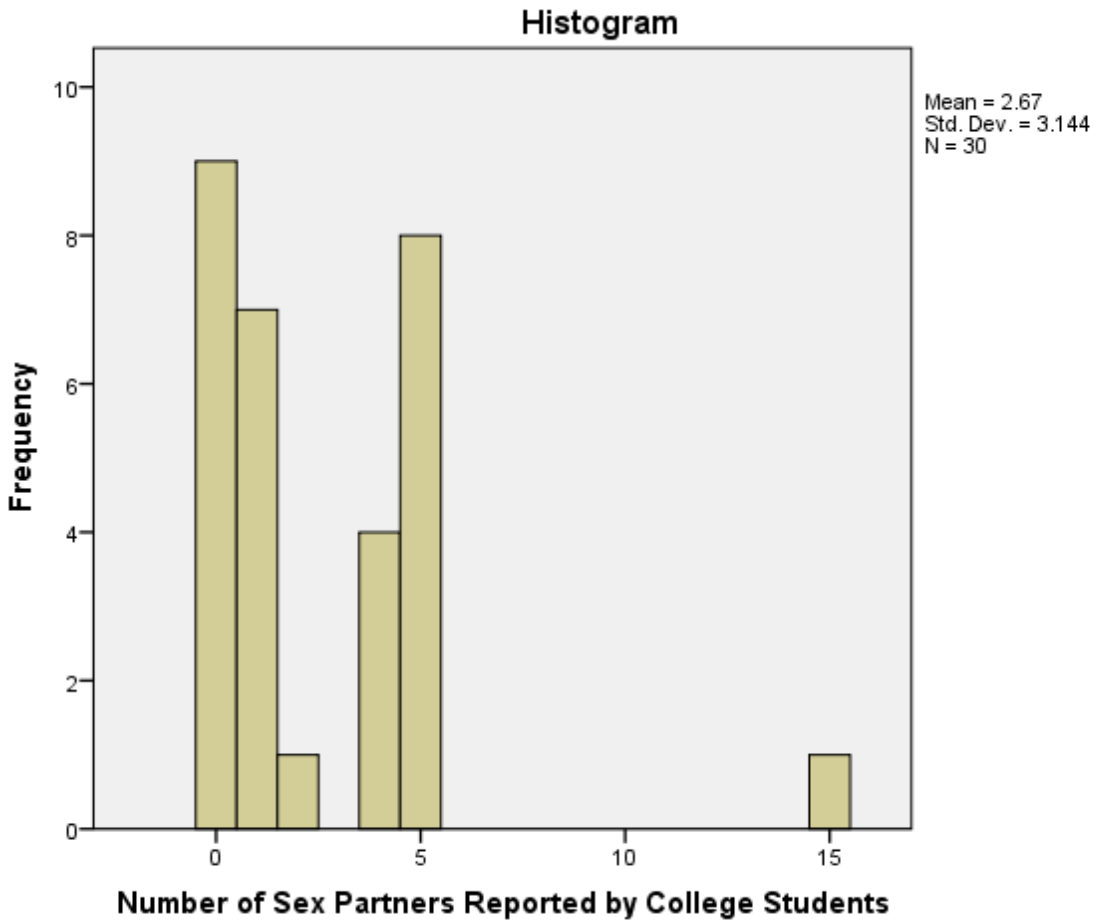
GRAPHICAL REPRESENTATIONS OF THE DATA

Line Graph



Bar Chart





Best practices in statistics recommend that we explore the data using different graphical formats, as each may provide a different (and useful) perspective on the data. In this case, it is clear that the histogram best identifies the outlier (15) by separating it from the rest (or body) of the data, and thereby, also indicating positive skewness (for these data).

The Box Plot (another graphical format, not shown here) is also useful for identifying outliers.

The Bar Chart allows us to see/identify meaningful subgroups within the distribution. For example, given that such data (number of sex partners) are generally used to help to determine the risk of acquiring sexually transmitted diseases (as well as psychosocial dysfunctions), practitioners (such as epidemiologists, public health specialists, and mental health practitioners) would be inclined to view this distribution as multimodal, as there are subgroups that tend to have different characteristics and profiles, and therefore, should be addressed/treated differently.

Accordingly, we can conclude that this distribution is multimodal, that is, it consists of three subgroups, those with 0, 1, and multiple sex partners.

The positive skewness observed can be accounted for by the value 15, which in this context, can be considered an outlier. It is recommended that this outlier (15) be removed and the data reanalyzed in order to obtain more reliable and representative descriptive statistics.

Okay, you have come to the end of the notes, and it's time to apply the knowledge and skills you have acquired, by attempting the assignment for this module. Please click on the assignment link and proceed.