Module 2

LEARNING ACTIVITY I

THE BASIC CONCEPTS OF STATISTICS

The basic concept of statistics and science is that of the variable. Very simply, a variable is a thing, characteristic, or phenomenon that can take different values; it can vary. For example, weight as a variable can be 10 lbs. or 10,000 lbs., sex has two values (usually), male and female. In contrast to the variable is the constant. A constant is a thing, characteristic, or phenomenon that has a fixed value. For example, pi (π) has a fixed value. We use the term observation to refer to any statement of value we make about a variable in an individual case. For example, on the variable of Empathy, John Jones scores 80 on the Empathy in Interviewing Instrument. The score of 80 is an observation.

When numbers or measurements are collected as a result of observations, we have data. The scores of all ADC clients on an alienation test would be data, or the head count of individuals preferring one type of cash payment over another would be another type of data.

The complete set of things (individual objects, etc.) that we wish to study or observe is known as a population or a universe. An example of population would be all the Public Welfare employees in the State of New York. A part or a subset of a population is known as a sample. So the Public Welfare workers in Chemung County are a sample of the Public Welfare workers in the State of New York.

Any characteristic of a population which we can measure is called a parameter. For example, the mean age of Public Welfare workers in New York (our population) would be a parameter. Parameters usually are represented by Greek letters. (The Greek letter μ (mu) is the symbol we use for the population mean). Any characteristic of a sample we term a statistic: usually italic letters are used to represent statistics (the letter X with a bar over the top, hence X bar is used to refer to the mean score of a sample.)

Suppose the State Commissioner of DSS is interested in finding out the I.Q. scores of unmarried pregnant teenage girls in New York State. In this statement we have one variable which is I.Q. and our population which is all of the unmarried pregnant teenage girls in New York State. To save money and time and to insure accuracy, this project is given to the department research unit. They choose by random means 10 counties in which they are going to test the I.Q. of all unmarried pregnant teenage girls. We now have a sample--the ten counties chosen at random are a subset of all the counties in New York State. They collect all the I.Q. scores of the individual sample members (observations) and transfer all of the observations onto IBM cards. Now we have data. They find that the average I.Q. of the sample group is 98 ($\bar{X}=98$). This we call a statistic (with a small "s"). From this statistic (the average I.Q. of the sample member)

they estimate that the population average will be 98 or $\mu=98$ (given three points either way). This estimation of the population value would be a parameter.

Now Try These:

Match up these terms with the following phrases:
A statistic, a parameter, data, variable, population, observation, sample.

_____ 1. Age of AFDC clients in NY State

_____ 2. Sample: Average age of AFDC client is 32

_____ 3. IBM deck of cards containing all of AFDC clients ages

_____ 4. Population: Average age of AFDC clients in NY State is 32

_____ 5. All AFDC clients in NY State

_____ 6. Age of AFDC client #46 in Albany County is 28

_____ 7. 18 counties chosen randomly to estimate average age of AFDC clients in NY State

Module 2

LEARNING ACTIVITY 2

IDENTIFYING STATISTICAL SYMBOLS AND OPERATIONS

The symbols we use to stand for variables and scores are usually X and Y (capitals). If X stands for height and Y for weight we may wish to see if there is any association. Or X may stand for frustration and Y for aggression and we may want to see if there is any casual relationship between the two (traditionally X stands for the independent variable and Y stands for the dependent variable). A subscript is a symbol, either a letter or number placed slightly below and to the right of the variable symbol: X_1 or X_2 would designate specific individuals. For example, if we were to correlate the height and weight of 5 grade school children, we would use X for height and Y for weight.

	VARIABLES			
	X	(Height)	Y	(Weight)
Individuals	X_1	73	Y_1	120
	X_2	64	Y_2	115
	X_3	70	Y_3	100
	X_4	71	Y_4	98
	X_5	72	Y_5	110

Another class of statistical symbols is called operators. We know these symbols from grade school and high school math. They tell us to add (+), subtract (-), multiply (X), or divide (÷), or they may tell us to take the square root of ($\surd$) or to square (2^2).

An operator that is peculiar to statistics is the Greek capital letter sigma (Σ) which means to sum up or to add up what follows the sign. For example, we could tell someone to add up the five scores on variable X as follows:

$$X_1 + X_2 + X_3 + X_4 + X_5$$

or we could shorten it to:

$$X_1 + X_2 + \ldots X_5$$

(the three dots mean "and so on")

or we could say:

$$\sum_{i=2}^{5} Xi$$

Here the sigma instructs you to add up everything that follows Xi starting with the case below the sigma (i=1) and ending up with the case specified above the sigma which is 5 (we usually use the subscript ith (i) to refer to an unspecified individual).

If we see this:

$$\sum_{i=3}^{7} Xi$$

It would instruct us to sum up the observations starting with the third observation and continuing until the seventh.

We might see this:

$$\sum_{i=1}^{N} Xi$$

which means to sum up all the observations from 1 through the Nth or last (capital N stands for the population number and the small n stands for the sample number).

Frequently we just see

$$\Sigma Xi \text{ or } \Sigma X$$

When we see these we know we are to sum up all the cases under consideration.

The symbols for constants are most commonly the numbers 1 2 3 4 ... Their value remains constant no matter what the problem may be.

The symbols for parameters will be Greek letters. The two we will encounter are μ and σ (mu and small sigma) μ= parameter mean and σ= population standard deviation (a value we will discuss later).

The symbols for statistics are usually letters. For example, $\bar{X}$ as stated before is a sample mean.

The last set of symbols we will discuss are called connectives. We use them to connect parts of equations.

=	Equal sign
<	Less than
>	More than
$\leqslant$	Less than or equal to
$\geqslant$	More than or equal to
$\neq$	Not equal to

These symbols connect statistics or parameters with the operations you wish to make on your raw data.

A formula for the population mean:

$$\mu = \frac{\Sigma\ X1}{N}$$

is broken down as follows:

μ	(mu) is the parameter
=	Connects the parameter with the operations and says it is equal to the result
X_1	Stands for the unspecified scores
Σ	Tells us to sum up all the scores from first to last
—	Tells us to divide by
N	Is the symbol for the number of cases in the population

Now Try These:

1. Match the following statistical symbols with the correct definition.

X	(a)_____	Sample mean
Σ	(b)_____	Variable or score symbol
μ	(c)_____	Summation sign
$\bar{X}$	(d)_____	Not equal to
$\neq$	(e)_____	Population mean

2. Write the notation for summing up 12 observations where the N=12.

MODULE 4

CENTRAL TENDENCY

Rationale:

One of the keystones of descriptive statistics is the calculation of the appropriate measure of central tendency. With this data the Staff Developer will have a single score that will give him or her a picture of where the average scores are in relationship to the spread of the distribution.

Competency/Terminal objective:

Given formula and data of the appropriate level the Staff Developer will be able to compute three measures of central tendency.

Enabling objective:

The Staff Developer given level of data and computational formula, will be able to:

1. match level of data and measure of central tendency
2. find the mode of a distribution
3. find the median score of a distribution
4. calculate the mean of a distribution.

MODULE 4

LEARNING ACTIVITY I

"MEASURES OF CENTRAL TENDENCY"

INTRODUCTION

Scores on tests have little meaning by themselves. A score of 23 by Joan Pitts on a training post-test tells us very little. Now the score would have more meaning if we knew where the typical trainee scored, then we could compare Joan's score with the average or typical trainee. The point we wish to make is that scores or figures to have any useful purpose must be related to some statistical criterion.

One such set of statistical criteria we call measures of central tendency. We typically group under this heading the mode, median, and mean.

The mode refers to the most frequent score in a distribution; the median refers to the middle score of the distribution; and the mean refers to the arithmetic average.

If we know what the measures of central tendency are for a group of scores we can do several things:

1. We can show where the "typical" score lies
2. We can compare other scores in terms of this typical score
3. We can compare scores on a pre and post basis
4. We can compare the mean achievement of two or more groups
5. We can compare the means of two or more groups on a pre/post basis.

The most powerful measure of central tendency is usually the mean, but in terms of the quality of our data, we can use the mean only when our data is of interval level.

The following table lists the appropriate level of measurement for each measure of central tendency.

Level of Measurement	Measure of Central Tendency
Nominal	Mode
Ordinal	Median (Mode)
Interval	Mean (Mode-Median)
Ratio	Mean (Mode-Median)

Appendix I lists the characteristics and use patterns for the measures of central tendency.

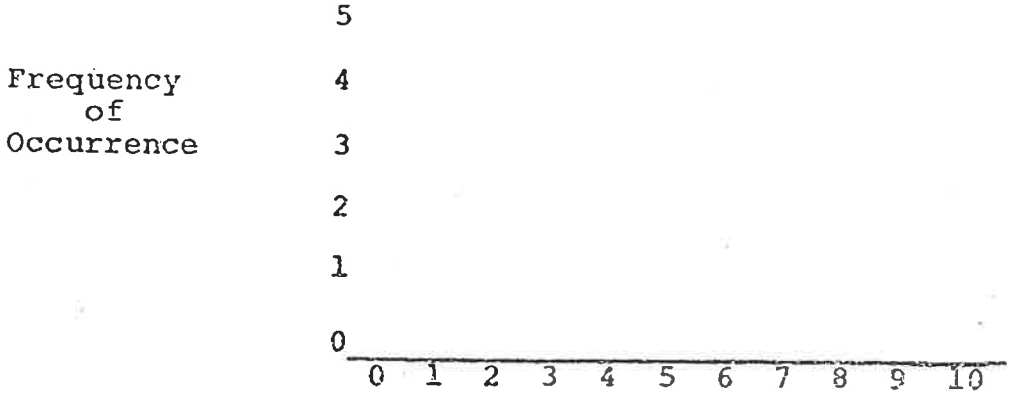
The Mode

Since we define mode as the score which occurs most often in a distribution of scores, try you hand at identifying the mode of the following distribution of scores

3 4 4 5 5 5 6 6 7 7 8 8 9

Mode_____1.

Now chart out the distribution on the matrix below



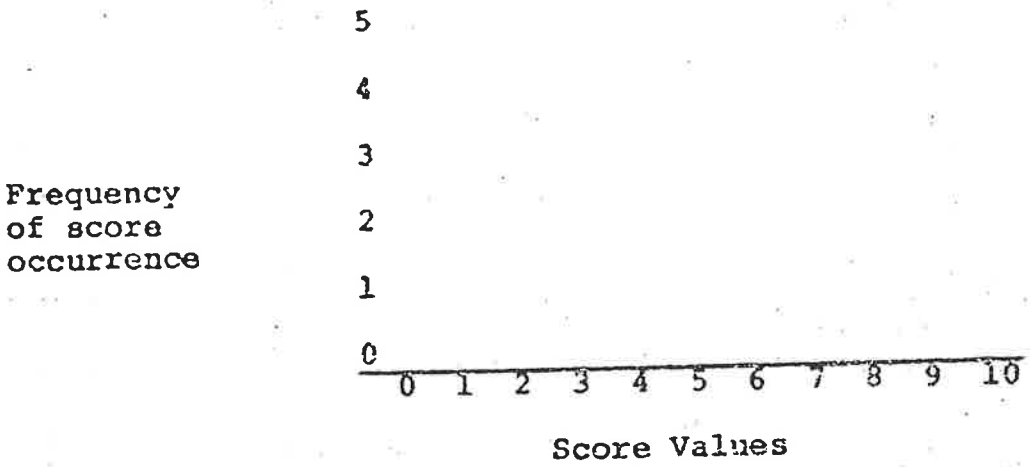
This distribution we term Unimodal.

Let's try our hand once more.

3 4 5 5 6 6 6 6 7 8 9 9 9 9 10

Mode_____2.

Chart out distribution on the matrix below



This distribution we term Bimodal.

When distributions of scores are bimodal it gives us hints at how well our material might be coming across. If the training population for this course was comprised of Ph.D.'s in Mathematics and Public Welfare Staff Development Coordinators, the distribution of scores on a pre-test involving correllation calculations might well be bimodal. We could then divide up the trainees into advanced and beginning levels. Of course Staff Development Coordinators would all be in the advanced group.

The Median

Since the median is the middle point of our distribution of scores, our first step is to arrange the scores in order. If there is an odd number of scores the median is the middle score.

4

8

12

16

18

24

26

Here 16 is our median because there are three scores above it and three scores below it.

Try to find the median for these scores.

6 3 8 20 5 9

Since we have an even number the median will fall between the 6 and the 8. The median is the point halfway between the two adjacent scores.

Here are some practice problems.

1. 1 7 5 3 2 Median = _____

2. 6 4 9 7 Median = _____

The computation of the median can get more complex especially when there is a repetition of the same score near the middle of the distribution. For our purposes we will not go further into this but the reader may want to look up the procedure in any of the statistics books listed in the references.

The Mean

The mean or arithmetic average can be put into formula terms as follows:

$$\text{Mean} = \frac{\text{Sum of scores}}{N}$$

Sometimes this appears as

$$\bar{X} = \frac{\xi X}{n} \quad \text{Sample Mean}$$

or

$$\mu = \frac{\xi X}{N} \quad \text{Population Mean}$$

Remember the Capital Sigma (ξ) means "to sum up" or "the sum of". In terms of operations it tells us to sum up the Xs or the Ys and then divide by the number of scores to get the mean.

Calculate the mean, mode, and median of the following scores:

X
23
21
18
17
15
14
14
12
9
7
ΣX =

Calculate the Mean_____

Median_____

Mode_____

Now let's use your calculator to find the mean of this distribution. Turn to page 1-20 of your handbook.

46
42
39
37
37
35
33
30
29
25
18
ΣX

$\bar{X}$ = ____

MODULE 4

Appendix I

Measure of Central Tendency	Features	For Use
MODE	a) A nominal statistic, but can be used for all levels of data	a) When data is nominal or bimodal
	b) The most frequently occurring value	b) When quick knowledge of distribution is desired
	c) Some data distributions may have more than one mode	
	d) Poor estimate of population values	
	e) Incapable of mathematical manipulation	
MEDIAN	a) An ordinal statistic but can be used with interval and ratio level data	a) When ordinal data is used
	b) Can yield a rank or position average	b) When you wish to pinpoint the middle score
	c) Is not sensitive to extreme values	c) In cases where extreme scores would distort a mean
	d) Capable of few mathematical manipulation	
	e) More stable than mode but less stable than mean	
MEAN	a) An interval or ratio level statistic	a) With interval/ratio level data (other use in "fear and trembling")
	b) An arithmetic average	b) When all scores should weigh equally
	c) Is affected by extreme score	c) When you intent to do further statistical computation
	d) Capable of many mathematical manipulations	
	e) It is the most stable of measures of central tendency	

MODULE 4

Answers to Module 4 - Learning Activity #1

Mode 5 1.

Mode 6 & 9 2.

Median 1. 3

Median 2. 6.5

Mean 15

Median 14.5

Mode 14

Mean 33.73

Module 5: Measures of Variability

Rationale:

The mean score (x) and the standard deviation are indispensable for summarizing distribution of evaluation scores. The mean gives us a one number summary of a distribution. It gives us the balance point;as such it is a type of average.

The standard deviation gives us a one number summary of how the scores are spread. It is useful because it is the most stable measure of variability (extreme scores do not bias it unduly) and it serves as a basis for further inferential statistical procedures.

It is necessary to compute a measure of variability if we wish to see how training scores are spread out around the mean. From this spread we can see how they do or do not approximate the normal curve.

Competency/Terminal Objective:

Given appropriate level of data and formula,the Staff Developer will be able to compute the appropriate measure of variability and relate this variability to the normal curve.

Enabling Objectives:

Given appropriate level of data and formula,the Staff Developer will be able to:

1. Compute the range scores
2. Compute the standard deviation
3. List the characteristics of the normal curve
4. Fit the concepts of standard deviation and the normal curve together.

Learning Activities:

Read attached pages:

1. Computing Measures of Variability
2. The Normal Curve or Gaussian Distribution

Module 5

LEARNING ACTIVITY I

COMPUTING MEASURES OF VARIABILITY

Besides knowing the central tendency of a distribution of training scores it is necessary to know how the scores are spread out or how they vary about the central value of a distribution. The purpose in computing these measures of dispersion is to:

1. first find the amount of spread;
2. secondly we can compare this spread with other distributions;
3. and we can compare the amount of spread with the same group on a pre and post basis.

The simplest measure of variability is the range, and it is simply the distance between the highest and lowest score.

Pre Test Training Scores		Post Scores	
X_1	95	X_1	98
X_2	80	X_2	95
X_3	70	X_3	90
X_4	50	X_4	92
X_5	40	X_5	93
X_6	35	X_6	90

We see the range on the pre-test scores is 60 points, from 95 to 35, while on the post the range is 8 points (98 to 90). We can see that training has cut down on the range of the scores.

The most valuable measure of variability, however, is the standard deviation (which is appropriate for interval level data, but this assumption is often violated). The standard deviation is based (like the mean) on all scores and is a must for calculating many inferential statistics. It is also a standard benchmark for use with the normal curve.

The formula for the standard deviation is

a) Population

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{(Little x)} \qquad [\sigma = \text{small sigma}]$$

b) Sample

$$SD = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{(Little x)}$$

When sample of less than 30 is used, we use N-1 in the denominator.

We calculate by setting our data up in a matrix.

Pre or Post Score

X	$x (X - \bar{X})$	x^2
23	(23-15) = 8	64
21	(21-15) = 6	36
18	(18-15) = 3	9
17	(17-15) = 2	4
15	(15-15) = 0	0
14	(14-15) = -1	1
14	(14-15) = -1	1
12	(12-15) = -3	9
9	(9-15) = -6	36
7	(7-15) = -8	64

$\Sigma X = 150$ $\Sigma X^2 = 224$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{150}{10} = 15$$

The steps in computing are as follows:

1. Sum the X's in the first column which is 150
2. Calculate the mean score which is 15
3. Subract mean from all the individual large X scores i.e. 23 - 15 = 8. The 8 is the score for little x (which is a mean deviation score)
4. Square the little x's for the last column (little x squared = x^2)
5. Sum the squares of the little x's for a total of $\Sigma x^2 = 224$.

Now we have the data necessary to plug into our standard deviation formula

$$\sigma = \sqrt{\frac{\Sigma x^2}{N-1}}$$

We use N-1 because our N is less than thirty.

$$\sigma = \sqrt{\frac{224}{10-1}} = \sqrt{\frac{224}{9}} = \sqrt{24.889} = 4.988$$

Range = 23-7 = 16

Now Try These:

1. Now try your hand at finding the mean, standard deviation and range of this distribution by hand.

X	x	x^2
10		
8		
6		
4		
2		

Mean =_____

SD =_____

Range=_____

2. Now we will try to compute the same statistics with our calculator.

X	x	x^2
26		
25		
24		
21		
20		
18		
17		
14		
14		
12		
11		
10		
8		
4		

Mean =_____

SD =_____

Range=_____

Module 5

LEARNING ACTIVITY 2

THE NORMAL CURVE OR GAUSSIAN DISTRIBUTION

The normal curve is not a distribution of actual data but a theoretical distribution derived from mathematical equations (although the original curve was built on data concerning errors). See page 6-5 of Calculator Decision-Making Sourcebook.

The normal curve isone of the most frequently used distributions we have in terms of describing the distribution of scores for populationsof at least one hundred. It is characterized by:

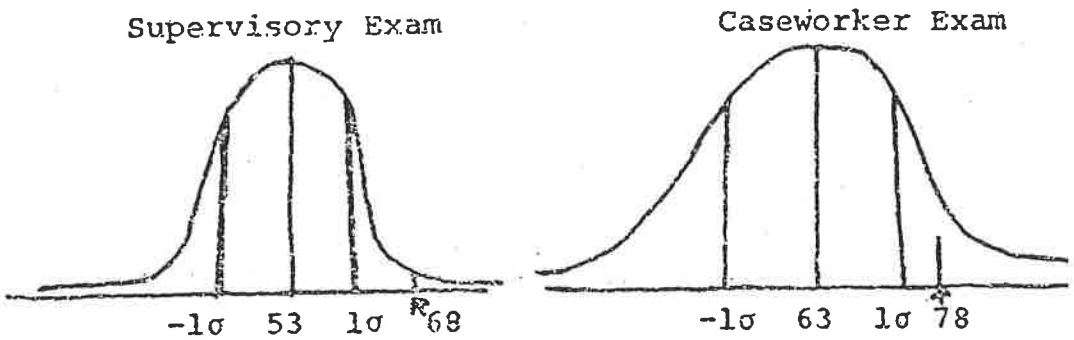
1. a smooth bell shape;
2. scores being distributed symmetrically about the mean;
3. the three measures of central tendency being identical;
4. the scores being distributed in a symmetrical pattern at various standard deviations from the mean.

See page 6-7 for a diagram of a normal curve. We can see from this figure that the standard deviation is an important concept for understanding the normal curve. If we make the assumption that our distribution is normally distributed then 68.26 percent of our cases will fall between plus or minus 1 standard deviation of the mean. In terms of mastery learning this concept is important for it states that scores on aptitude tests are normally distributed and highly correlated with achievement scores. So if we know our trainees' aptitude scores we can predict what the achievement scores will be. For mastery learning our efforts are to gear our training to move the achievement score mean over to the vicinity of the + 2 standard deviation.

The standard deviation is also a good estimate of our error. The more compactly the scores are distributed around the mean, the less our error will be in prediction. The larger our standard deviation the larger our error in prediction. We will compute a standard error in the module on correlation and regression.

The standard deviation helps us in describing individual differences in a group. Suppose you take two civil service tests, one for a caseworker position and one for a supervisory position in I.M. You receive a score of 78 on the caseworker exam and a score of 68 on the supervisory exam. On what exam did you do better? While the mean score on the supervisory exam was 53, the mean score on the caseworker exam was 63. This might indicate that it was easier to get a point on the caseworker exam than on the supervisory exam. In either case you are about fifteen points above the mean. Does that mean you did equally well on both tests? For that answer we need to calculate the standard deviation and we find that for the supervisory exam you are 2.6 standard deviation above the mean and for the caseworker exam

you are 1.2 standard deviation above the exam mean.



In conclusion then, relative to this, you did much better on the supervisory exam than you did on the caseworker exam.