LECTURE 4: INTRODUCTION TO DESCRIPTIVE STATISTICS

"***Descriptive statistics*** can help to provide a simple summary or overview of the data, thus allowing researchers to gain a better overall understanding of the data set" (Mackey & Gass, 2005, p. 292)[1]. Because raw data are not in and of themselves revealing, they must be organized and described in order to be informative. This lecture will present an overview of three different types of descriptive statistics: (1) measures of frequency; (2) measures of central tendency; and (3) measures of variability or dispersion.

1) Measures of frequency

Measures of frequency are used to indicate how often a particular behavior or phenomenon occurs. For example, in second language studies, researchers might be interested in tallying how often learners make errors in forming the past tense, or how often they engage in a particular classroom behavior. This is usually done through a table format where the researcher presents frequencies and percentages. The following table, for example, presents the frequency of phonemic pronunciation errors made by beginner EFL students in a spontaneous speech task.

Participants	Gender	Vowel errors	Consonant errors	Totals
Student 1	Male	17	3	20
Student 2	Female	15	8	23
Student 3	Female	25	9	34
Student 4	Male	18	6	24
Student 5	Female	14	2	16
Totals *(%)*	*n = 5 (2 male, 3 females)*	**89** *(76.07%)*	**28** *(23.93%)*	**117** *(100%)*

Measures of frequency can also be visualized through pie charts, bar graphs, and line charts.

2) Measures of central tendency

Second language researchers often use one or more measures of central tendency to provide precise quantitative information about the typical behavior of learners with respect to a particular phenomenon. There are three commonly used measures of central tendency, namely: 1) the *mode*, 2) the *median*, and 3) the *mean*.

A) The mode: is the most frequent score obtained by a particular group of learners. For example, if the ESL proficiency test scores recorded for a group of students were 92, 78, 92, 74, 89, and 80, the mode would be 92 because two students in this sample obtained that score.

B) The median: is the score at the center of the distribution—that is, the score that splits the group in half. For example, in our series of ESL proficiency test scores (92, 78, 92, 74, 89, and 80), we would find the median by first ordering the scores (74, 78, 80, 89, 92, 92) and then finding the score at the center. Since we have an even number of scores in this case (i.e., six), we would take the midpoint between the two middle scores (80 and 89), or 84.5.

C) The mean (average): is derived from adding up all the numbers (*Sum*) and dividing by the total number of observations (e.g., participants). The following table presents the formulas needed for calculating the mean.

	FORMULA	EXPLANATION
Population mean (μ)	$\mu = \frac{(\Sigma\, xi)}{N}$	$Population\ Mean\ = \frac{Sum}{Population\ size}$
Sample mean (x̄)	$\bar{x} = \frac{(\Sigma\, xi)}{n}$	$Sample\ Mean\ = \frac{Sum}{Sample\ size}$

[1] Mackey, A., & Gass, S.M. (2015). Second Language Research: Methodology and Design (2nd ed.). Routledge. Retrieved from https://doi.org/10.4324/9781315750606

For our scores (92, 78, 92, 74, 89, and 80), the mean would be the sum of all scores divided by the number of observations, or (505 /6 =) 84.17. It should be kept in mind that even though the mean is commonly used, it is sensitive to extreme scores especially if the number of participants is small.

3) Measures of spread (dispersion):

"Measures of dispersion describe variability of the numeral data away from the central tendency" (Phakiti, 2010, p. 44)[2]. "Measures of dispersion [particularly standard deviation] can serve as a quality control for measures of central tendency; the smaller the standard deviation, the better the mean captures the behavior of the sample." (Mackey & Gass, 2015, p. 303)[1]. In statistics, there are two main measures of dispersion, the *variance* and *standard deviation*.

A) The variance: is the expectation of the squared deviation of a random variable from its population mean or sample mean. The variance is a measure of how far a set of numbers is spread out from their average value. The more spread the data, the larger the variance is in relation to the mean. It is calculated by dividing the sum of squared deviations from the mean by the population or sample size.

	FORMULA	EXPLANATION
Population variance (σ2)	$\sigma^2 = \frac{\Sigma\ (xi - \mu)^2}{N}$	$Population\ Vriance\ = \frac{Sum\ of\ (observation\ value - population\ mean)^{Squared}}{Population\ size}$
Sample variance (S²)	$S^2 = \frac{\Sigma\ (xi - \bar{x})^2}{n-1}$	$Sample\ Vriance\ = \frac{Sum\ of\ (observation\ value - sample\ mean)^{Squared}}{Sample\ size - 1}$

B) The Standard deviation (*SD*): "is the average point from the mean which indicates on average how much the individual scores spread around the mean." (Phakiti, 2010, p. 44)[2]. The standard deviation is "high if the sample is heterogeneous and contains extreme scores, whereas [it is] low in a homogeneous sample with all the scores clustered around the mean." (Dörnyei, 2007, p. 214)[3]. The standard deviation is obtained by calculating the square root (√) of the variance.

	FORMULA	EXPLANATION
Population standard deviation (σ)	$\sigma = \sqrt{\frac{\Sigma\ (xi - \mu)^2}{N}}$	$Population\ SD\ = \sqrt{\frac{Sum\ of\ (observation\ value - population\ mean)^{Squared}}{Population\ size}}$
Sample standard deviation (S)	$S = \sqrt{\frac{\Sigma\ (xi - \bar{x})^2}{n-1}}$	$Sample\ SD\ = \sqrt{\frac{Sum\ of\ (observation\ value - sample\ mean)^{Squared}}{Sample\ size - 1}}$

LET'S PRACTICE

Calculate the sample variance and standard deviation for the ESL proficiency scores (i.e., 92, 78, 92, 74, 89, 80).

$S^2 =$ _____ $S^2 =$ _____

$S^2 =$ _____ $S^2 =$ _____ $S^2 =$

$S = \sqrt{S^2}$ $S = \sqrt{\ }$ $S =$

[2] Phakiti, A. (2010). Analyzing quantitative data. In B. Paltridge & A. Phakiti (Eds.), *Continuum Companion to Research Methods in Applied Linguistics* (pp. 39-49). London: Continuum.

[3] Dörnyei, Zoltán. (2007). Research methods in applied linguistics. Oxford university press.