**Data Analysis (Quantitative data)**

1. **Need for Analysis of Data or Treatment of Data**

After administering and scoring research tools scripts, data collected and organized. The collected data are known as ‘raw data. The raw data are meaningless unless certain statistical treatment is given to them. Analysis of data means to make the raw data meaningful or to draw some results from the data after the proper treatment. The ‘hypotheses’ are tested with the help of analysis data so to obtain some significant results.

Thus, the analysis of data serves the following main functions:

1. To make the raw data meaningful

2. To test the hypothesis

3. To obtain the significant results

4. To draw some inferences or make generalization

5. To estimate parameters

**2. Kinds of Data**

* **Continuous Data:** the data that can take values between whole numbers such as time, temperature, and length; someone can be 36.33 years old.
* **Discrete Data:** the data that are meaningful as whole numbers, you cannot have 2.4 children.

**3. Levels of Measurement**

We can aspire to four different levels of measurement – Nominal, Ordinal, Interval and Ratio–and these are listed from the weakest level (nominal) to the strongest level (ratio).

* **Nominal data** give the least information, recording merely the name of the group category to which an individual or item belongs.
* **Ordinal data** indicate the rank order to which items are placed but give no indication of distance between ranks. The fact that a child may be the second child in a family reveals nothing about distance in age from the older sibling. Rating scales which use classifications produce ordinal data.
* **Interval data** indicate the rank order to which measures are placed but have equal intervals or equal distances between each of the measures on the scale. However, with interval data there is no absolute zero point, and so it is not possible to divide or to multiply one score by another. Measurement of temperature are an example of interval data, where the temperature of 40℃ cannot be divided by a temperature of 20℃ to claim that 40℃ represents twice as much heat as 20℃. There is what is called a zero, at 0℃, but this is merely a point of reference against which other temperature measures can be set.
* **Ratio** where we have ratio data, we can draw conclusions about the relative size or worth of the data. For example, a person’s income, regardless of the units in which it is measured, can be expressed as a ratio of someone else’s. So, for example, we can say that people earning $30.000 have twice as much as people earning $15.000.

**4. Analysis of Data**

Analysis of data involves breaking down existing complex factors into simpler parts and putting the parts together in new arrangements for the purpose of interpretation.

**4.1** **Statistical Analysis of Data**

Statistics is the body of mathematical techniques or processes for gathering, describing, orgainising, and interpreting numerical data. Since research often yields such quantitative data, statistics is a basic tool of measurement and research. Research in education may deal with two types of statistical data application:

* Descriptive Statistical Analysis
* Inferential Statistical Analysis

**4.1.1 Descriptive Statistical Analysis**

Descriptive statistical analysis is concerned with numerical description of a particular group observed and any similarity to those outside the group cannot be taken for granted. The data describe one group and that one group only. Much simple educational research involves descriptive statistics and provides valuable information about the nature of a particular group or class.

Most commonly used methods of analysis data statistically are:

1. Calculating frequency distribution usually in percentages of items under study.

2. Testing data for normality of distribution, skewness and kurtosis.

* Where data samples are large enough, acceptable deviations from normality for distributions can be found by looking at scores for skewness and excess kurtosis.
* Skewness is a measure of how much a distribution’s asymmetry deviates around the mean.
* If the skewness value is zero this implies that the distribution is symmetric about the mean.
* Data that is positively skewed tends to cluster to the left and the curve has a long tail to the right.
* The opposite is the case when the data is negatively skewed tending to cluster to the right.
* Kurtosis is a measure of flatness or steepness.
* Excess kurtosis = 0 for a normal distribution.
* Positive excess kurtosis (>0) implies the distribution has heavier tails and a sharper peak.
* Negative excess kurtosis (<0) implies lighter tails and a flatter peak.

3. Calculating percentiles and percentile ranks.

* A percentile is a value below which a certain percentage of observation lie.

4. Calculating measures of central tendency-mean, median and mode.

* The mean is simply the average of all the data.
* The median is the middle value of the ordered data, i.e. when the data are arranged from the smallest up to the largest, or vice versa.
* The mode is the value that is most common in the data set.

5. Calculating measures of standard deviation.

* It is a measure how much data scatter (disperse) around the mean. In other words, standard deviation measures the spread of a group of data, such as scores, by examining the distance of the individual scores from the mean.

While analyzing their data investigations usually make use of as many of the above simple statistical devices as necessary for the purpose of their study.

**4.1.2 Inferential Statistical Analysis**

Inferential statistics has two aspects

* Hypothesis testing
* Population parameters estimation from the sample

The primary purpose of research is to discover principles that have universal application. However, to study a whole population in order to arrive at generalization would be impracticable if not impossible. A measured value based upon sample data is statistic. A population value estimated from a statistic is a parameter. A sample is a small proportion of a population selected for analysis. By observing the sample, certain inferences may be made about the population. Samples are not selected haphazardly, but are chosen in a deliberate way so that the influence of chance or probability can be estimated.

**4.1.2.1 Types of Errors**

When researchers are drawing conclusions from a sample, there is the danger of two different types of incorrect conclusions being drawn from the evidence available. The null hypothesis (H0) may be rejected when it is true (Type 1/*α* known asthe *significance level*) or the null hypothesis may not be rejected when in fact it is false (Type 2/*β*). Whenever a hypothesis is being tested, the probability that either of these errors will occur can be calculated.

*The observed significance level* or *P value (the probability value)* is normally used to indicate the exact point at which H0 is either rejected or not rejected. Again, the decision of to reject or not to reject H0 may be made by the researchers on the basis of the chosen value for *α.* If the P value is less than *α,* then H0 is rejected.

In statistical tests, the most common cut-off point for statistical significance is at 5% (0.05) level P value. This implies that if you get a result that has less than 5% significance, the effect you are looking for exists in your data. Greater than this 5% significance then the effect you are looking for cannot be detected. The 5% P value significant level means that you are 95% (0.95) confident in whatever question you are asking is true. This probability is based on an assumption that only 5% of the time you would be making an error by assuming that this is true.

For example, if *a* has been set at 0.05, and *P* is calculated to be 0.15, then H0 would not be rejected.

**4.1.2.2 Consideration for Statistical Analysis**

The major types of tests are employed for analyzing data so as to interpret the results. There are:

1. Parametric statistics or tests: they assume that the whole population that the sample data has drawn from is normally distributed.

2. Non-parametric statistics or tests: they are used if the data is not normally distributed.

**a. Considerations for Parametric Statistics**

This type of statistical analysis may be employed effectively in the following conditions:

* Probability or representative sample has been employed in the investigation.
* Variables of the study can be qualified at interval scale
* Specific assumptions are fulfilled. The obtained data are normally distributed or not free distribution.
* The population of the study has been clearly defined.
* Objectives of the research study

Under this approach the following statistical techniques are employed :

1. To study the descriptive relationship of two or more variables:

a1 – Pearson’s product moment method of correlation (two variables)

a2 – Multiple correlation (more than two variables)

a3 – Partial correlation (more than two variables)

a4 – Factor analysis-extracting factors or estimating psychological or factorial validity of tests.

1. To analyse the functional relationship of the variables:

b1 – Main effect of two treatments ‘t’ test

b2 – Main effect of more than two treatments F-test

b3 – Interaction effect of two or more variables-Two or more ways analysis variance techniques

b4 – Gain or loss of more than two treatments-Analysis of covariance and correlated ‘t’ test.

| **Statistical Technique** | **Example in Language Education** |
| --- | --- |
| **Pearson’s Correlation** | Relationship between students’ **grammar knowledge** scores and their **writing proficiency** scores (continuous interval data). |
| **Multiple Correlation** | Relationship among students’ **vocabulary knowledge**, **reading habits**, and **reading comprehension performance** (three continuous variables). |
| **Partial Correlation** | Relationship between **vocabulary knowledge** and **reading comprehension**, controlling for students’ **age** as a confounding variable. |
| **Factor Analysis** | Analyzing students' responses to a **language anxiety questionnaire** to identify underlying factors (e.g., fear of speaking, fear of failure). |
| **t-Test** | Comparing the **mean writing scores** of students taught through peer feedback vs. traditional teacher feedback (two independent groups). |
| **F-Test (ANOVA)** | Comparing the **mean vocabulary scores** of students exposed to three teaching methods: flashcards, games, and rote learning. |
| **Two-Way ANOVA** | Examining how **teaching method** (flashcards vs. games) and **gender** interact to influence **listening comprehension scores**. |
| **ANCOVA** | Comparing **mean reading scores** of two teaching methods, while controlling for **pre-test scores** to account for students’ prior knowledge. |
| **Paired t-Test** | Comparing **pre-test and post-test speaking scores** for students who completed a 6-week oral presentation training program |

**b. Considerations for Non-parameteric Statistics**

This type of statistical analysis may be used effectively in the following situations:

* When non-probability sample is selected in the research study.
* The variables of the study are quantified at any level of measurement, mainly, nominal and ordinal scale. It may be in the discrete form.
* No assumption is required for this approach.
* Free distribution of data, may be skewed or may be normally distributed.
* Objectives of the study.

In this approach the following statistical techniques are generally used:

1. To study the relationship of two or more variables:

a1 – Spearman’s Rank correlation is appropriate for ordinal or interval data and measures the relationships of two variables. It is usually used when the sample size is small.

a2 – Chi-square test (*x2*) is used when two or more variables are taken. The data may be nominal or ordinal scale.

1. To analyse the difference between two or more groups:

b1 – Median test: used to compare medians of two or more groups, suitable for small or large groups.

b2 – *x2* test for comparing categorical data in large samples also for small sample if assumptions are met.

b3 – Mann- Whitney U-test compares two groups using ordinal data or non-normally distributed interval data.

| **Statistical Technique** | **Example in Language Education** |
| --- | --- |
| **Spearman’s Correlation** | Relationship between time spent reading English books and comprehension scores (ordinal data). |
| **Chi-Square Test** | Association between students’ first language (L1) and preferred learning strategies (nominal data). |
| **ANOVA** | Differences in listening comprehension scores among students using three different teaching methods. |
| **Median Test** | Median vocabulary scores comparison between game-based learning and memorization methods. |
| **Run Test** | Testing randomness in language test scores (e.g., by gender or testing time). |
| **Mann-Whitney U-Test** | Writing scores comparison between students using peer review vs. teacher feedback. |
| **Sign Test** | Consistent improvement in speaking fluency scores after debate training (paired pre-test and post-test). |

b4 – Sign test is used to compare paired data, often in small samples. It focuses on the direction of differences rather than their magnitude. It is applicable for ordinal.