Statistical Procedure

**When and how to use statistics**

**Statistics and communication sciences**

Statistics is a way of organizing, describing and making inferences from data and statistical methods are used throughout the physical, natural and social sciences. Statistics is a way of thinking, a language, and a means of making as argument based on data available (Ahelson, 1995).

There are two mutually interdependent concepts—research method and the use of statistics--of the scientific process. The first is the research design: approaches take to collecting the data to ensure your questions. The second one is the process of data analysis-- when and how to use statistics to evaluate the data with respect to the original questions that motivated the research in the first place.

Once you have developed the ability to understand and then properly apply the statistical principles and concepts to your scientific investigation, you will find that your every day thinking has changed as well. You will be able to participate in existing world of knowledge generation that is the field of communication. How is measurement done? It is a careful deliberate observation for the purpose of describing objects and events in terms of the attributes composing a variable.

**Importance**

Statistical procedures are used in communication sciences to both describe study results and to make inference. The use of statistics involves some kind of mathematical operation on numbers and than a translation of those numbers into something interpretable in the context of the study design and its purpose (Hayes, 2005).

* .The numbers in the data set being analyzed are almost always the result of some kind of measurement of the unit on one or more constructs.
* A construct is simple a concept that we use to describe a psychological, cognitive, and behavioral process or attribute of the object being studied.
* Media exposure, involvement, attitude towards action, political participation, aggression, communication, gender, population density, environment, nonverbal sensitivity, and family communication pattern are among the many constructs.
* It is through the measurement process that we are able to do such things as quantify the association between constructs, compare groups in their behavior, thoughts, or feeling, or otherwise test hypothesis that our intuition, curiosities or theories suggest.

**Statistics in Research**

* Statistics can play a very important role in answering your research questions in such a manner that you are able to quantify, measure, place a level of confidence on the findings, and make an assessment of the contribution-- each variable is to bringing out change.
* Statistics measures that association and relationship between variables, help to predict what is likely to happen in light of current trends (Kumar, 2005).
* Dispersion in individual responses, it come extremely difficult to understand the patterns in data, so it is important to summarize the data.
* Some simple statistical measures such as percentage, means, standard deviation, coefficients of correlations can reduce the volume of data and make easier to understand.
* Statistics play a vital role in understanding the relationship between variables, particularly when there are more than two.
* Statistics help you to ascertain the strength of relationship. They confirm or contradict what you read from a piece of information and provide indications of the strength of relationship and the level of confidence that can be placed in findings.
* If more than two variables, then statistics help in understanding the interdependency among them and their contribution to phenomenon or event.
* Indirectly knowledge of statistics helps you at each step of the research process. Knowledge of problems associated with data analysis, the types of statistical test that can be applied to certain types of variables.
* The calculations of summary statistics in relation to the measurement scale used play an important role in a research endeavor (Ahelson, 1995; and Kumar, 2005)

Typically, there are two general types of statistic that are used to describe data: descriptive statistics and inferential statistics. Descriptive statistics are used to synopsize (summarize) data from a sample exercising the mean or standard deviation. Inferential statistics are used when data is viewed as a subclass of a specific population.

**Descriptive Statistics**

Descriptive statistics is the discipline of quantitatively describing the main features of a collection of information, or the quantitative description itself. D**escriptive statistics are** use to observe data to predict what is true of areas beyond the **inferential statistics** data. Descriptive statistics simply presented our raw data. It enables us to present the data in a more meaningful way, which allows simpler interpretation of the data (Kille, 2015).

For example, if we had the results of 100 pieces of students' coursework, we may be interested in the overall performance of those students. We would also be interested in the distribution or spread of the marks. Descriptive statistics allow us to do this.

Measures of **central tendency**: these are ways of describing the central position of a frequency distribution for a group of data. In this case, the frequency distribution is simply the distribution and pattern of marks scored by the 100 students from the lowest to the highest.

Measures of **spread out scores**: these are ways of summarizing a group of data by describing how the scores are spread out. For example, the mean score of our 100 students may be 65 out of 100. However, not all students will have scored 65 marks. Some will be lower and others higher.

Descriptive statistics is also useful to summarize our group of data using a combination of tabulated description-- tables, graphical description (graphs and charts) and statistical commentary-- a discussion of the results), (see for details, Wimmer & Dominic, 2015; Bailey,1998).

**Types of descriptive statistics**

* Measures of frequency: count, percent, frequency.
* Measures of central tendency: mean, median, and Mode.
* Measures of dispersion or variation: range, variance, standard deviation.
* Measures of position: percentile ranks, quartile ranks (ibid).

**Standard deviation** provides insight into how much variation there is within a group of values. It measures the deviation (difference) from the group’s mean (average).

**Average**, **mean** terms are synonymous (identical), and refer to the average value of a group of numbers. Add up all the figures, divide by the number of values, and that’s the average or mean. **Median** is the central value, and and if there’s an extremely high or low value in a collection of values, or a cutting point between high and low values of a variable (question items).

**Mode** is the frequently occurred figure (number) in the dispersion of score on single question’s response categories

**Inferential statistics**

Inferential statistics makes inferences and predictions about a population based on a sample of data taken from the population in question. With inferential statistics you take that sample data from a small number of people and try to determine if the data can predict, whether the drug will work for everyone (i.e. the population) (Kille, 2015). Inferential statistics use statistical models to help you compare your sample data to other samples or to previous research.

The most common methodologies in inferential statistics are hypothesis tests, confidence intervals, and regression analysis. Interestingly, these inferential methods can produce similar summary values as descriptive **statistics**, such as the mean and standard deviation.

**Goal**

The goal of the inferential statistics is to draw **conclusions** from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, Analysis of variance etc.

A **sample** is a portion of an entire **population**. Inferential statistics seek to make predictions about a population based on the results observed in a sample of that population.



There are two primary types of population samples: **random**and **stratified**. For a random sample, study subjects are chosen completely by chance, while a stratified sample is constructed to reflect the characteristics of the population at large (gender, age or ethnicity, for example). There are a wide range of sampling methods, each with its advantages and disadvantages.

Attempting to extend the results of a sample to a population is called **generalization**. This can be done only when the sample is truly representative of the entire population.

Generalizing results from a sample to the population must take into account **sample variation**. Even if the sample selected is completely random, there is still a degree of**variance**within the population that will require your results from within a sample to include a**margin of error (Kille, 2015).**.

Sample selected is completely random, there is still a degree of**variance**within the population that will require your results from within a sample to include a**margin of error**.

**For example**

The greater the sample size, the more representative it tends to be of a population as a whole. Thus the margin of error falls and the **confidence level**rises. One variable rises or falls, the other does as well — caloric intake and weight, for example**. Negative correlation** indicates that two variables move in opposite directions — say, vehicle speed and travel time. So if a scholar writes “Income is negatively correlated with poverty rates,” what he or she means is that as income rises, poverty rates fall (ibid).

**Causation**is when change in one variable alters another. For example, air temperature and sunlight are correlated (when the sun is up, temperatures rise), but causation flows in only one direction. This is also known as**cause and effect**.

**Regression analysis**

**It** is a way to determine if there is or isn’t a correlation between two (or more) variables and how strong any correlation may be. At its most basic, this involves plotting data points on a X/Y axis (in our example cited above, vehicle weight and fatal accidents) looking for the **average causal effect.** This means looking at how the graph’s dots are distributed and establishing a **trend line**. Again, correlation isn’t necessarily causation. There are main areas of inferential statistics:

* 1. Estimating parameter taking a statistic form your sample data 9for example, the same mean) and using it to say something about a population parameter (i.e. population).
	2. Hypothesis tests: this is where you can use sample data to answer research question. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in school (sample data, sample mean). (Glen, 2014) .

**Examples** include: "How did you arrive at that conclusion?" Asking how and why questions help you weigh the merits of the answers. Variance and standard deviation require the mean to be calculated, which is not appropriate for categorical variables as they have no numerical value. Inferential statistics: making inferences about quantitative data from a sample, estimates or projections for the total population can be produced.

**Conclusion**

**Descriptive statistics** are the basic measures used to describe survey data.  They consist of summary **descriptions of single variables**(also called “univariate” analysis) and the associated survey sample.  Examples of descriptive statistics for survey data include frequency and percentage response distributions, measures of central tendency (which include the mean, median and mode), and dispersion measures such as the range and standard deviation, which describe how close the values or responses are to central tendencies.

**Inferential statistics** offer more powerful analyses to be performed on your online web survey data.  As the names suggests, this branch of statistics is concerned with making larger inferences about social phenomena.  This can include **associations between variables**, how well your sample represents a larger population, and cause-and-effect relationships.  Some examples of inferential statistics commonly used in survey data analysis are t-tests that compare group averages, analyses of variance, correlation and regression, and advanced techniques such as factor analysis, cluster analysis and multidimensional modeling procedures

**Statistical tests**

A **chi-squared test**, also written as ***χ*2 test**

It is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis, specifically Pearson's chi-squared test and variants thereof. Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table (William, 1952).

Chi-squared tests often refers to tests for which the distribution of the test statistic approaches the *χ*2 distribution asymptotically, meaning that the sampling distribution (if the null hypothesis is true) of the test statistic approximates a chi-squared distribution more and more closely as sample sizes increase (Wikipedia)

Test statistics that follow a *χ*2 distribution occur when the observations are independent and normally distributed, which assumptions are often justified under the central limit theorem. There are also *χ*2 tests for testing the null hypothesis of independence of a pair of random variables based on observations of the pairs. It is commonly used to explore the direction relationship (whether positive or negative).

## T-Test

A t-test is a type of inferential statistics used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances. A t-test is used as a hypothesis testing tool, which allows testing of assumption applicable to a population (Kenton, 2020).

* The t-test is one of many tests used for the purpose of hypothesis testing in statistics.
* Calculating a t-test requires three key data values. They include the difference between the mean values from each data set (called the mean difference), the standard deviation of each group, and the number of data values of each group.
* There are several different types of t-test that can be performed depending on the data and type of analysis required.

**F-Test**

The F-test of overall significance indicates whether your linear regression model provides a better fit to the data than a model that contains no independent variables. R-squared  tellsyou how well your model fits the data, and the F-test is related to it. An F-test is a type of statistical test that is very flexible (Wikipedia).

The F-Statistic: Variation between sample means / variation within the samples. The F-statistic is the test statistic for F-tests. In general, an F-statistic is a ratio of two quantities that are expected to be roughly equal under the null hypothesis, which produces an **F**-**statistic** of approximately (Minitab Blog Editor, 2016).

An ***F*-test** is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data are sampled. Exact "*F*-tests" mainly arise when the models have been fitted to the data using least squares.

Common examples of the use of *F*-tests include the study of the following cases:

* The hypothesis that the means of a given set of normally distributed of the populations, all having the same standard deviation, are equal. This is perhaps the best-known *F*-test, and plays an important role in the analysis of variance (ANOVA).
* The hypothesis that a data set in a regression analysis follows the simpler of two proposed linear models that are nested within each other.

**Difference (t-test & f-test)**

# The main difference between the t-test and f-test is, that t-test is used to test the hypothesis whether the given mean is significantly different from the sample mean or not. On the other hand, an F-test is used to compare the two standard deviations of two samples and check the variability.

**Z-Test**

A z-test is a statistical test to determine whether two population means are different when the variances are known and the sample size is large. It can be used to test hypotheses in which the z-test follows a normal distribution. A z-statistic, or z-score, is a number representing the result from the z-test (Chen, 2020).

A z-test is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large. The test statistic is assumed to have a normal distribution, and nuisance (imitation) parameters such as standard deviation should be known in order for an accurate z-test to be performed.

* A z-test is a statistical test to determine whether two population means are different when the variances are known and the sample size is large.
* It can be used to test hypotheses in which the z-test follows a normal distribution.
* A z-statistic, or z-score, is a number representing the result from the z-test.
* Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size.
* Also, t-tests assume the standard deviation is unknown, while z-tests assume it is known.

## Analysis of Variance (ANOVA)

# Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study (Kenton 2020).

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples (Mackenzie, 2018).

* Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
* A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
* If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

##  How to Use ANOVA

A researcher might, for example, test students from multiple disciplines of a University to see if students from one of the departments consistently outperform students from the other departments. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

# A one-way ANOVA is primarily designed to enable the equality testing between three or more means. A one-way ANOVA only involves one factor or independent variable, whereas there are two independent variables in a two-way ANOVA.

In a one-way ANOVA, the one factor or independent variable analyzed has three or more categorical groups. A two-way ANOVA is designed to assess the interrelationship of two independent variables on a dependent variable (see for difference the below).

|  |  |  |
| --- | --- | --- |
|  | One-Way ANOVA | Two-Way ANOVA |
| Definition | A test that allows one to make comparisons between the means of three or more groups of data. | A test that allows one to make comparisons between the means of three or more groups of data, where two independent variables are considered.  |
| No. of Independent Variables | One. | Two.  |
| What is Being Compared? | The means of three or more groups of an independent variable on a dependent variable. | The effect of multiple groups of two independent variables on a dependent variable and on each other.  |
| Number of Groups of Samples  | Three or more. | Each variable should have multiple samples. |

 **Adopted by Mackenzie, 2018**

**Univariate** **analysis**

It involves analyzing one variable at a time. Frequency distributions show the numbers and percentages of people or items that fall into different categories. Categories can be nominal, ordinal, interval, or ratio. ... Other tests are available for ordinal, interval, and ratio data (Segal, 2015).

* Frequency distributions show the numbers and percentages of people or items that fall into different categories.
* Categories can be nominal, ordinal, interval, or ratio.

A variable is any characteristic that can be observed or measured on a subject. In clinical studies a sample of subjects is collected and some variables of interest are considered. Univariate descriptive analysis of a single variable has the purpose to describe the variable distribution in one sample and it is the first important step of every research study.

It is the distribution of respondents’ scores on a single variable (question) response categories,

**For example**: viewers television watching habit of infotainment programs.

How often do you usually prefer to watch television infotainment programs?

 1 2 3 4 5

Very often often sometimes rarely never. N==100.

 15 25 30 18 12

**Mean** (100÷5=**20**)  **Mode**: the most frequent figure **3 Median**: splitting point, 1 & 2=high, and 3+4+5=low (15+25=**40** high, and 30+18+12=**60** low)

**Bivariate frequency**

The classes of one variable may be arranged horizontally, and the classes of another variable may be arranged vertically in the two way table. By going through the pairs of values of X and Y, we can find the frequency for each cell. The whole set of cell frequencies will then define a bivariate frequency distribution. In other words, a bivariate frequency distribution is the frequency distribution of two variables.

The following table shows the frequency distribution of two variables, namely, age and marks obtained by 50 students in an intelligent test. Classes defined for marks are arranged horizontally (rows) and the classes defined for age are arranged vertically (columns). Each cell shows the frequency of the corresponding row and column values. For instance, there are 5 students whose age fall in the class 20 – 22 years and their marks lie in the group 30 – 40. (KS academy, 2019).



 **Uploaded by KS Academy**

Table shows the frequency distribution of two variables, namely, age and marks obtained by 50 students in an intelligent test. Classes defined for marks are arranged horizontally (rows) and the classes defined for age are arranged vertically (columns). Each cell shows the frequency of the corresponding row and column values

**Multivariate statistics**

**Multivariate statistic** is a subdivision of statistics encompassing the simultaneous observation and analysis of more than one outcome variable. Multivariate statistics concerns understanding the different aims and background of each of the different forms of multivariate analysis, and how they relate to each other. The practical application of multivariate statistics to a particular problem may involve several types of univariate and multivariate analyses in order to understand the relationships between variables and their relevance to the problem being studied (Hidalgo, & Goodman, M. (2013).

In addition, multivariate statistics is concerned with multivariate probability distributions, in terms of both: how these can be used to represent the distributions of observed data; and how they can be used as part of statistical inference, particularly where several different quantities are of interest to the same analysis. There are a set of probability distributions used in multivariate analysis that play a similar role the corresponding set of distribution is appropriate to dataset. These multivariate distribution are: multivariate normal distribution **(**Johnson; et al (2007).

**Example.**  A researcher has collected data on three psychological variables, four academic variables (standardized test scores), and the type of educational program the student is in for 600 high school students. A doctor has collected data on cholesterol, blood pressure, and weight (ibid). Multivariate analysis (MVA) is based on the principles of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time.

**Regression statistics**

## What Is Regression?

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables), (Brian (2020).

The two basic types of regression are simple **linear regression** and **multiple linear regression**, although there are non-linear regression methods for more complicated data and analysis. Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome (ibid)

Regression takes a group of random variables, thought to be predicting Y, and tries to find a mathematical relationship between them. This relationship is typically in the form of a straight line (linear regression) that best approximates all the individual data points. In multiple regression, the separate variables are differentiated by using subscripts.

* Regression helps investment and financial managers to value assets and understand the relationships between variables

Regression can help finance and investment professionals as well as professionals in other businesses.