BASICS OF STATISTICS FOR SURGEONS

Dr. Subashini R, Sri Ramachandra Medical College, Chennai

doctor.subashini@ymail.com

Dr. Vimalendu Brajesh, Medanta Hospital, Gurugram

drvbrajesh@yahoo.co.in

ISSHACADEMICS

Data is the sword of the 21st century, those who wield it well, the Samurai-

Jonathan Rosenberg

Statistics is not a comfortable domain for many surgeons. However, in modern medicine where publications and research are becoming an essential part of one's growth it is imperative to get familiar with statistics. It not only allows one to appropriately analyse a study and accept or reject its proposal but also enables one to present their work in a more impressive and convincing manner. This article provides a basic introduction to various terminologies used in statistics and introduces various statistical tests and methods used in common study designs.

INTRODUCTION

Statistics is the science of collecting, summarizing, and interpreting data (variables), and of using this data to estimate the size and strengths of associations between variables. It's a tool for converting data into meaningful information. Simply observing data may indicate towards certain findings, but to conclude that the findings are real and not due to chance we need to resort to appropriate statistical analysis. The type of statistical analysis that needs to be applied to a given data depends on the data available and study design.

The sequence of a study can be summarised as shown in (Figure 1).

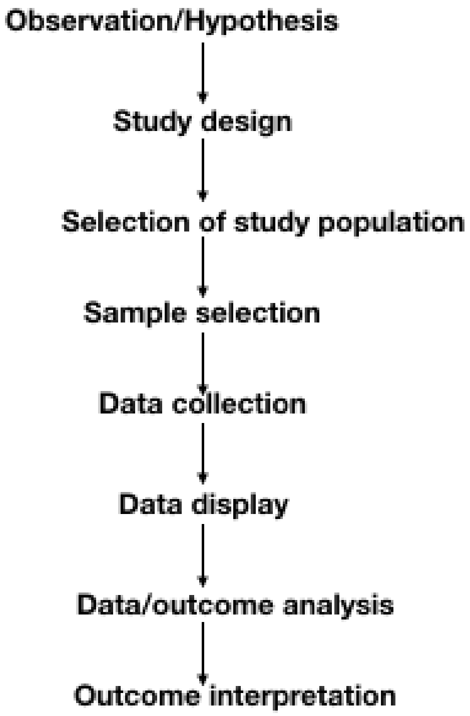

Figure – 1: Flow chart of sequence of events in a scientific study.

HYPOTHESIS/QUESTION/OBSERVATION

It all starts with a question in the mind of an investigator, or an observation made by the investigator. Based on this question or observation the investigator formulates a hypothesis. To further approve or disapprove this hypothesis the investigator does a scientific enquiry called statistical analysis.

STUDY DESIGN

Study design is a set of methods and procedures used to collect and analyze data on variables specified in a particular study. They can be broadly divided into descriptive and analytic studies. (Figure-2)

Descriptive study – These studies are easy to conduct. The focus of a descriptive study is to elaborate on one or few variables. They neither try to establish relation between exposure and outcome, nor they try to answer any specific question. Case reports and case series are the most commonly used descriptive studies used across the board by surgeons.

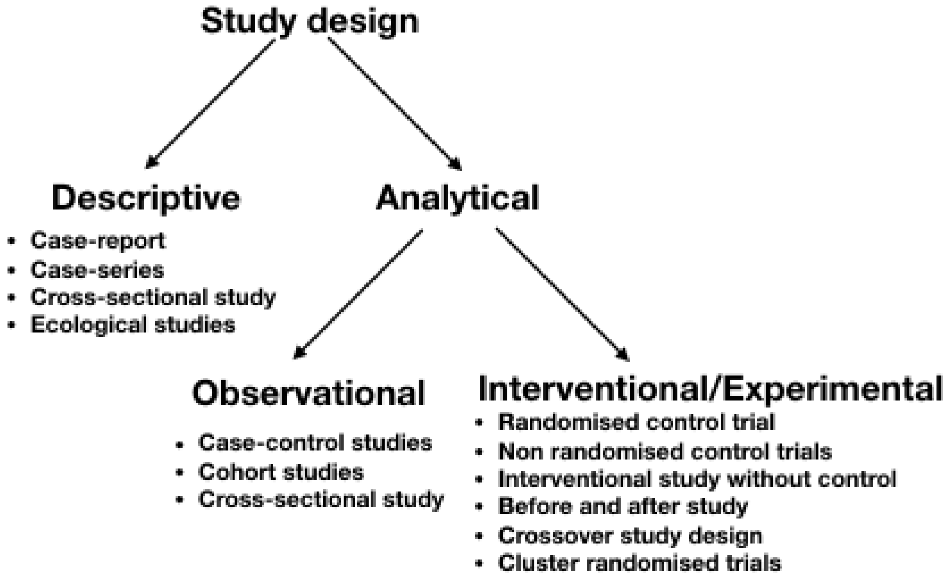

Figure – 2: Different study designs in medical statistics.

Analytical observational study – These studies try to identify relationship between variables (exposure and outcome variables). Two commonly performed observational analytical study are case control and cohort study.

In a case control study, outcome is predetermined or known, and the investigator does a retrospective study to establish relationship between exposure and outcome. Contrary to this, in a prospective cohort study exposure variable and outcome variable to be registered are determined and the investigator progresses in a prospective manner. (Figure-3)

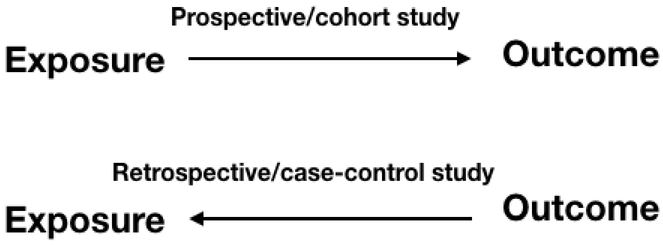

Figure – 3: Pictorial representation of two most commonly done studies.

Analytical Interventional/experimental study – The investigators perform an intervention and determine the outcome of such intervention. Example- Studying the effect of a new drug/surgical technique. The different designs of this study have been mentioned in figure 2.

TYPES OF DATA

Raw data are sets of observations made in a sample from the target population. Each individual observation is known as a variable.

They can be divided into two broad categories- qualitative/ quantitative (Figure – 4)

1. Qualitative data or Categorical data- consists of categories and not numerical data.

   a) Nominal data: consists of "naming" observations and classifying them into different and exclusive classes. e.g.- gender [male/ female/ others], smoker or non-smoker.

   b) Ordinal data: when the classes are not only different but also have a definite order as per some criterion. e.g. - stages of carcinoma [progressively worsening from Stage I to IV], stages of a SLAC wrist.

2. Quantitative data – consists of numerical observations rather than just categories. It can be divided into Discrete or Continuous based on whether only whole numbers constitute the data or not.

   a) Interval data: consists of observations with a defined interval, which is to say that the distance between any to measurements is known. The difference between measurements of 10 and 20 is equal to the difference between measurements of 20 and 30. E.g. – Celsius or Fahrenheit scales for temperature. Interval data do not contain a true zero point, e.g. - Zero degree Celsius does not mean an absence of heat, since it has been arbitrarily defined to be the freezing point of water. This also implies that ratios are not meaningful, since 20^{o}C cannot be said to be twice as hot as 10^{o}C.

   b) Ratio data: similar to an interval scale, but also has a true zero point. This makes ratios meaningful. E.g.- height or weight. Zero grams implies a complete absence of weight and 20 grams is equal to twice 10 grams.

The type of data that is measured in a study is important to be understood clearly, since this has a bearing on the type of test that can be used for studying significance later. Categorical data tests are entirely different from tests used for Quantitative data.

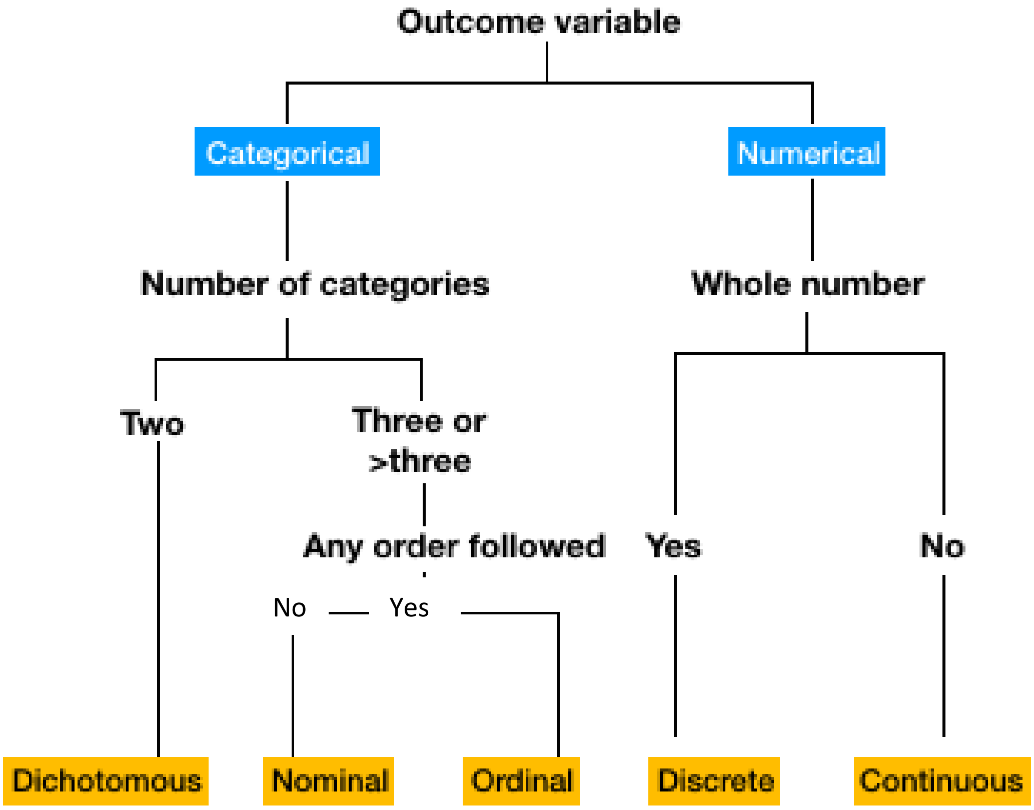

Figure – 4: Types of Data

Methods of Data Collection-

- In depth Interviews (individual interviews)
- Focus group discussions- open ended group interviews
- Participant Observation- researcher becomes participant in social event or group under study and records observation.
- Web survey

SUMMARIZING DATA

For numerical outcomes

Any set of numerical outcome variable can be summarized with two types of measures – *measures of central tendency* and *measures of dispersion*.

Measures of central tendency give us a single value that is considered to by typical of the whole data. The three most commonly used are the Mean, Median and the Mode.

- Mean – defined to be the obtained by adding up all the values and dividing by the number of values. Usually designated by μ.
- Median- middle value of distribution when the data is arranged in ascending or descending order. This value divides the set of values into two equal parts.
- Mode- most frequent value in a data set.

Measures of dispersion convey information about the variability present in a set of data.

- Range - difference between minimum and maximum value
- Mean deviation- average of the absolute deviations of the observations from the arithmetic mean.
- Variance – attempts to measure dispersion in relation to the mean. The deviation from the mean of each observation is squared up and then all of them added, which is divided by the number of values. It can be described as the *average square deviation*. Usually designated as σ^2
- Standard deviation – the variance represents squared units since each deviation is squared. To represent dispersion in the original units, the square root of the variance is used and called the standard deviation. Usually designated by σ.

For categorical outcomes

- Proportion – It is the number of individuals who experience an event divided by total number in the sample. The sample proportion is used to estimate the probability or risk.
- Risk - It is the ratio of probability of exposure outcome under consideration to all possible outcomes.
- Risk ratio (**Relative risk**, RR)- it is the ratio of risk in exposed group and risk in non-exposed group usually measured from a cohort study.

Presentation of data of a Cohort Study (**Relative Risk**):

	Diseased	*Non- Diseased*	*Total*
Exposed	a	b	a+b
Unexposed	c	d	c+d
	a+c	b+d	a+b+c+d

$$Relative\ Risk = \frac{Incidence\ of\ disease\ in\ exposed}{Incidence\ of\ disease\ in\ unexposed} = \frac{a/(a+b)}{c/(c+d)}$$

If Relative Risk =1, Incidence in both group is same, that is exposure is not associated with disease.

If Relative Risk >1, Incidence in exposed is higher, that is exposure is positively associated with disease.

If Relative Risk<1, Incidence in exposed is lower, that is exposure is negatively associated with disease.

- Odds Ratio (OR)- is a measure of risk used for a retrospective study like a case control study. The term ‘odds’ denotes the ratio of probability of success to probability of failure. The odds for being a case in the Exposed and the Unexposed groups is calculated and the ratio of these is termed the Odds Ratio [OR]. In other words, it is the ratio of odds in exposed groups to odds in unexposed group.

Presentation of data of a Case Control Study (**Odds Ratio**):

	Cases	*Controls*	*Total*
Exposed	a	b	a+b
Unexposed	c	d	c+d
	a+c	b+d	a+b+c+d

$$Odds\ Ratio = \frac{Odds\ of\ case\ in\ Exposed\ group}{Odds\ of\ case\ in\ Unexposed\ group} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

If Odds Ratio=1, odds of exposure among case and control are same.

If Odds Ratio>1, odds of exposure among case is higher, that is exposure is positively associated with disease.

If Odds Ratio<1, odds of exposure among case is lower, that is exposure is negatively associated with disease.

- Difference between odds ratio and relative risk – The relative risk calculation requires that all exposed to a risk factor be included in the denominator. This would be possible in a prospective study like a cohort study. However, in a retrospective case control study, the study begins from identification of cases and finds out whether they had been exposed or not. In such a study, all the people exposed to said risk factor is not known and the RR cannot be calculated. For such a situation the OR is a better measure.

POPULATION AND SAMPLE

Clinical studies are conducted to reach a conclusion about a population, on the basis of information obtained from a sample drawn from that population.

- Study population- The study population is the population to which the results of the study will be inferred.
- Sample- Subset of the population that will be actually studied.

For example, in a study conducted on distal radius fracture patients, the population would be *all* distal radius fracture patients and the sample would be the subjects of that particular study. The population is usually designated N and the sample by n.

- Sampling- procedure by which some members of the population [*the sample*] are selected as representatives of the entire population.
- Types of sampling-

  1. Nonprobability sampling- selected by researcher’s choice. Few examples are:
     a) Purposive sampling- participants from whom information can be easily obtained are selected.
     b) Convenience sampling- ease of accessibility. This is actually the most commonly used method of sampling in clinical practice due to the ease and convenience and also it being least expensive method.
  2. Probability sampling- probability to be selected is known. Few examples are:

a) Simple random sampling- number all participants and randomly draw numbers. Each participant has the same probability of selection.
b) Systematic random sampling- every n^{th} unit is taken
c) Stratified random sampling- Population is classified into homogenous subgroups (strata). From each strata, a simple random sample is taken and all are combined.

Confidence Interval

Any study conducted on a sample is an attempt to estimate properties of the entire population. For example, in a study of 50 distal radius fracture patients the mean DASH score can be assumed to be nearabout the mean score of all such patients. However, it is important to realize that this mean (m) is only an estimate of the population mean (μ). Similarly, the sample standard deviation (s) is an estimate of the population standard deviation (σ).
{PS- All symbols of the population are in Greek and Sample in English}

The Central Limit Theorem – when repeated samples of size 'n' are drawn from a population (N), the sample means (m_1,m_2,m_3,…m_n) are approximately normally distributed. This distribution of sample means has a mean equal to the population mean (μ) and standard deviation $\sigma/\sqrt{n}$. This particularly important property allows us to estimate the 'μ' when the sample mean 'm' is known. The value $\sigma/\sqrt{n}$ is called the *standard error of the mean [SEM]*.

In all practical studies, we obtain information about the mean 'm' and standard deviation 's' of a particular measure. The standard error is usually considered to be approximately $s/\sqrt{n}$. From the properties of the normal distribution, it is known that 95% of all values fall between 1.96 standard deviations on either side of the mean. This allows to say with 95% certainty [or 'confidence'] that the population mean μ would lie in the range $m \pm 1.96 \text{ x SEM} = m \pm 1.96\ s/\sqrt{n}$. So, this range is called the *95% Confidence Interval* [95% CI] for the mean. It is important to note that the width of the interval depends only on the SEM, which in turn decreases with increasing sample size [$SEM = s/\sqrt{n}$]. Therefore, increasing the sample size makes the 95% CI narrower. In other words, increasing the sample size allows to estimate the population mean with greater precision.

SAMPLE SIZE

The sample size estimate provides the minimum required size for a particular study. A larger than required sample size is a strain on resources, though with an advantage of increased precision (see above). A smaller than required sample, on the other hand, ends up with unreliable conclusions.

When designing a study, we can setup an allowable margin of error in our estimation. This allowable error can be considered as the width of the confidence interval that we would consider acceptable. Designating this error margin as 'd',

$$d = 1.96\,SEM = 1.96(\frac{\sigma}{\sqrt{n}})$$

Again, this reinforces the idea that greater sample size allows lesser error and, therefore, better precision.

From the above equation, solving for n gives,

$$n = (1.96)^2 \mathrm{X}\ \frac{\sigma^2}{d^2}$$

The margin of error can be set up by the investigators, say 5% or 10% (d = 0.05 or 0.1), but the standard deviation 'σ' needs to be assumed. In practice, this can be either from a prior study or a small pilot study using the same intended methodology of the study proper.

STATISTICAL ANALYSIS OF DATA

Descriptive studies-

In descriptive study, the results are obtained as estimates. The result is reported in % confidence interval. The general formula for % CI is-

$$Confidence\ interval = mean \pm z\ \frac{s}{\sqrt{n}}$$

z is the number of standard deviations in the width of the CI. For e.g., z = 1 gives a CI of 68%. z = 1.96 gives the earlier formula for the 95% CI.

Analytical Studies-

Analytical studies involve testing of a hypothesis. It begins with an assumption that there is no difference between the outcome of two or more groups, called the *Null Hypothesis*. The Alternative Hypothesis is often that these groups are different. The null hypothesis is the hypothesis to be tested, which basically involves statistical tests that help in deciding whether the null hypothesis is to be accepted or rejected.

The appropriately chosen test assigns a p-value to the data collected. The p-value is the probability that the difference between groups is entirely due to chance. Hence it is also the probability that the null hypothesis is true. Most studies choose a p-value of 0.05 as the cutoff for rejecting the null hypothesis and accepting the alternative hypothesis.

Errors- all hypothesis testing methods have inherent errors which are not entirely avoidable.

Type I error (α error)- rejecting null hypothesis when it is actually true. When the cutoff p-value is set at 5% or 0.05, this means that we would wrongly reject the null hypothesis 5% of the time. Hence, the standard α error rate is 0.05.

Type II error (β error)- accepting null hypothesis when it is actually false. Hence, the value 1-β (one minus β) gives us the probability of correctly rejecting the null hypothesis and is called the ***Power*** of the study.

While the α error is known and is a matter of choice mostly, the β error values is more complex and depends on various factors like the population and sample means, sample size, α error etc. β error rate of 20% or 0.2 is often considered acceptable and this corresponds to a Power of 80%.

Condition of Null Hypothesis

Possible Action		TRUE	FALSE
	ACCEPT	Correct (1 - α)	Type II error (β)
	REJECT	Type I error (α)	Correct (1 - β)

Figure-5. Depicting Types of errors in hypothesis testing

STATISTICAL TESTS

Type of statistical analysis to be used depends on type of sampling distribution and dependency of the variable.

1. Sampling Distribution- it can be normal (Gaussian or Bell shaped or parametric) or non-normal. For data that are considered normally distributed, parametric tests are performed. If the data are non-normally distributed, then non-parametric tests are more appropriate. Tests that determine the normality of data are the Kolmogorov - Smirnov test, Lilliefors test, Shapiro-Wilk test etc.

2. Dependent or paired and Non-dependent or unpaired Sample- Paired or dependent sample means a test subject has two sample values.

The statistical test which can be done for numeric outcome [Interval or Ratio scale data] has been shown in the chart below.

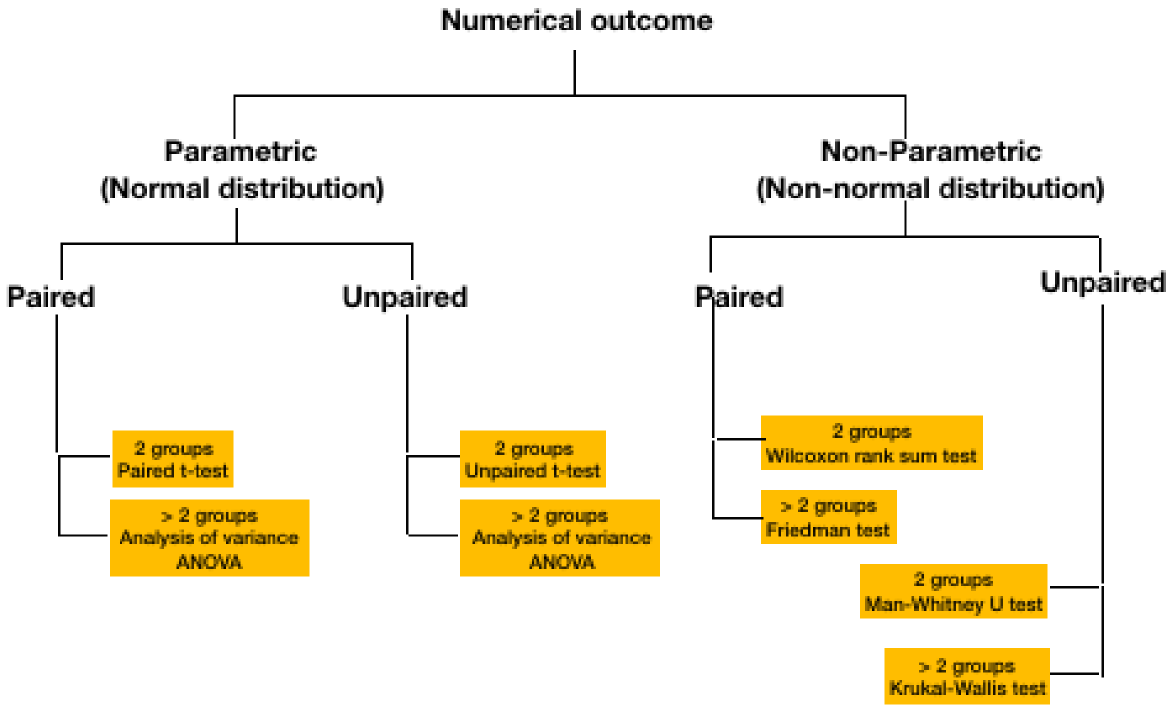

Figure-6: Choice of test for numerical data

The statistical test which can be done for Categoric outcome [Nominal or Ordinal scale data] has been shown in the chart below.

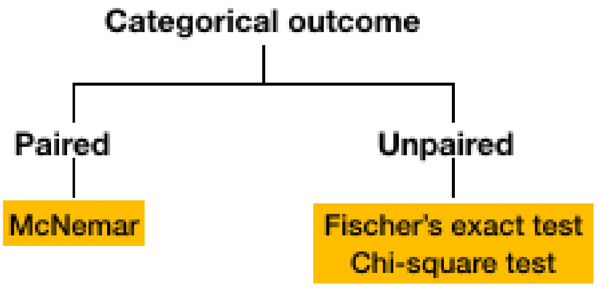

Figure-7: Choice of test for categorical outcome

The statistical tests commonly used are briefly described below:

- **Student T Test**- To analyse small samples with 't' distribution. T distribution is similar to normal distribution symmetrical with equal mean, median and mode but more spread out. Unpaired t test and Paired t test are used for unpaired and paired samples respectively.
- **Chi Square test**- To analyse significance of difference between two proportions. It is used for nominal and ordinal data. Different variants of Chi square test are Fisher's exact test, McNamara test and Cochran Mantel Haenszel test.
- **ANOVA** (Analysis of Variance)- used to compare means of three or more mutually exclusive groups. There is one dependent and one independent variable in One way ANOVA. There is one dependent and two independent variable in two way ANOVA. In Repeated measure ANOVA, the dependent variable is measured repeatedly at different time or under different circumstance. In Friedman's ANOVA median value is compared instead of mean. Multivariate Analysis of variance is used for continuous variable.
- **Mann-Whitney U test**- To analyse differences between the medians of two different data sets (unpaired sample). This is an alternative of unpaired t test.
- **Wilcoxon Matched Pairs Signed Rank Test**- To analyse difference between medians of paired sample. This is an alternative of Paired t test.
- **Correlation**- Correlation represents relationship between two variables. Correlation Coefficient express the degree and direction of the relationship. *Causation is not determined.* Correlation Coefficient varies from -1 through 0 to +1, where -1 represents negative correlation and +1 represents positive correlation. Pearson's Correlation Coefficient is most often used.
- **Regression**- Regression predicts how much change is caused in other variable for one unit change in one variable. Regression coefficient is a measure of change of one dependent variable with one unit change in dependent variable. Linear Regression is often used to predict the outcome of continuous data. Logistic Regression is used to predict the outcome of discrete data.

CONCLUSION- Complex statistical analysis remains a specialised domain to be interpreted with the help of a professional statistician but the basic knowledge is essential for all surgeons.

For conducting a study, the groundwork with respect to data collection, study design, formulation of hypothesis still must be performed by the research team of surgeons. Basic understanding of statistics is also required to properly interpret the analysis and results of any published study.

ACKNOWLEDGEMENT- Dr. Ajeesh Sankaran for finetuning the final manuscript.

REFERENCES

1. Essential medical statistics-Betty R Kirkwood and Jonathan A.C. Stern
2. Krousel-Wood MA, Chambers RB, Muntner P. Clinicians' guide to statistics for medical practice and research: part I. Ochsner J. 2006;6(2):68-83.
3. Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014 Feb 15;24(1):12-8. doi: 10.11613/BM.2014.003. PMID: 24627710; PMCID: PMC3936971.
4. du Prel JB, Röhrig B, Hommel G, Blettner M. Choosing statistical tests: part 12 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010 May;107(19):343-8. doi: 10.3238/arztebl.2010.0343. Epub 2010 May 14. PMID: 20532129; PMCID: PMC2881615.
5. Duquia RP, Bastos JL, Bonamigo RR, González-Chica DA, Martínez-Mesa J. Presenting data in tables and charts. An Bras Dermatol. 2014 Mar-Apr;89(2):280-5. doi: 10.1590/abd1806-4841.20143388. PMID: 24770505; PMCID: PMC4008059.
6. Goldberg RJ, McManus DD, Allison J. Greater knowledge and appreciation of commonly-used research study designs. Am J Med. 2013 Feb;126(2): 169.e1-8. doi: 10.1016/j.amjmed.2012.09.011. PMID: 23331447; PMCID: PMC3553494.
7. Indranil S, Bobby P. Essentials of Biostatistics. 2nd Edition. Academic Publishers 2017
8. Health research methodology : a guide for training in research methods. 2nd ed.. Manila:WHO Regional Office for the Western Pacific. https://apps.who.int/iris/handle/10665/206929
9. Basic Course in Biomedical research – NPTEL
10. Lopes, Bernardo & Ramos, Isaac & Ribeiro, Guilherme & Correa, Rosane & Valbon, Bruno & Luz, Allan & Salomão, Marcella & Lyra, João & Jr, Renato. (2014). Biostatistics: Fundamental concepts and practical applications. Revista Brasileira de Oftalmologia. 73. 10.5935/0034-7280.20140004.