Original Research Paper

Volume - 12 | Issue - 06 | June - 2022 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Statistics

UNDERSTANDING BIOLOGICAL DATA AND PROBABILITY DISTRIBUTIONS FOR MEDICAL STATISTICS.

Nandkishore singh Statistician and Associate Professor, Dept.Com.Med, SHKMGMC Nalhar, Mewat, Haryana.

Dr. Neha Nehra* Dept.Com.Med, SHKMGMC Nalhar, Mewat, Haryana. *Corresponding Author

ABSTRACT Bio-Statistics remains an enigma for physicians, even though the knowledge is essential not only for research but also for understanding and interpreting information relevant to the practice of medical science. At the base of all high-end statistics presently in use in the medical field is the understanding of the type of data dealt with. The data generated from the observation of and intervention in the human body are as diverse as all the anatomical, physiological, and biochemical parameters. Added to the complexity of human biology is the environment with which humans interact and, therefore, has very intricate relationship with health. In the endeavour to understand the causes of illnesses and finding appropriate solutions for the same, man has, from time immemorial, generated volumes of information based on data. But with the advent of modern medicine and sophisticated technology both for investigations and treatment of disease conditions, generating evidence for the use of simplest to most complicated procedures has come to be known as evidence-based medicine. Therefore, all medical professionals are expected to have a sound knowledge in the medical statistics to be able to practice their craft efficiently. As the foundation of medical statistics is knowledge of the data generated by day-to-day practice and research, this paper attempts at understanding the Biological data and choosing the appropriate Probability distribution.

KEYWORDS :

INRODUCTION:

In statistics, a categorical variable (also called qualitative variable) is a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property. In computer science and some branches of mathematics, categorical variables are referred to as enumerations or enumerated types. Commonly (though not in this article), each of the possible values of a categorical variable is referred to as a level.The probability distribution associated with a random categorical variable is called a categorical distribution.

Categorical data is the statistical data type consisting of categorical variables or of data that has been converted into that form, for example as grouped data. More specifically, categorical data may derive from observations made of qualitative data that are summarised as counts or cross tabulations, or from observations of quantitative data grouped within given intervals. Often, purely categorical data are summarised in the form of a contingency table. However, particularly when considering data analysis, it is common to use the term "categorical data" to apply to data sets that, while containing some categorical variables, may also contain non-categorical variables.

A categorical variable that can take on exactly two values is termed a binary variable or a dichotomous variable; an important special case is the Bernoulli variable. Categorical variables with more than two possible values are called polytomous variables; categorical variables are often assumed to be polytomous unless otherwise specified. Discretization is treating continuous data as if it were categorical. Dichotomization is treating continuous data or polytomous variables as if they were binary variables. Regression analysis often treats category membership with one or more quantitative dummy variables.

Data and variables:-

1. Nominal variable is one where the values fall into unordered categories (e.g. gender, religion, color of the eyes, or blood groups). Even though numbers are used to represent the categories, like 0 for males and 1 for females, for performance of statistical procedures in computers, it is important to remember that there is no order in these categories and there cannot be an average of 0.5 genders.

2. Ordinal variable has ordered categories, but the differences between the categories cannot be considered equal. For example, though numbers are used to denote the stages of cancer, the difference in the severity of disease between stages I and II may not be the same as between stages II and III. The numbers only signify the order. Other examples are power of muscles, severity of dyspnea, and class of occupation.

3. An interval variable has all the characteristics of an ordinal variable, and also, the differences or distances between any two values on the scale are equal, but the zero point is arbitrary. Examples are temperature measured as Celsius and the intelligence quotient (IQ). The difference between IQ 50 and 70 is same as the difference between IQ 90 and IQ 110; but because the zero point is artificial and movable, IQ 100 is not twice as high as IQ 50.

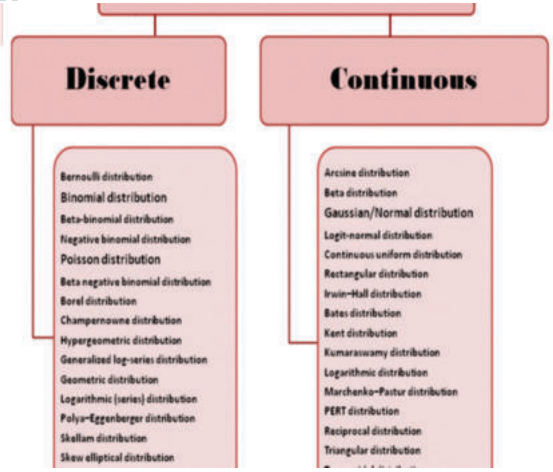
4. A ratio variable has all the characteristics of an interval variable and, in addition, has a true zero point. Only when the zero point is meaningful, the ratios between the numbers are also meaningful. Height, weight, and most laboratory test values are ratio data.

Variables can also be classified as qualitative and quantitative. Nominal and ordinal variables are otherwise known as qualitative, and interval and ratio variables, which can either be discrete or continuous, are called quantitative. One thing to remember is that for quantitative variables, it is the measurement that is expressed numerically; but in qualitative variables, the numbers are frequencies of the categories in the variable. For quantitative data, a continuous variable like systolic blood pressure 114 mm Hg is the measurement; for quantitative data, a continuous variable like systolic blood pressure, 114 mm Hg is the measurement. But for qualitative data, a nominal variable like recovered from a disease or dead, the numbers who have died is the count or frequency. Nominal and ordinal variables are measured in categories and, therefore, are also called categorical variables. In experimental research, one variable is manipulated and effects are observed in another variable. The outcome of interest, which changes in response to the intervention, is the dependent variable and the intervention is the independent variable. Hypertension is dependent on factors like salt intake, exercise, and stress. Therefore, hypertension is a dependent variable and salt intake, exercise, and stress are independent variables.

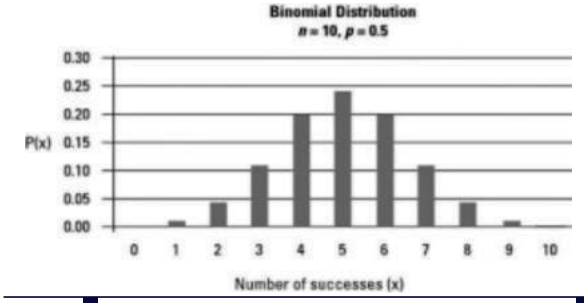
Distribution of Data:

On measuring any variable in a large number of individuals, we call the pattern of values obtained a distribution. In biology, most continuous variables in interval or ratio scale are distributed normally when the number of measurements is fairly large (i.e. more than 30), and the frequency polygon tends to get smoother . A normal distribution is a bell-shaped symmetrical curve where the mean, median, and mode coincide as 50% of the values are above the mean and most of the values are close to the mean. In addition to all these, in a normally distributed data, the mean and variance are not dependent on each other. Even if the mean changes, the proportion of values that lie between the mean ± 1SD (68.2%), mean ± 2SD (95.4%), and mean ± 3SD (99.8%) remains the same in a normal distribution. The name "normal" is a little unfortunate as the distributions, which do not fit into this shape, are in no way abnormal. [7] Contrary to the belief, many variables in medical science may not follow normal distribution, for example, the number of children a family can have or the timing of patients' arrival at the emergency department. Lack of symmetry in the frequency distribution is called skewness. A frequency distribution that has a long tail extending to the right is known as positive skewness or skewed to the right and the one that has a long tail extending to the left is known as negative skewness or skewed to the left . But in the history of statistical methods, the first techniques of inference that were developed and the ones most commonly used, including z-test, Student's t-test, analysis of variance (ANOVA), correlation, and regression, are based on assumptions that the data follow a normal distribution. If the assumption of normality is violated, interpretation and inference may not be reliable or valid. Therefore, it is necessary to assess the normality of data before proceeding with these statistical procedures. Normality of data can be assessed either visually by use of normal plots, the numerical methods which include the skewness and kurtosis coefficients, or by using significance tests. Though visual inspection of data is less reliable, it is preferable that normality be assessed both visually and through normality tests. The normality tests are supplementary to graphical assessment. The frequency distribution (histogram), stem-and-leaf plot, boxplot, P-P plot (probability-probability plot), and Q-Q plot (quantile-quantile plot) are used for checking the normality visually. The most common tests used for the assessment of normality are Kolmogorov-Smirnov (K-S) test, Lilliefors corrected K-S test, Anderson-Darling test, and Shapiro-Wilk test. If the test is significant, the distribution is non-normal. But these tests should be used cautiously for small samples (30 and below), as they have less power for small sample sizes. Of all these tests, Shapiro-Wilk test is the most powerful test for all types of distribution and sample sizes, whereas Kolmogorov-Smirnov test is the least powerful test.

Application and distribution of data


Discrete
Continuous
Bernoulli distribution
Binomial distribution
Beta-binomial distribution
Negative binomial distribution
Poisson distribution
Beta negative binomial distribution
Borel distribution
Champernowne distribution
Hypergeometric distribution
Generalized log-series distribution
Geometric distribution
Logarithmic (series) distribution
Polya–Eggenberger distribution
Skellam distribution
Skew elliptical distribution
Arcsine distribution
Beta distribution
Gaussian/Normal distribution
Logit-normal distribution
Continuous uniform distribution
Rectangular distribution
Irwin–Hall distribution
Bates distribution
Kent distribution
Kumaraswamy distribution
Logarithmic distribution
Marchenko–Pastur distribution
PERT distribution
Reciprocal distribution
Triangular distribution

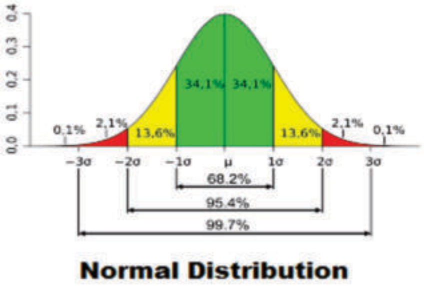
Distribution is a most commonly used concept in research and acquaints with the shape of the data. Most of the time when a researcher collects data, s/he wants to know about the characteristics of the study sample. Let's say, an investigator is interested to know the cancer site of first 100 patients diagnosed in a tertiary care hospital in a specified year. S/he would just select 100 patients, gather information from the lab about the diagnosis, and summarize data. These data might give statistics such as the average age, gender, eating habit, and smoking pattern, number of breast cancer cases and so on. All of these statistics simply describe characteristics of the sample and are therefore called descriptive statistics. The probability distributions are the part of descriptive statistics to describe the shape of the data and possibly predict the probability of an event. Although there are a number of probability distributions as shown in the figure, majorly three distributions are used in medical research studies i.e. Binomial, Poisson and Gaussian/Normal distribution.



Binomial distribution has very wide application and it is important as it allows us to deal with the outcome belongs to two categories such as accepted/rejected, yes/no or male/female. It is a probability model for a discrete outcome with dichotomous nature. It gives you the probability of m successes among n trials of any event. For example, there are 40 kidney cancer patients and we are looking for 5-year survival of 16 patients. The individual patient outcomes are independent and if we assume that the probability of survival is p = 0.20 or 20% for all patients then the required probability will be 0.2%.

Similarly, you may find several options such as what is the probability of at least 16 patients survives, more than 16 patients survives etc.

Poisson distribution describes the behaviour of rare events (with small probabilities) such as patients arriving at an emergency room, decaying radioactive atoms, bank customers coming to their bank, number of suicide cases in adolescence. Let us say your local hospital registered a mean of 2.3 patients arriving at the emergency department on Saturday second half. You can find the probability of exactly four patients arrives on a randomly selected Saturday's second half.



The most important continuous probability distribution is the Gaussian or Normal Distribution. It is a bell-shaped slider and also known as symmetrical distribution. The normal distribution has some very nice properties. If two random variables have a normal distribution, their sum has a normal distribution. In general, all kinds of sums and differences of normal variables have normal distributions.

The normal distribution has only two parameters, the mean and the standard deviation (SD). By definition, about 67% of the values of a normal distribution is within ±1 SD of the mean, and about 95% are within ± 2 SDs. Skewed distributions (non-normal) are not appropriately described with the mean and the SD. The median and the interquartile range are more appropriate for describing non-normally distributed data as most of the time biological data is not normally distributed. Also arithmetic mean should not be used to average normalized numbers. In this case, the geometric mean is the right statistics to average normalized numbers.

Almost 90% statistical inference is based on normal distribution. There are hundreds of statistical tests, and tests will not give accurate results if their assumptions are not met. All parametric tests have an assumption of normally distributed data. So the most common application of normal distribution is to identify whether to select a parametric test or non-parametric test. There are many ways to identify whether data is normally distributed such as histogram, box-plot, outliers, normal quantile plot, Shapiro-Wilk test etc. However, most of the statistician assume data to be normal if SD is less than half of the mean.

Other applications of a normal distribution are to find probabilities from given values of a random variable and to find cut-off values of a random variable from a given probability. The entire theory of linear regression is based on the assumption of normality.

REFERENCES

1. Gravetter FJ, Wallnau LB. Statistics for Behavioural Sciences. 5 th ed. Australia: Wadsworth ThomsonLearning;2000.p.7,583-605,637-56.
2. Norman GR, Steiner DL. Biostatistics: The Bare Essentials. 2 nd ed. London: BC Decker Inc; 2000. p. 2-5.
3. Pagano M, Gauvreau K. Priciple of Biostatistics. 2 nd ed. Australia: Duxbury, Thomson Learning; 2000.p.7-11,38-43.
4. Driscoll P, Lecky F, Crosby M. An Introduction to everyday statistics - 1. J Accid Emerg Med 2000;17:205-11.
5. Bland M. An Introduction to Medical Statistics. 3 rd ed. Oxford: Oxford University Press; 2000. p. 46-62.
6. Altman DG, Bland JM. Statistics notes: The normal distribution. BMJ 1995;310:298.
7. Hill AB, Hill ID. Bradford Hill′s Principles of Medical Statistics. 12 th ed. New Delhi: BI Publications Pvt.Ltd;1993.p.81.
8. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov,

Lilliefors and Anderson-Darlingtests.JOSMA2011;2:21-33.

9. Field A. Discovering Statistics Using SPSS. 3 rd ed. London: SAGE Publications Ltd; 2009. p. 822.
10. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: A guide for non-statisticians. Int JEndocrinolMetab2012;10:486-9. 11.Elliott AC, Woodward WA. Statistical Analysis Quick Reference Guidebook with SPSS Examples.1 st ed.London:SagePublications;2007.
12. Siegel S, Castellan NJ. Nonparametric Statistics for the Behavioral Sciences. 2 nd ed. New York, NY:McGraw-Hill,SiegelandCastellan;1988.p.35. 13.Six Sigma Material. Data Classification. Available from: http://www.six-sigma-material.com/Data-Classification.html. [Last accessed on 2014 Apr 04].