BIOSTATISTICS SERIES

Statistics Corner: Fundamental of Biostatistics

Kamal Kishore[1], Vidushi Jaswal[2]

Reality check

Health researchers routinely collect a subset of data (sample) to study disease characteristics. However, the purpose almost always is to generalize findings to the population. Inferential statistics plays a significant role in the generalization of conclusions from sample to population. The inferential statistics uses probability theory to quantify uncertainty for generating evidence in favor or against the intervention. There are several statistical software for data analysis. However, the lack of fundamental statistical concepts may lead to flawed assumptions and incorrect data analysis. The erroneous statistical analysis precedes both inaccurate reporting and interpretation. Thus, it is essential to understand the fundamentals of statistical jargon before delving into the statistical analysis. In this context, the researcher wants to understand few concepts such as:

- What is the difference between random selection and random allocation?
- What are the differences between standard deviation (SD) and standard error (SE)?
- Are α and *p* value the same or different?
- What are type-I and type-II errors?

Keywords: Hypothesis formulation, *p* value interpretation, Random allocation, Random selection, Type of errors.

Journal of Postgraduate Medicine, Education and Research (2021): 10.5005/jp-journals-10028-1449

Introduction

Academic publication in peer-reviewed journals is prestigious and professionally gratifying. Therefore, graduate students and young faculty are encouraged to publish early in their scholarly careers. The publication of study findings also establishes the credibility of research and researcher among the scientific fraternity. However, there is often a considerable gap in researching and publishing the same in peer-reviewed journals. Literature highlights the numerous barriers to successful publication.[1–3] Besides being scope and novelty, statistical analysis is one of the foremost reasons for the manuscripts' rejection.

The twin hallmark of science is replication and objective evaluation of evidence. Therefore, scientists across the globe carefully plan, document, and execute the studies to generate evidence. The researcher rigorously analyses and evaluates study data before sharing it with the scientific fraternity. Thus, the statistical analysis is a vital ingredient of quality scientific evidence. The statistical issues for data analysis may vary from single to multiple such as non-optimal sample size, unclean and untidy data, inappropriate graphs and tables, flawed assumptions, incorrect statistical tests, wrong conclusions, and interpretations. Thus, many applied researchers collaborate and work with statisticians due to a lack of rigorous statistical training, time, and core subject area priorities.

This manuscript will try to explain the subtle differences in fundamental statistical jargon with an example. Flowchart 1 displays a general list of steps involved in conducting many scientific studies. However, the steps in the figure are not exclusive and exhaustive. Authors need to think and tailor the diagram for their studies and specialization to figure out other issues. Although study design and sample size calculation are at the heart of any research. The discussion about study designs and sample size calculation is beyond the current manuscript's scope and will be dealt with separately.

[1]Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India

[2]Mehr Chand Mahajan DAV College for Women, Chandigarh, India

Corresponding Author: Kamal Kishore, Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India, Phone: +91 9591349768, e-mail: kkishore.pgi@gmail.com

How to cite this article: Kishore K, Jaswal V. Statistics Corner: Fundamental of Biostatistics. J Postgrad Med Edu Res 2021;55(3):149–151.

Source of support: Nil

Conflict of interest: None

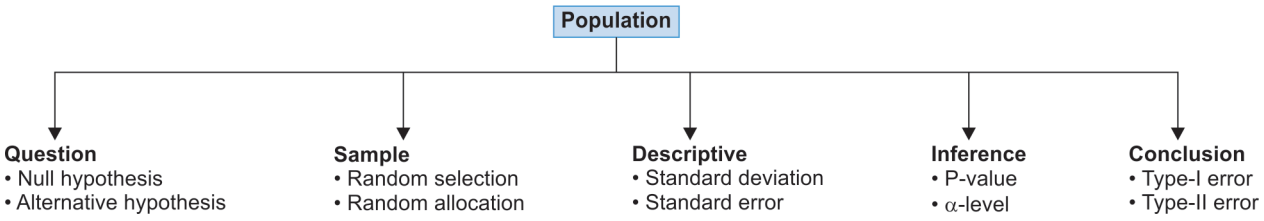
Research Question

A researcher wants to know if there is any evidence of an association between vaccination and clotting, or is vaccination a risk factor for blood clotting? After deciding the research question, the usual first step is to formulate a specific and straightforward statement, also known as a hypothesis.

Hypothesis

A hypothesis is a testable statement that specifies the independent (IV), dependent variables (DV), and relationship (such as mean outcome level, association, correlation) between variables of the study. A hypothesis is a statement or phenomenon observed in the population and is not about the sample. There are two types of hypotheses known as the null and alternate hypotheses. The alternate hypothesis is also known as the researcher's hypothesis or hypothesis of interest. An alternate hypothesis is framed first, and researchers often find compelling reasons to articulate a one-tailed hypothesis before or after data analysis. However, the literature advises against stating a one-tailed hypothesis.[4,5] The theory of hypothesis is based on rejection rather than acceptance

Flowchart 1: General list of the tasks involved in conducting an experiment



of proof. It is almost impossible to prove that all swans are white despite observing the population because only one black swan is enough to disapprove of the theory. Therefore, standard textbooks and journal articles use the phrase "*reject or do not reject the null hypothesis*" rather than accept the hypothesis.

Null Hypothesis (H_0)

The null hypothesis states no difference between the comparison groups, i.e., no difference, no association, or effect. In our example, H_0 is "*There is no significant association (relationship) between vaccination (IV) and blood clots (DV) formation*".

Alternative Hypothesis (H_a or H_1)

An alternate hypothesis states a significant difference between groups, i.e., an association or "effect". The H_a is "*There is a significant (higher or lower effect) association between vaccination and blood clots formation*".

Randomization

Randomized control trials are the gold standard of medical research. Fisher proposed randomization in the 1920s for agricultural experiments, and the same concept but in a different and unrelated fashion was popularized two decades later by Hill in medical research.[6] Randomization is vital for the generalization of the study findings. Randomization has a particular meaning contrary to the belief that it is haphazard. It is a characteristic in which each participant has an equal and known chance of being selected. Random selection is possible when the researcher knows the sampling frame and has access to the population. The typical example of a random sample is the lottery method, house allocation, etc. Some researchers still use a lottery, coin tossing, dice, or odd-even days, but these are non-verifiable techniques and leave the chance to introduce researcher bias. Researchers involved in medical sciences typically do not have access to sampling frames. Researchers make a crucial assumption that participants coming to health facilities represent the diseased people in the community. Thus, participants are assessed based on inclusion criteria and randomly allocated to intervention and control groups.

Random Selection

Random selection is selecting a subset of participants from a list of all the participants. The researcher needs to know the sampling frame for selecting participants. Probability sampling techniques such as simple random sampling, stratified random sampling, and cluster sampling are typical examples of random selection. The principle of random selection will assume that the population includes COVID-19 positive patients who are vaccinated and non-vaccinated. The participants will be randomly selected from both vaccinated and non-vaccinated COVID infected groups to assess each group's frequency of blood clotting. However, in medical sciences, it is rare to have complete information such as the total number of COVID-19 patients.

Random Allocation

Random allocation allocates participant reporting in the health facility after assessing inclusion and exclusion criteria. In our situation, the researcher will assign each COVID positive patient to either the intervention group (drug in controversy) or control group (placebo/other drugs not in debate). Subsequently, the infection rate in both groups will be compared to find the association between vaccination and blood clots formation. Typically, researchers in health settings use simple randomization, block randomization, and stratify randomization to allocate participants to intervention and control groups. Block randomization is ideal for distributing equal sample sizes among groups for a small sample size. Stratify randomization is superior when known confounding factors such as age and sex affect the outcome and need to be balanced between groups.

Variability

Data analysis succeeds data collection. Descriptive statistics is the first natural step in data analysis as it defines the characteristics of data. It is becoming friendly with data regarding data range and restrictions, assumptions met and unmet, and selecting appropriate reporting standards. In the current times, almost anyone can analyze data with the help of available statistical software. However, the software will not tell the researcher about data cleaning, assumptions validation, and wrong statistical test selection. Software often gives multiple outputs such as standard deviation (SD) and standard error (SE). However, researchers not friendly with statistical jargon tend to confuse and report SD in place of SE or *vice versa*.

Standard Deviation

The SD of a variable is a quantity that measures the variability of a set of observations around the mean value. It explains the variability in measurements for a variable (sample); demographic tables reporting characteristics of the sample report SD.

Standard Error

The quantity measures the uncertainty in estimating a population parameter from a sample. It is the average deviation of each sample's estimate from the population parameter. Therefore, it is reported primarily with inferential statistics such as odds ratio (OR), risk ratio, hazard ratio, and regression coefficients while generalizing the sample findings to the population.

p value Conundrum

Inferential statistics is a crucial final step in data analysis. The fundamental idea of inductive reasoning dominates the majority of biomedical research. In inductive reasoning, the researcher makes an assertion and validates it in a sample. Subsequently, researchers generalize the argument to the population after testing of hypotheses. The researcher uses the *p* value approach to declare the significance of importance. However, many researchers find it challenging to interpret and discriminate between *p* value and α.

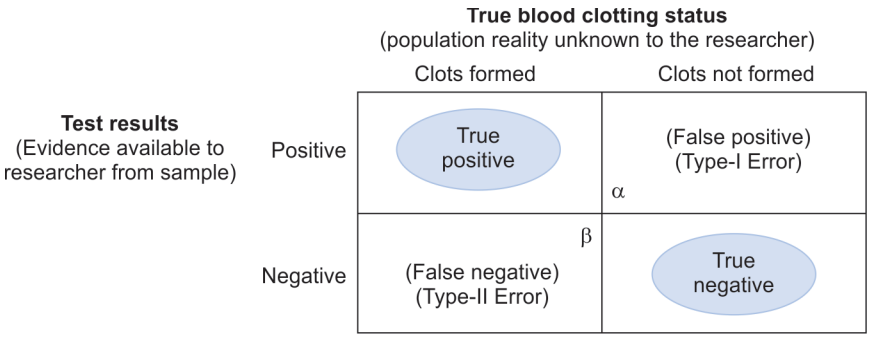


Fig. 1: The paradigm of inductive reasoning and type of errors

p value

The *p* value is the probability of obtaining the observed (difference of means or proportions between the groups) or more extreme difference, given the null hypothesis is true. Assume researcher obtain OR = 2 with $p = 0.04$ after analyzing sample data. If we repeat our experiment multiple times under similar conditions, we can expect four times out of 100 to see an OR = 2. It can happen under two situations:

- The null hypothesis is correct, and the researcher obtains a small *p* value due to unusual or wrong sample selection.
- The null hypothesis is incorrect, and a small *p* value is not obtained due to unusual or wrong sample selection.

α-value

An α-value (0.05 or 0.01 or 0.001) is an arbitrary value selected by the researcher at the study's beginning to reject or not reject a null hypothesis. An α-value is a comparator against which the *p* value is compared. The usual rejection rule for the null hypothesis is when $p < \alpha$. In other words, the results are unlikely by chance if $p < \alpha$. The researcher concludes that there is sufficient evidence to reject the null hypothesis of no difference. α-value is always selected before data analysis.

Types of Errors

The interpretation and conclusion of study results from data analysis is a vital step. However, due to the inductive nature of evidence generation, there is always a possibility of a wrong conclusion, no matter how so ever small it is. Researchers call it type-I and type-II errors. Figure 1 displays the situation leading to 2-type of errors. Scientists usually call type-I error as serious among two errors and thus give it precedence over type-II error. The typical type-I error value used by researchers are 0.05, 0.01, and 0.001. It is impossible to know whether researchers have committed type-I and type-II errors as the same are conceptual. However, researchers can minimize and control both the errors after considering their seriousness by increasing or decreasing sample size.

Type-I Error

Rejecting (COVID-19 vaccination is associated with clots formation) a null hypothesis (COVID-19 vaccination is not associated with clots formation) when it should not be rejected. It means there is no association between COVID-19 vaccination and clots formation in the population (unknown reality). Still, the researcher infers from the sample (collected data) that there is a statistically significant association. It is also known as false-positive (FP) results. It can occur either due to erroneous rejection of the null hypothesis or the researcher obtaining statistically significant but clinically non-relevant results.

Type-II Error

Not rejecting a null hypothesis when it should be rejected. The researcher infers from the sample that there is no significant association between COVID-19 vaccination and clots formation. Although the association is present (unknown reality) in the population. The researchers consider type-II error a less severe error than type-I error and are usually selected or fixed after selecting type-I. It is known as a false-negative (FN) result. It can occur either due to the null hypothesis's erroneous non-rejection or obtain clinically significant differences but do not attain statistical significance. Typically type-II errors arise due to the small sample size.

Conclusion

Statistics is the Achilles' heel of many applied researchers. Despite inadequate training, many people self-analyze and report statistical analysis due to the wide availability of statistical software. However, when it comes to publication in peer-reviewed journals, finer details make a lot of difference. Therefore, it is ideal for collaborating with statisticians while designing the study. However, many adventurous researchers will take in their stride to learn and perform statistical analysis independently. The current article explained few fundamental statistical tricks that will help both adventure seekers and naïve alike.

Acknowledgments

We acknowledge Dr Manoj Jaiswal and Meenakshi Sharma for their valuable time and inputs to improve the readability of the article.

References

1. Ali J. Manuscript rejection: causes and remedies. J young Pharm JYP 2010;2(1):3. DOI: 10.4103/0975-1483.62205.
2. Ehara S, Takahashi K. Reasons for rejection of manuscripts submitted to AJR by international authors. Am J Roentgenol [Internet] 2007;188(2):W113–W116. Available from: https://doi.org/10.2214/AJR.06.0448.
3. Eassom H. 9 Common Reasons for Rejection [Internet]. [cited 2021 Jun 1]. Available from: https://www.wiley.com/network/researchers/submission-and-navigating-peer-review/9-common-reasons-for-rejection.
4. Phillips A, Haudiquet V. ICH E9 guideline 'Statistical principles for clinical trials': a case study. Stat Med 2003;22(1):1–11. DOI: 10.1002/sim.1328.
5. ICH E9 Expert Working Group. Statistical Principles for Clinical Trials: ICH Harmonized Tripartite Guideline, https://database.ich.org/sites/default/files/E9_Guideline.pdf (1998). p. 25.
6. Armitage P. Fisher, Bradford hill, and randomization. Int J Epidemiol [Internet] 2003;32(6):925–928. Available from: 10.1093/ije/dyg286.