Statistical Bias

See online here

The article aims to describe the concept of information bias in the experimental research studies and to throw light on the issues concerning the bias, the types of bias, and the methods of disposal of these errors.

Introduction to Statistical Bias

The ultimate aim of epidemiology is to investigate the cause of diseases and it is one of the important fields of medicine. The health of the human is subjected to the adversities caused by different viruses, bacteria, and microorganisms, which basically are the pathogens disrupting the milieu of the body.

In addition to the microorganisms, there are many chronic diseases like diabetes mellitus and autoimmune diseases, which also cause significant mortality and morbidity. The aim of epidemiology and biostatistics is to determine the evidence for the cause, the routes of dissemination, the frequency of occurrence of particular diseases, the consequences of a disease, and the different therapeutics available to target the disease.

Epidemiology studies the causes of the disease (specifically the pathogens responsible for the disease), the effects, the cure for the disease, and the methods for improving people's lives overall (targeting for better health, which will lead to a peaceful society).

Epidemiology is based on the combination of 3 Greek words:

1. *Epi* refers to ‘among/upon’.
2. *Demos* relates to ‘study of population’.
3. *Logos* relates to ‘scientific research (study)’.

Epidemiology, as a science, studies the influence of pathogens on individuals suffering from the disease, caused by the pathogens and their drastic effects on the human body. In many of the diseases, it is very difficult to establish a causative link for the disease to occur. In these cases, the epidemiology plays a key role in determining the causative role of a pathogen pertaining to the particular disease.

In addition, it also determines the rate at which the disease spreads over a locality, the distribution pattern, and the locations over which the disease spreads predominantly. All these in turn, again help in determining the unexplored causal elements.

All the above-mentioned features of epidemiology and biostatistics are carried out by taking into consideration the pattern and the details of the disease and then based on these observations reaching a conclusion.

The sampling and experimentations are also taken into account. The experimentations, in turn, involve a diverse array of functions like data collection to analyze the collected data and based on the conclusions, hypothesizing the theories behind the disease.

Bias

A bias is an inclination of a person to hold or present a perspective (which in real life may be true or false). It is someone’s natural tendency to act or feel in a particular way based on their inclination.

When we use bias in the field of statistics, it refers to an error. Statistical error is the error that one cannot correct by repeating the process again and again, but by taking the average of the results.

Example

The aim of our experiment is to determine the area of a mall. The actual area of the mall is approx. 150 yards. Now let us repeat the experiment and consider the area of the mall. The aim is to see how much it varies from the actual area? If you have a random survey and ask people “what is the area of that mall?”, the answer remains “140 yards” from anyone answering the survey, which indicates that the answer will always have an error of 10 yards, and if we squared, the answer is always 100 (never considering the squared errors of the units).

Now, let’s assume another example, where the Americans are certain about the area, with a mean of 90 yards, and a standard deviation of 10 m. Now if you randomly poll two different people, and one answer is “90 yards”, and the other says “70 yards”, that means the first answer has an error of -10 yards and the second contains an error of -30 yards.

Notice that for the first case, the perturbation leads the person towards correction. However, the second answer was even farther away from the actual answer. Thus, the perturbation can have an influence on both sides of the coin (both take our result towards the correction or take it further away).

Selection Bias

Selection bias is a type of error that involves the selection of a specific population related to the trial or in the case of surveys and observation studies, the population from which we get the view.

Selection bias is an unfair type of selection. It results when all the participants are not equally interested in the experiment; i.e. some of the participants are willing, and others are forced to participate.

Selection bias occurs at the following stages:

- The stage of allocation of the participants.
- Keeping the participants engaged in the activity.
- Along the course of the experiment.

Selection bias involves the following persons:

- Self-selected participants
- Selection of samples to propose the hypothesis for the respective experiment.

Sources of selection bias

During randomization

- Subversion of randomization due to lack of allocation concealment.

After randomization

- Attrition (reducing the strength or effectiveness of something through effective attack or pressure).

Types of selection bias

There are following types of selection bias:

- Sampling bias
- Time interval
- Exposure
- Data
- Studies
- Attrition
- Observer selection

Sampling bias

Sampling bias has resulted because of a non-random sample of participants; i.e. some of the participants are involved completely, while others are forced to get involved, which results in the sampling bias.

There are methods like the random generation sequencer used to select the sample. It should also be remembered that the laws of inferential statistics can be applied and hold true only when the randomization was done properly. It is a subtype of selection bias, but some researchers consider it as a separate type of bias.

A major difference between selection bias and sampling bias is that selection bias involves internal validity, i.e. the similarities or the differences found in the sample, while

sampling bias involves external validity, i.e. the ability for generalization of the result to the remaining population.

Examples of sampling bias:

- Self-selection
- Pre-selection of trial participants
- Discounting trial subjects that didn't undergo completion of the study.
- Migration bias by omitting subjects who were recently added or removed from the study.

Time interval

Early cessation of a trial at a particular time, when its results reinforce the desired conclusion. One more condition when the trial may be terminated prematurely is when the adverse effect is of severe nature.

A trial may be terminated early at an uttermost value, but the resulting extreme value will have a large 'variance' even if the 'mean' of all the variables are similar.

Exposure

Clinical susceptibility bias: when the person gets affected with one disease and if that particular disease increases the vulnerability of the second disease then clinical susceptibility bias occurs (in those cases, the treatment of the first disease can erroneously appear to make the person vulnerable to the second disease). For example, "postmenopausal syndrome" makes somebody vulnerable for "endometrial cancer". Hence, estrogen taken for postmenopausal syndrome may result in a higher likelihood of endometrial cancer.

Protopathic bias: when nursing for the first symptom of a disease or another outcome appears to cause the outcome. It can be alleviated by "lagging," i.e. prohibition of exposure that occurred before diagnosis.

Indication bias: a potential mess between cause and effect when exposure is relying on indication, e.g., a treatment was given to a person with a high risk of getting a disease, potentially causing a predominance of treated people among those acquiring the disease. This may cause an erroneous appearance of the disease.

Data

- Partitioning data and then analyzing tests designed for blindly chosen subdivisions.
- Preference (cherry-picking)
    - Rejection
    - Post hoc: Amendment of the data inclusion based on temporary or subjective reasons including

Studies

Designation of the studies involved in the meta-analysis. A proper meta-analysis should contain all the available evidence on the subject of the interest. The omission of the key studies would falsely alter the results.

Repetition of the experiments and reporting of only the most favorable results; this is commonly done.

Demonstrating only the most expressive results of data fishing.

Another method is the data dredging, where all attempts are made to do multiple statistical analyses (this increases the error rate of the study). It may be done as a part of the study with the normal intention or may be carried out to determine which item/group in the comparative analysis is of statistical significance. The statistical test is to be mentioned clearly in the initial protocol and is to be followed in order to avoid this.

Attrition

- The process of gradually reducing the size of work or assignment through pressure.
- Attrition bias results are given by attrition. It includes dropout, non-response departure, and protocol divergence.

Observer selection

Data are selected not only for study design and measurement but also by the important prerequisite that there has to be someone doing some study about it.

Examples of selection bias:

- Bias due to the non-implementation or improper implementation of the allocation concealment.
- Randomized control trial (RCT) on thrombolysis with alternating day concealment.
- Bias due to attrition.
- RCT comparing medical versus surgical management of the cerebrovascular disease.
- Some disease circumstances can influence the circulation of blood to the brain (atherosclerosis can cause cerebrovascular disease).

A good researcher should have methods to overcome the shortcomings resulting from the selection bias; i.e one is not exposing some participants for a certain duration and others for a different duration of time. The researcher will explain the possible errors that can result in his report or survey.

As explained above, it is very essential to follow the randomization principles in the study. Ultimately, selection bias is unavoidable. So, the researcher should try to minimize its effects and quote important errors.

Information Bias

Also known as observational bias/scrutinization bias, information bias relates to the bias which arises due to an error in the measurement process.

Information bias is a cognitive bias (systematic deviation from norms), which refers to the distorted evaluation or data analyzed, which in turn relates to non-differential or differential misclassification.

Information bias includes:

- Classification error
- Differential and non-differential bias
- Direction of bias
- Misclassification of covariables

Classification/measurement error

The sources for the classification/measurement error include:

- Manager
- Instrumentation involved in the experiment.
- Experimentation environment (laboratory)
- Questionnaire (in the case of a survey)
- Researcher
- Participants

Differential and non-differential bias

Non-differential misclassification

This includes errors that are approximately the same if we compare two or more groups.

Differential misclassification

Differential misclassification is an error in which the frequency is relatively higher in one of the groups being studied.

In real life, the non-differential misclassification of exposure is much more epidemic as compared to the differential.

Direction of bias:

- Upward
- Downward
- Towards the null
- Away from null

NULL is ‘0’ for differences and ‘1’ for ratios.

The direction of the bias is usually unpredictable.

Response bias

It is a terminology for the capability of a person to answer a questionnaire ambiguously.

Reporting bias

It refers to the results from particular diagnostic phenomena or a particular type of instrument. Reporting bias can be a cause of great concern, especially in the case of observational studies.

Detection bias

It is one of the most complicated phases in the analysis because the outcome may not be systemized.

Confounding bias

It refers to the state of an affair in which association between the subjection and the outcome is unclear due to the presence of another variable. Unless the effect of the confounding variable is nullified and analyzed, the proper causation cannot be

established.

Confounding can be of two types–positive and negative.

Positive confounding

It is a type of confounding in which the researcher observes that the association is biased away from the null.

Negative confounding

It is a type of confounding in which the researcher observes that the association is biased towards the null.

Confounder

The confounder is an irrelevant variable to that of causation that partially or completely affects the causative analysis in determining the risk factor of the disease. The results become inaccurate due to the presence of a confounder.

For a variable to be declared as a confounder, it needs to satisfy the following conditions:

- Probability factor of the disease
- It is affiliated with a hypothetical risk factor
- Not in causal mechanism between subjection and disease.

The last two conditions can be tested using the appropriate statistical tests on the data, whereas the first condition is more biological and conceptual. One of the examples of the statistical tests is the application of the hierarchical logistic regression for determining the confounding factor among a group of variables.

Potential confounders can be determined by our:

- Knowledge
- Experience
- Conditions

Causation

Causation is important in epidemiology. **Though there is no exact quote/definition for causation, the following 5 categories can be portrayed:**

1. Generation
2. Necessary and sufficient
3. Sufficient-component
4. Counter of facts
5. Probabilistic

To get a hold on the proposed characteristics of causation (which in turn will be the right definition, too), supremacy and weaknesses of all these categories are to be scrutinized by the reader.

Two classes–era and counterfactual, are thoroughly important in the meaning of causation. The essential and adequate definition assess that every one of the causes is deterministic. The adequate part clarifies that the variable is indispensable and by itself

can cause the disease. Henceforth, as per both the perspectives, overwhelming smoking can be cited as a reason for lung disease, just when the nearness of the obscure deterministic variable is suspected.

Some of the causative factors though essential, may not be adequate to cause disease; there may be additional factors required to cause the full-blown disease.

Correlation

Correlation tells us the relationship between the 2 variables. The relationship between the 2 variables could be present or absent. When the relationship is present, it can be either positive or negative.

The correlation coefficient (r) represents the strength of the correlation. There are 2 methods of determining the correlation coefficient; the choice of the method depends on the parametricity of the data in hand. If the data is parametric then generally the Pearson correlation coefficient is used; when the data is nonparametric then the Spearman correlation coefficient is used.

The relationship between the 2 correlated variables could be a causal association or it may be in any other context taken with regards to the experiment.

Positive correlation

In the case of positive correlation, the dependent and independent variables fluctuate together, i.e. if one increases, the other variable also increases, and vice versa.

Negative correlation

For a negative correlation, the dependent and independent variables fluctuate opposite from one another, i.e. if one increases, the other decreases, and vice versa.

Conclusion

The epidemiology and biostatistics are deeply interlinked and depend on each other in their principles. The rules of both disciplines are indispensable to any researcher and should be followed with full regard in their study. The biases mentioned in this document should be noted and taken care of by the researcher.

References

DiCenso, A. G. (2014). Evidence-based nursing: A guide to clinical practice. Elsevier Health Sciences.

Kramer, M. S. (2012). Clinical epidemiology and biostatistics: a primer for clinical investigators and decision-makers. Springer Science & Business Media.

Nieuwenhuijsen, M. J. (2015). Exposure assessment in environmental epidemiology. Oxford University Press, USA.

Walsh, M. D. (2013). Relationship between Intraoperative Mean Arterial Pressure and Clinical Outcomes after Noncardiac SurgeryToward an Empirical Definition of Hypotension. The Journal of the American Society of Anesthesiologists, 119(3), 507-515.

Wassertheil-Smoller, S. &. (2015). Biostatistics and epidemiology: a primer for health and

biomedical professionals. Springer.

Wassertheil-Smoller, S. (2013). Biostatistics and epidemiology: a primer for health professionals. Springer Science & Business Media.

Notes