Selection bias

From Wikipedia, the free encyclopedia

Selection bias is the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed.[1] It is sometimes referred to as the **selection effect**. The phrase "selection bias" most often refers to the distortion of a statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

Contents

Types

There are many types of possible selection bias, including:

Sampling bias

Sampling bias is systematic error due to a non-random sample of a population,[2] causing some members of the population to be less likely to be included than others, resulting in a biased sample, defined as a statistical sample of a population (or non-human factors) in which all participants are not equally balanced or objectively represented.[3] It is mostly classified as a subtype of selection bias,[4] sometimes specifically termed *sample selection bias*,[5][6][7] but some classify it

as a separate type of bias.[8]

A distinction of sampling bias (albeit not a universally accepted one) is that it undermines the external validity of a test (the ability of its results to be generalized to the rest of the population), while selection bias mainly addresses internal validity for differences or similarities found in the sample at hand. In this sense, errors occurring in the process of gathering the sample or cohort cause sampling bias, while errors in any process thereafter cause selection bias.

Examples of sampling bias include self-selection, pre-screening of trial participants, discounting trial subjects/tests that did not run to completion and migration bias by excluding subjects who have recently moved into or out of the study area.

Time interval

- Early termination of a trial at a time when its results support a desired conclusion.
- A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.

Exposure

- *Susceptibility bias*
  - *Clinical susceptibility bias*, when one disease predisposes for a second disease, and the treatment for the first disease erroneously appears to predispose to the second disease. For example, postmenopausal syndrome gives a higher likelihood of also developing endometrial cancer, so estrogens given for the postmenopausal syndrome may receive a higher than actual blame for causing endometrial cancer.[9]
  - *Protopathic bias*, when a treatment for the first symptoms of a disease or other outcome appear to cause the outcome. It is a potential bias when there is a lag time from the first symptoms and start of treatment before actual diagnosis.[9] It can be mitigated by lagging, that is, exclusion of exposures that occurred in a certain time period before diagnosis.[10]
  - *Indication bias*, a potential mix up between cause and effect when exposure is dependent on indication, e.g. a treatment is given to people in high risk of acquiring a disease, potentially causing a preponderance of treated people among those acquiring the disease. This may cause an erroneous appearance of the treatment being a cause of the disease.[11]

Data

- Partitioning (dividing) data with knowledge of the contents of the partitions, and then analyzing them with tests designed for blindly chosen partitions.
- Post hoc alteration of data inclusion based on arbitrary or subjective reasons, including:
  - Cherry picking, when specific subsets of data are chosen to support a conclusion (e.g. citing examples of plane crashes as evidence of airline flight being unsafe, while ignoring the far more common example of flights that complete safely. See: Availability heuristic)
  - Rejection of "bad" data on (1) arbitrary grounds, instead of according to previously stated or generally agreed criteria or (2) discarding "outliers" on statistical grounds that fail to take into account important information that could be derived from "wild" observations.[12]

Studies

- Selection of which studies to include in a meta-analysis (see also combinatorial meta-analysis).
- Performing repeated experiments and reporting only the most favorable results, perhaps relabelling lab records of other experiments as "calibration tests", "instrumentation errors" or "preliminary surveys".
- Presenting the most significant result of a data dredge as if it were a single experiment (which is logically the same as the previous item, but is seen as much less dishonest).

Attrition

Attrition bias is a kind of selection bias caused by attrition (loss of participants),[13] discounting trial subjects/tests that did not run to completion. It is closely related to the survivorship bias, where only the subjects that "survived" a process are included in the analysis. It includes *dropout*, *nonresponse* (lower response rate), *withdrawal* and *protocol deviators*. It gives biased results where it is unequal in regard to exposure and/or outcome. For example, in a test of a dieting program, the researcher may simply reject everyone who drops out of the trial, but most of those who drop out are those for whom it was not working. Different loss of subjects in intervention and comparison group may change the characteristics of these groups and outcomes irrespective of the studied intervention.[13]

Observer selection

Data are filtered not only by study design and measurement, but by the necessary precondition that there has to be someone doing a study. In situations where the existence of the observer or the study is correlated with the data observation selection effects occur, and anthropic reasoning is required.[14]

An example is the past impact event record of Earth: if large impacts cause mass extinctions and ecological disruptions precluding the evolution of intelligent observers for long periods, no one will observe any evidence of large impacts in the recent past (since they would have prevented intelligent observers from evolving). Hence there is a potential bias in the impact record of Earth.[15] Astronomical existential risks might similarly be underestimated due to selection bias, and an anthropic correction has to be introduced.[16]

Mitigation

In the general case, selection biases cannot be overcome with statistical analysis of existing data alone, though Heckman correction may be used in special cases. An assessment of the degree of selection bias can be made by examining correlations between exogenous (background) variables and a treatment indicator. However, in regression models, it is correlation between *unobserved* determinants of the outcome and *unobserved* determinants of selection into the sample which bias estimates, and this correlation between unobservables cannot be directly assessed by the observed determinants of treatment.[17]

Related issues

Selection bias is closely related to:

- publication bias or reporting bias, the distortion produced in community perception or meta-analyses by not publishing uninteresting (usually negative) results, or results which go against the experimenter's prejudices, a sponsor's interests, or community expectations.
- confirmation bias, the distortion produced by experiments that are designed to seek confirmatory evidence instead of trying to disprove the hypothesis.
- exclusion bias, results from applying different criteria to cases and controls in regards to participation eligibility for a study/different variables serving as basis for exclusion.

See also

- Berkson's paradox
- Black Swan theory
- Cherry picking (fallacy)
- Funding bias
- List of cognitive biases
- Reporting bias
- Sampling bias
- Self-fulfilling prophecy
- Publication bias
- Participation bias
- Survivorship bias

References

1. Dictionary of Cancer Terms → selection bias (http://www.cancer.gov/dictionary?CdrID=44087). Retrieved on September 23, 2009.
2. Medical Dictionary - 'Sampling Bias' (http://www.medilexicon.com/medicaldictionary.php?t=10087) Retrieved on September 23, 2009
3. TheFreeDictionary → biased sample (http://medical-dictionary.thefreedictionary.com/Sample+bias). Retrieved on 2009-09-23. Site in turn cites: Mosby's Medical Dictionary, 8th edition.
4. Dictionary of Cancer Terms → Selection Bias (http://medical.webends.com/kw/Selection%20Bias). Retrieved on September 23, 2009.
5. Ards, Sheila; Chung, Chanjin; Myers, Samuel L. (1998). "The effects of sample selection bias on racial differences in child abuse reporting". *Child Abuse & Neglect*. **22** (2): 103–115. doi:10.1016/S0145-2134(97)00131-2. PMID 9504213.
6. Cortes, Corinna; Mohri, Mehryar; Riley, Michael; Rostamizadeh, Afshin (2008). "Sample Selection Bias Correction Theory" (PDF). *Algorithmic Learning Theory*. **5254**: 38–53. doi:10.1007/978-3-540-87987-9_8.
7. Cortes, Corinna; Mohri, Mehryar (2014). "Domain adaptation and sample bias correction theory and algorithm for regression" (PDF). *Theoretical Computer Science*. **519**: 103–126. doi:10.1016/j.tcs.2013.09.027.
8. Fadem, Barbara (2009). *Behavioral Science*. Lippincott Williams & Wilkins. p. 262. ISBN 978-0-7817-8257-9.
9. Feinstein AR; Horwitz RI (November 1978). "A critique of the statistical evidence associating estrogens with endometrial cancer". *Cancer Res*. **38** (11 Pt 2): 4001–5. PMID 698947.
10. Tamim H; Monfared AA; LeLorier J (March 2007). "Application of lag-time into exposure definitions to control for protopathic bias". *Pharmacoepidemiol Drug Saf*. **16** (3): 250–8. doi:10.1002/pds.1360. PMID 17245804.
11. Matthew R. Weir (2005). *Hypertension (Key Diseases) (Acp Key Diseases Series)*. Philadelphia, Pa: American College of Physicians. p. 159. ISBN 1-930513-58-5.
12. Kruskal, William H. (1960). "Some Remarks on Wild Observations". *Technometrics*. **2** (1): 1–3. doi:10.1080/00401706.1960.10489875.
13. Jüni, P.; Egger, Matthias (2005). "Empirical evidence of attrition bias in clinical trials". *International Journal of Epidemiology*. **34** (1): 87–88. doi:10.1093/ije/dyh406.
14. Bostrom, Nick (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York: Routledge. ISBN 0-415-93858-9.
15. Ćirković, M. M.; Sandberg, A.; Bostrom, N. (2010). "Anthropic Shadow: Observation Selection Effects and Human Extinction Risks". *Risk Analysis*. **30** (10): 1495. doi:10.1111/j.1539-6924.2010.01460.x.

16. Tegmark, M.; Bostrom, N. (2005). "Astrophysics: Is a doomsday catastrophe likely?". *Nature*. **438** (7069): 754. doi:10.1038/438754a. PMID 16341005.

17. Heckman, J. J. (1979). "Sample Selection Bias as a Specification Error". *Econometrica*. **47**: 153. doi:10.2307/1912352. JSTOR 1912352.

Categories: Sampling (statistics) | Bias | Scientific method | Causal inference