10. SAMPLE BIAS, BIAS OF SELECTION AND DOUBLE-BLIND

10.1 SAMPLE BIAS: In statistics, **sampling bias** is a bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others. It results in a**biased sample**, a non-random sample[1] of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected.[2] If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling.

Medical sources sometimes refer to sampling bias as **ascertainment bias**.[3][4] Ascertainment bias has basically the same definition,[5][6] but is still sometimes classified as a separate type of bias

Types of sampling bias

- Selection from a **specific real area**. For example, a survey of high school students to measure teenage use of illegal drugs will be a biased sample because it does not include home-schooled students or dropouts. A sample is also biased if certain members are underrepresented or overrepresented relative to others in the population. For example, a "man on the street" interview which selects people who walk by a certain location is going to have an overrepresentation of healthy individuals who are more likely to be out of the home than individuals with a chronic illness. This may be an extreme form of biased sampling, because certain members of the population are totally excluded from the sample (that is, they have zero probability of being selected).
- **Self-selection** bias, which is possible whenever the group of people being studied has any form of control over whether to participate. Participants' decision to participate may be correlated with traits that affect the study, making the participants a non-representative sample. For example, people who have strong opinions or substantial knowledge may be more willing to spend time answering a survey than those who do not. Another example is online and phone-in polls, which are biased samples because the respondents are self-selected. Those individuals who are highly motivated to respond, typically individuals who have strong opinions, are overrepresented, and individuals that are indifferent or apathetic are less likely to respond. This often leads to a polarization of responses with extreme perspectives being given a disproportionate weight in the summary. As a result, these types of polls are regarded as unscientific.
- **Pre-screening** of trial participants, or **advertising** for volunteers within particular groups. For example a study to "prove" that smoking does not affect

fitness might recruit at the local fitness center, but advertise for smokers during the advanced aerobics class, and for non-smokers during the weight loss sessions.

- **Exclusion** bias results from exclusion of particular groups from the sample, e.g. exclusion of subjects who have recently migrated into the study area (this may occur when newcomers are not available in a register used to identify the source population). Excluding subjects who move out of the study area during follow-up is rather equivalent of dropout or non response, a selection bias in that it rather affects the internal validity of the study.
- **Healthy user bias**, when the study population is likely healthier than the general population, e.g. workers (i.e. someone in ill-health is unlikely to have a job as manual laborer).
- **Berkson's fallacy**, when the study population is selected from a hospital and so is less healthy than the general population. This can result in a spurious negative correlation between diseases: a hospital patient without diabetes is *more* likely to have another given disease such as cholecystis, since they must have had some reason to enter the hospital in the first place.
- **Overmatching**, matching for an apparent confounder that actually is a result of the exposure. The control group becomes more similar to the cases in regard to exposure than the general population.

Symptom-based sampling

The study of medical conditions begins with anecdotal reports. By their nature, such reports only include those referred for diagnosis and treatment. A child who can't function in school is more likely to be diagnosed with dyslexia than a child who struggles but passes. A child examined for one condition is more likely to be tested for and diagnosed with other conditions, skewing comorbidity statistics. As certain diagnoses become associated with behavior problems or intellectual disability, parents try to prevent their children from being stigmatized with those diagnoses, introducing further bias. Studies carefully selected from whole populations are showing that many conditions are much more common and usually much milder than formerly believed.

Symptom-based sampling

The study of medical conditions begins with anecdotal reports. By their nature, such reports only include those referred for diagnosis and treatment. A child who can't function in school is more likely to be diagnosed with dyslexia than a child who struggles but passes. A child examined for one condition is more likely to be tested for and diagnosed with other conditions, skewing comorbidity statistics. As certain diagnoses become associated with behavior problems or intellectual

disability, parents try to prevent their children from being stigmatized with those diagnoses, introducing further bias. Studies carefully selected from whole populations are showing that many conditions are much more common and usually much milder than formerly believed.

The caveman effect

An example of selection bias is called the "caveman effect". Much of our understanding of prehistoric peoples comes from caves, such as cave paintings made nearly 40,000 years ago. If there had been contemporary paintings on trees, animal skins or hillsides, they would have been washed away long ago. Similarly, evidence of fire pits, middens, burial sites, etc. are most likely to remain intact to the modern era in caves. Prehistoric people are associated with caves because that is where the data still exists, not necessarily because most of them lived in caves for most of their lives.

Problems caused by sampling bias

A biased sample causes problems because any statistic computed from that sample has the potential to be consistently erroneous. The bias can lead to an over- or underrepresentation of the corresponding parameter in the population. Almost every sample in practice is biased because it is practically impossible to ensure a perfectly random sample. If the degree of underrepresentation is small, the sample can be treated as a reasonable approximation to a random sample. Also, if the group that is underrepresented does not differ markedly from the other groups in the quantity being measured, then a random sample can still be a reasonable approximation.

The word bias has a strong negative connotation. Indeed, biases sometimes come from deliberate intent to mislead or other scientific fraud. In statistical usage, bias merely represents a mathematical property, no matter if it is deliberate or either unconscious or due to imperfections in the instruments used for observation. While some individuals might deliberately use a biased sample to produce misleading results, more often, a biased sample is just a reflection of the difficulty in obtaining a truly representative sample.

Some samples use a biased statistical design which nevertheless allows the estimation of parameters. The U.S. National Center for Health Statistics, for example, deliberately oversamples from minority populations in many of its nationwide surveys in order to gain sufficient precision for estimates within these groups. These surveys require the use of sample weights (see later on) to produce proper estimates across all ethnic groups. Provided that certain conditions are met

(chiefly that the sample is drawn randomly from the entire sample) these samples permit accurate estimation of population parameters.

Historical examples

Example of biased sample, claiming as of June 2008, that only 54% of web browsers (Internet Explorer) in use do not pass theAcid2 test. The statistics are from visitors to one website comprising mostly web developers.[15]

A classic example of a biased sample and the misleading results it produced occurred in 1936. In the early days of opinion polling, the American *Literary Digest* magazine collected over two million postal surveys and predicted that the Republican candidate in the U.S. presidential election, Alf Landon, would beat the incumbent president, Franklin Roosevelt, by a large margin. The result was the exact opposite. The Literary Digest survey represented a sample collected from readers of the magazine, supplemented by records of registered automobile owners and telephone users. This sample included an over-representation of individuals who were rich, who, as a group, were more likely to vote for the Republican candidate. In contrast, a poll of only 50 thousand citizens selected by George Gallup's organization successfully predicted the result, leading to the popularity of the Gallup poll.

Another classic example occurred in the 1948 presidential election. On election night, the Chicago Tribune printed the headline *DEWEY DEFEATS TRUMAN*, which turned out to be mistaken. In the morning the grinning president-elect, Harry S. Truman, was photographed holding a newspaper bearing this headline. The reason the Tribune was mistaken is that their editor trusted the results of a phone survey. Survey research was then in its infancy, and few academics realized that a sample of telephone users was not representative of the general population. Telephones were not yet widespread, and those who had them tended to be prosperous and have stable addresses. (In many cities, the Bell System telephone directory contained the same names as the Social Register.) In addition, the Gallup poll that the Tribune based its headline on was over two weeks old at the time of the printing.[16]

Statistical corrections for a biased sample

If entire segments of the population are excluded from a sample, then there are no adjustments that can produce estimates that are representative of the entire population. But if some groups are underrepresented and the degree of underrepresentation can be quantified, then sample weights can correct the bias.

For example, a hypothetical population might include 10 million men and 10 million women. Suppose that a biased sample of 100 patients included 20 men and 80 women. A researcher could correct for this imbalance by attaching a weight of 2.5 for each male and 0.625 for each female. This would adjust any estimates to achieve the same expected value as a sample that included exactly 50 men and 50 women, unless men and women differed in their likelihood of taking part in the survey.

10.2 SELECTION BIAS: Selection bias is a statistical bias in which there is an error in choosing the individuals or groups to take part in a scientific study. 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.

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, 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. It is mostly classified as a subtype of selection bias, sometimes specifically termed *sample selection bias*, but some classify it as a separate type of bias.

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.
  - *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. It can be mitigated by lagging, that is, exclusion of exposures that occurred in a certain time period before diagnosis.
  - *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.
- Rejection of "bad" data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
- Rejection of "outliers" on statistical grounds that fail to take into account important information that could be derived from "wild" observations.

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), discounting trial subjects/tests that did not run to completion. It includes *dropout*, *non response* (lowerresponse rate), *with drawal* 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.

Observer selection

Data is 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.

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. Astronomical existential risks might similarly be underestimated due to selection bias, and an anthropic correction has to be introduced.

Avoidance

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 informal 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.

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.