Bias

Thomas R. Knapp

Introduction

The word "bias" means different things to different people. The Free Online Dictionary lists seven meanings for bias as a noun, one as an adjective, and two as a transitive verb. Porta's (2008) epidemiology dictionary has 15 entries for the term. Bethel Powers and I (2011) provided several meanings that bias has in quantitative nursing research and in qualitative nursing research.

Many of the meanings, e.g., the use of the term in sewing (a diagonal cut on a piece of cloth), have nothing to do with scientific research. In what follows I shall concentrate on several kinds of bias that can arise in research design, instrumentation, and data analysis.

Biased estimator

The easiest context of the term to deal with is the inferential statistical matter of a biased estimator. A sample statistic is said to be a biased estimator of a population parameter if the arithmetic mean of its sampling distribution is not equal to that parameter. Under the typical assumptions of inferential statistics, the sample variance with the sample size n in the denominator is a biased estimator of the population variance. (The statistic with n-1, rather than n, in the denominator is an unbiased estimator.) But that is not to say that division by n is necessarily bad. Division by n is the way we usually get a mean; and division by n also produces a maximum likelihood estimator of the population variance. (The statistic itself is an estimator; the number so calculated is an estimate.)

Measurement bias

Measurement bias occurs when a measuring instrument is found to produce "scores" that differ substantially from subgroup to subgroup. For example, a particular instrument for measuring blood pressure that yields systematically higher systolic pressure readings for women than it does for men is said to be biased, IF it is known that the two sexes have equal systolic pressures on the average.

But it is in the social sciences, not the physical or biological sciences, in which measurement bias is most often found. The principal context is the measurement of intelligence. Test A might be said to be biased against women; Test B might be said to be biased against Blacks; etc. There was an interesting "debate" on the internet (Whiting & Ford, n.d.) during which it was claimed that the factor structure underlying an intelligence test should even be the same for all subgroups of people to whom it is administered; otherwise the test is biased.

Publication bias

This is the biggie. The literature is replete with articles whose authors claim (sometimes with empirical evidence, sometimes without) that there is a bias on the part of reviewers and editors to prefer to include in their journals manuscripts for studies in which there is at least one statistically significant finding. A statistical significance test can result in a correct decision to reject a false null hypothesis, a Type I error (the rejection of a true null hypothesis), a correct decision to not reject a true null hypothesis, or a Type II error (the non-rejection of a false null hypothesis). If studies in which a null hypothesis is not rejected are seldom published, an excess of Type I errors over Type II errors is expected.

Biased sample

This is not the same thing as a biased estimator based upon a sample. It's actually much worse. A sample is said to be biased if it is not representative of the population to which inferences are desired to be made. An obvious example is a sample of college students in an introductory psychology course, if inferences are to be made to adults in general. (But you might be surprised to know how often such samples are used for that purpose.)

Some people argue (again sometimes with empirical evidence, sometimes without) that a small sample can never be representative of a large population. They would reject out of hand the typical sample used in a Gallup poll that is used to get a snapshot of the opinions of the American people at a given point in time. Such polls are often based upon samples of a few thousand people, out of approximately 200 million adults, are usually randomly drawn, but often have response rates less than 50 percent.

The term "representative" is itself controversial: Representative with respect to what? It is often not even used by statisticians, who talk only about probability samples (simple random samples or more complex random samples) vs. non-probability samples ("convenience" samples of various kinds). Random samples are representative enough for their work, since chance is a great equalizer and a protector against many biases.

Experimenter bias

In a true experiment (randomized clinical trial) that is not double-blinded, those carrying out the experiment might, consciously or unconsciously, give greater attention to those who receive the experimental treatment; or perhaps the other way 'round (greater attention to the controls). The "treatment effect" could thus be some combination of an actual effect and an attention effect.

Assignment bias

In a quasi-experiment where the assignment to treatment is not random, a researcher might favor certain groups or certain individuals over others. For example, subjects who are sicker might be chosen to receive the experimental treatment, in the hope that they would get better and their lives would be prolonged. Or subjects who are healthier might get preference, in the hope that the experimental treatment might work better for them.

Rater bias

This type of bias can come up in a variety of contexts, ranging from educational situations in which teachers grade their students on performance and/or attitude, to laboratory settings in which researchers rate the interactions between mothers and their children. It is often claimed that teachers favor girls. It is similarly claimed that some researchers find certain mother/child combinations to be "cuter" than others, and thus tend to give them higher ratings.

Contamination bias

If you think about it, all experiments should be run with each person in an isolation booth (just like in the old days of TV quiz shows), so that no person in the control group gets exposed to any feature of the experimental group, and vice versa. In the real world that is usually not feasible, but to the extent that control subjects are free to mingle with experimental subjects, contamination is possible and even likely. This is one of the reasons that the unit of analysis is often shifted from the individual person to the school, the hospital, or whatever.

Non-response bias

Non-response bias plagues almost all of survey research (some people sampled at random might refuse to be part of the sample, for example) but can also be a problem in other types of research (e.g., a randomized clinical trial in which they are told that by chance they might or might not get a particular experimental treatment).

Although it is not usually called non-response bias, there is the associated problem of missing data. Some people might agree to participate in a study but for a variety of reasons they don't provide all of the data that they are asked to provide. It might be background data (e.g., refusal to divulge their ages); it might be data for an important variable in the study itself (e.g., refusal to have blood drawn); or it might be something as simple as the omission of one or more items on an achievement test.

A very interesting dilemma can occur in a randomized clinical trial when some subjects who are randomly assigned to an experimental treatment participate to a limited extent, are switched to another treatment, or drop out of the study altogether. By virtue of the random assignment the treatment groups are comparable at the beginning of the study, but if there is any limited participation, treatment switching, or dropout, that might no longer be the case. When it comes to an analysis of the data, there are two schools of thought. One is the Intent(ion)-to-treat (ITT) approach, in which every subject's data are associated with the treatment to which he(she) was assigned, regardless of how little or how much of that treatment he(she) actually received. The other is the per-protocol (PP) approach, in which the data for all subjects are associated with the treatment they got, which is not necessarily the treatment to which they were assigned. Whichever approach is chosen there will be a missing-data problem. For ITT what are missing are the data that would be otherwise associated with the opposite treatment. For PP some subjects might have to be eliminated entirely from the analysis. See Gross and Cobb (2004) for a good discussion of the biases associated with both of those approaches.

Digit preference bias

People seem to like numbers that end in zero or five (we're not sure why that is). When asked how old they are, some people who are age 29 or age 32 might say "30", for example. (Demographers call this phenomenon "age heaping".) The result is that age is occasionally not measured as accurately as it could be if such bias didn't exist. (Asking for date of birth would avoid that problem, but might create another one. Some people don't know their dates of birth or don't think about them as often as they think about their ages in years.)

So what should we do about bias?

Reference has already been made to using n-1 rather than n to get an unbiased estimate of a population variance. Well-written manuscripts for well-designed studies should be accepted for publication whether or not statistically significant results are obtained. (I remember once reading about a recommendation that researchers submit only the design aspects of a study for initial publication consideration. If the manuscript is accepted the results are subsequently provided. Nice.) Better training of experimenters, raters, and the like should help in the prevention or minimization of such biases. But for most of the other biases the solutions are more difficult. It could be argued that "bias", in the sense of discrimination of some sort, is inherent in human nature. A woman discriminates against Peter if she chooses to marry Paul (if both are suitors). Some editors and some reviewers don't like certain authors' writing styles (some don't like mine, for example). You can't force people to participate in a study, so non-response will always be a problem. However, there is an ingenious approach called the randomized response technique that has been designed to improve non-response to sensitive questions. See, for example, Campbell & Joiner, 1973. Also nice.

Additional reading

For a brief description of each of 32 different kinds of bias that might arise in medical research, see the section entitled "Varieties of bias to guard against" in Indrayan's (2012) textbook that is available via the medicalbiostatistics.com website.

References

Campbell, C., & Joiner, B.L. (1973). How to get the answer without being sure you've asked the question. The American Statistician, 27 (5), 229-231.

Gross, D., & Fogg, L. (2004). A critical analysis of the intent-to-treat principle. The Journal of Primary Prevention, 25 (4), 475-489.

Indrayan, A. (2012). Basic methods of medical research (3rd. ed.). Delhi: AITBS Publishers.

Porta, M. (Ed.) (2008). A dictionary of epidemiology (5th ed.). New York: Oxford University Press.

Powers, B.A., & Knapp, T.R. (2011). Dictionary of nursing theory and research (4th ed.). New York: Springer.

Whiting, G., & Ford, D. (n.d.) Cultural bias in testing. Retrieved from the internet.