# Misuse of statistics

From Wikipedia, the free encyclopedia

A **misuse of statistics** occurs when a statistical argument asserts a falsehood. In some cases, the misuse may be accidental. In others, it is purposeful and for the gain of the perpetrator. When the statistical reason involved is false or misapplied, this constitutes a **statistical fallacy**.

The false statistics trap can be quite damaging to the quest for knowledge. For example, in medical science, correcting a falsehood may take decades and cost lives.

Misuses can be easy to fall into. Professional scientists, even mathematicians and professional statisticians, can be fooled by even some simple methods, even if they are careful to check everything. Scientists have been known to fool themselves with statistics due to lack of knowledge of probability theory and lack of standardization of their tests.

## Contents

## Types of misuse

### Discarding unfavorable data

In product quality control terms all a company has to do to promote a neutral (useless) product is to find or conduct, for example, 40 studies with a confidence level of 95%. If the product is really useless, this would on average produce one study showing the product was beneficial, one study showing it was harmful and thirty-eight inconclusive studies (38 is 95% of 40). This tactic becomes more effective the more studies there are available. Organizations that do not publish every study they carry out, such as tobacco companies denying a link between smoking and cancer, or miracle pill vendors, are likely to use this tactic.

Another common technique is to perform a study that tests a large number of dependent (response) variables at the same time. For example, a study testing the effect of a medical treatment might use as dependent variables the probability of survival, the average number of days spent in the hospital, the patient's self-reported level of pain, etc. This also increases the likelihood that at least one of the variables will by chance show a correlation with the independent (explanatory) variable.

## Loaded questions

The answers to surveys can often be manipulated by wording the question in such a way as to induce a prevalence towards a certain answer from the respondent. For example, in polling support for a war, the questions:

- Do you support the attempt by (the war-making country) to bring freedom and democracy to other places in the world?
- Do you support the unprovoked military action by (the war-making country)?

will likely result in data skewed in different directions, although they are both polling about the support for the war.

Another way to do this is to precede the question by information that supports the "desired" answer. For example, more people will likely answer "yes" to the question "Given the increasing burden of taxes on middle-class families, do you support cuts in income tax?" than to the question "Considering the rising federal budget deficit and the desperate need for more revenue, do you support cuts in income tax?"

## Overgeneralization

If you have a statistic saying 100% of apples are red in summer, and then publish "All apples are red", you will be overgeneralizing because you only looked at apples in summertime and are using that data to make inferences about apples in all seasons. Often, the overgeneralization is made not by the original researcher, but by others interpreting the data. Continuing with the previous example, if on a TV show you say "All apples are red in summer" many people will not remember you said "in summer" if asked weeks later.

With a subject of which the general public has no personal knowledge, you can fool a lot of people. For example you can say on TV "Most autistics are hopelessly incurable if raised without parents or normal education" and many people will only remember the first part of the claim, "Most autistics are hopelessly incurable". This problem is especially prevalent on TV, where talk show hosts interview one individual as representative of a whole class of people.

## Biased samples

## Misreporting or misunderstanding of estimated error

If a research team wants to know how 300 million people feel about a certain topic, it would be impractical to ask all of them. However, if the team picks a **random** sample of about 1000 people, they can be fairly certain that the results given by this group are representative of what the larger group would have said if they had all been asked.

This confidence can actually be quantified by the central limit theorem and other mathematical results. Confidence is expressed as a probability of the true result (for the larger group) being within a certain range of the estimate (the figure for the smaller group). This is the "plus or minus" figure often quoted for statistical surveys. The probability part of the confidence level is usually not mentioned; if so, it is assumed to be a standard number like 95%.

The two numbers are related. If a survey has an estimated error of ±5% at 95% confidence, it might have an estimated error of ±6.6% at 99% confidence. The larger the sample, the smaller the estimated error at a given confidence level.

Most people assume, because the confidence figure is omitted, that there is a 100% certainty that the true result is within the estimated error. This is not mathematically correct.

Many people may not realize that the randomness of the sample is very important. In practice, many opinion polls are conducted by phone, which distorts the sample in several ways, including exclusion of people who do not have phones, favoring the inclusion of people who have more than one phone, favoring the inclusion of people who are willing to participate in a phone survey over those who refuse, etc. Non-random sampling makes the estimated error unreliable.

On the other hand, many people consider that statistics are inherently unreliable because not everybody is called, or because they themselves are never polled. Many people think that it is impossible to get data on the opinion of dozens of millions of people by just polling a few thousands. This is also inaccurate. A poll with perfect unbiased sampling and truthful answers has a mathematically determined margin of error, which only depends on the number of people polled.

Another problem that crops up is the "re-sampling" problem. For example, a survey of 1000 people may contain 100 people from a certain ethnic or economic group. The people taking the survey would then "re-sample" their results focusing on that group. They then make claims that a percentage of that group believes X, or whatever the survey is about.

Unfortunately, re-sampling reduces the statistical reliability of the data. The larger the total sample is, then the more likely the sample represents the population. So, consider the case where a sampling is done and has a margin of error of only 3%. Those reporting the statistics from the sample can say the answers to the questions represent population results plus or minus 3%. However, that error rate does not hold to sub-sets of the samples. If you take a sub-group that uses only 10% of the samples (say 100 samples from a 1000-sample finding) the error grows significantly higher than the original 3%. To accurately find statistics for a sub-group, that sub-group would have to be sampled by itself (e.g. the sub group would need the same number of samples as the original group).

Resampling errors seem to occur all the time in news coverage of survey data.

The problems mentioned above apply to all statistical experiments, not just population surveys.

There are also many other measurement problems in population surveys.

*Further information: Opinion poll, Statistical survey*

## False causality

When a statistical test shows a correlation between A and B, there are usually four possibilities:

1. A causes B.
2. B causes A.
3. A and B are both caused by a third factor, C.
4. The observed correlation was due purely to chance.

The fourth possibility can be quantified by statistical tests that can calculate the probability that the correlation observed would be as large as it is just by chance if, in fact, there is no relationship between the variables. However, even if that possibility has a small probability, there are still the three others.

If the number of people buying ice cream at the beach is statistically related to the number of people who drown at the beach, then nobody would claim ice cream causes drowning because it's obvious that it isn't so. (In this case, both drowning and ice cream buying are clearly related by a third factor: the number of people at the beach).

This fallacy can be used, for example, to prove that exposure to a chemical causes cancer. Replace "number of people buying ice cream" with "number of people exposed to chemical X", and "number of people who drown" with "number of people who get cancer", and many people will believe you. In such a situation, there may be a statistical correlation even if there is no real effect. For example, if there is a perception that the chemical is "dangerous" (even if it really isn't) property values in the area will decrease, which will entice more low-income families to move to that area. If low-income families are more likely to get cancer than high-income families (this can happen for many reasons, such as a poorer diet or less access to medical care) then rates of cancer will go up, even though the chemical itself is not dangerous. It is believed that this is exactly what happened with some of the early studies showing a link between EMF (electromagnetic fields) from power lines and cancer.

In well-designed studies, the effect of false causality can be eliminated by assigning some people into a "treatment group" and some people into a "control group" at random, and giving the treatment group the treatment and not giving the control group the treatment. In the above example, a researcher might expose one group of people to chemical X and leave a second group unexposed. If the first group had higher cancer rates, the researcher knows that there is no third factor that affected whether a person was exposed because he controlled who was exposed or not, and he assigned people to the exposed and non-exposed groups at random. However, in many applications, actually doing an "experiment" in this way is either prohibitively expensive, infeasible, unethical, illegal, or downright impossible. (For example, it is highly unlikely that an IRB would accept an experiment that involved intentionally exposing people to a dangerous substance in order to test its toxicity.)

## Proof of the null hypothesis

In a statistical test the null hypothesis (H0) is considered valid until enough data proves it to be wrong. When this occurs H0 is rejected and the alternative hypothesis (HA) is considered to be proven as correct. By chance this can happen, although H0 is true, with a probability denoted alpha, the significance level. This can be compared by the judical process, where the accused is considered innocent (H0) until proven guilty (HA) beyond reasonable doubt (alpha).

But if data does not give us enough proof to reject H0, this does not automatically prove that H0 is correct. If, for example, a tobacco producer wishes to demonstrate that his/her products are safe (s)he can easily conduct a test with a small sample of smokers vs a small sample of non-smokers. Since it is unlikely that any of them will develope lung cancer (and even if they do, the difference between the groups has to be very big in order to reject H0). Therefore it is likely that - even when smoking is dangerous - that our test will not reject H0. If H0 is accepted it does not automatically follow that smoking is proven harmless. The test is having a to small power to be able to reject H0 and therefore the test is useless and the value of the "proof" of H0 is also null.

This can - using the judicial analogue above - be compared with the true guilty defendant who is released just because the proof is not enough for a verdict. This does not prove his/her innocence, but only that there is not proof enough for a verdict.

## Data dredging

Data dredging is an abuse of data mining. In data dredging, large compilations of data are examined in order to find a correlation, without any pre-defined choice of a hypothesis to be tested. Since the required confidence interval to establish a relationship between two parameters is usually chosen to be 95% (meaning that there is a 95% chance that the relationship observed is not due to random chance), there is a thus a

5% chance of finding a correlation between any two sets of completely random variables. Given that data dredging efforts typically examine large datasets with many variables, and hence even larger numbers of pairs of variables, spurious but apparently statistically significant results are almost certain to be found by any such study.

Note that data dredging is a valid way of *finding* a possible hypothesis but that hypothesis *must* then be tested with data not used in the original dredging. The misuse comes in when that hypothesis is stated as fact without further validation.

## Data manipulation

Data manipulation is the presentation of scientific data in a misleading way to support a hypothesis that is actually without merit. Informally called "fudging the data," this practice includes selective reporting (see also publication bias) and even simply making up false data.

Examples of selective reporting abound. The easiest and most common examples involve choosing a group of results that follow a pattern consistent with the preferred hypothesis while ignoring other results or "data runs" that contradict the hypothesis.

Psychic researchers have long disputed studies showing people with ESP ability. Critics accuse ESP proponents of only publishing experiments with positive results and shelving those that show negative results. A "positive result" is a test run (or data run) in which the subject guesses a hidden card, etc., at a much higher frequency than random chance.

The deception involved in both cases is that the hypothesis is **not** confirmed by the totality of the experiments - only by a tiny, selected group of "successful" tests.

Scientists, in general, question the validity of study results that cannot be reproduced by other investigators. However, some scientists refuse to publish their data and methods.

## Linguistically asserting unit measure when it is empirically violated

*Unit measure* is an axiom of probability theory which states that, when an event is certain to occur, its probability is 1. This axiom is consistent with the empirical world, if the relation from a set of events that are certain to occur to a set of physical objects is one-to-one, but not otherwise. In the latter case, *unit measure* is scientifically invalidated.

Christensen and Reichert (1976), Oldberg and Christensen (1995) and Oldberg (2005) report observations of systems in which the relation is not one-to-one. A result of the lack of one-to-one-ness is that the following elements of statistical terminology are not defined for the associated systems: a) "population", b) "sampling unit", c) "sample", d)"probability", e) any term that assumes probability theory. A misuse of statistics arises when any of these terms are used in reference to a system that lacks one-to-one-ness, for *unit measure* is linguistically asserted and empirically violated. Oldberg and Christensen (1995) and Oldberg (2005) report observations of this type of misuse.

## See also

- *How to Lie with Statistics*, a 1954 book by Darrel Huff

## References

- Christensen, R. and T. Reichert, 1976 "Unit Measure Violations in Pattern Recognition, Ambiguity and Irrelevancy," *Pattern Recognition*, vol. 4, pp. 239-245. Pergamon Press.
- Hooke, R., 1983, *How to tell the liars from the statisticians*; Marcel Dekker, Inc., New York, NY.
- Jaffe, A.J. and H.F. Spirer, 1987, *Misused Statistics*; Marcel Dekker, Inc., New York, NY.
- Campbell, S.K., 1974, *Flaws and Fallacies in Statistical Thinking*; Prentice Hall, Inc., Englewood Cliffs, NJ.
- Oldberg, T., "An Ethical Problem in the Statistics of Defect Detection Test Reliability," 2005, Speech to the Golden Gate Chapter of the American Society for Nondestructive Testing. Published on the Web by ndt.net at http://www.ndt.net/article/v10n05/oldberg/oldberg.htm.
- Oldberg, T. and R. Christensen, 1995, "Erratic Measure" in *NDE for the Energy Industry 1995*; The American Society of Mechanical Engineers, New York, NY. Republished on the Web by ndt.net at http://www.ndt.net/article/v04n05/oldberg/oldberg.htm.

Retrieved from "http://en.wikipedia.org/wiki/Misuse_of_statistics"
Categories: Misuse of statistics | Laymen and statistics
Hidden categories: All articles with unsourced statements | Articles with unsourced statements since March 2007 | Articles with unsourced statements since December 2008 | Articles with unsourced statements since January 2009

- This page was last modified on 1 April 2009, at 18:36 (UTC).