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

<table>
<thead>
<tr>
<th>Discarding unfavorable data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loaded questions</td>
</tr>
<tr>
<td>Overgeneralization</td>
</tr>
<tr>
<td>Biased samples</td>
</tr>
<tr>
<td>Misreporting or misunderstanding of estimated error</td>
</tr>
<tr>
<td>Proof of the null hypothesis</td>
</tr>
<tr>
<td>Data dredging</td>
</tr>
<tr>
<td>Data manipulation</td>
</tr>
</tbody>
</table>

### 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 Organizations that do not publish every study they carry out, such as tobacco companies denying a link between smoking and cancer, or miracle pill vendors, are likely to use this tactic.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 USA to bring freedom and democracy to other places in the world?
• Do you support the unprovoked military action by the USA?

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.

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**

In statistics sampling bias is causing some members of the population to be 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 participants are not equally balanced or objectively represented.[2] If the bias makes estimation of population parameters impossible, the sample is a non-probability sample. If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling. It is a form of sampling error, that is, an error caused by observing a sample instead of the whole population. [3] However, sampling error also includes non-systematic errors that can be decreased by increasing sample size.

1. Selection from a specific 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 over-representation 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).

2. 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.

3. 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 centre, but advertise for smokers during the advanced aerobics class, and for non-smokers during the weight loss sessions.
4. 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 nonresponse, a selection bias in that it rather affects the internal validity of the study.

5. 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).

6. 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.

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.

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 also has an estimated error of ±6.6% at 99% confidence. ±(x)% at 95% confidence is always ±(1.32x)% at 99% confidence. The smaller the estimated error, the larger the required sample, 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.

However, often only one margin of error is reported for a survey. When results are reported for population subgroups, a larger margin of error will apply, but this may not be made clear. For example, a survey of 1000 people may contain 100 people from a certain ethnic or economic group.

\[\text{Confidence intervals are equivalent to p-values of probability; a CI of 95% corresponds to a p-value of } p < 0.05.\]
The results focusing on that group will be much less reliable than results for the full population. If the margin of error for the full sample was 4%, say, then the margin of error for such a subgroup could be around 13%.

There are also many other measurement problems in population surveys. The problems mentioned above apply to all statistical experiments, not just population surveys.

**False causality**

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

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

The fifth 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 four 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[who?] 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 judicial 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 versus a small sample of non-smokers. It is unlikely that any of them will develop 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 - 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 too small of 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. In other words, "absence of evidence" does not imply "evidence of absence".

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 data sets 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

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[1].
See also


References


Ercan I, Yazici B, Ocakoglu G, Sigirli D, Kan I *Review of Reliability and Factors Affecting the Reliability*, InterStat, 2007 April, 8