MedicalBiostatistics.com

STATISTICAL FALLACIES

Large number of statistical fallacies occurring in medical literature are discussed in Indrayan's book Medical Biostatistics, Second Edition (2008), published by Chapman & Hall/CRC Press. The following are some of them.

Data analysis can always be geared to serve an ulterior motive but our concern in this note is with interpretation and particularly the reporting. Statistical fallacies in reporting are quite common. Although some of these occur inadvertently but some could be deliberate. Avoid both.

MEAN OR PROPORTIONS

Depending upon the preference the results can be provided in terms of mean or distribution of blood pressure (BP) levels or in terms of prevalence of hypertension, in terms of haemoglobin (Hb) level or in terms of anaemia, in terms of plasma glucose level or in terms of diabetes, etc. No such name is available for a measurement such as total lung capacity but this can be categorized as low (<4.0 l), medium (4.0-5.9 l), and high (≥6.0 l). All quantitative measurements can be converted to qualities. The summary measure for quantities generally is mean and for qualities is proportion. Which one should you use in your report?

Means can lead to a conclusion different from the one reached by proportions. This provides a leeway to the investigator to try out both, and report the one that suits a particular hypothesis. In addition, for anaemia for example various categories can be tried: <12 g/dl, <11 g/dl or <10 g/dl; and the one that looks 'favourable' can be adopted. Thus the report may not truly reflect the findings. The general statistical advice is to stick to the original metric measurements and not convert them into categories. This is more exact and removes the subjectivity in devising categories. When categories are essential, state them before hand preferably at the time of writing the protocol and justify those categories. Do not change them later on unless there are strong reasons to do so.

MISUSE OF PERCENTAGES

If six patients out of eight respond to a new treatment, is it proper to say that the response rate is 75 percent? What about stating one out of two as 50 percent? Isn't preposterous to call zero out of one as nil and one out of one as complete? These are extreme examples but illustrate that percentages based on small *n* can be misleading. Generally no percentage should be stated when *n is* less than 30.

While assessing gestational age for 300 births, if 60 women do not know the date of last menstrual period, the percentage of births with gestation, say, 32-34 weeks should be calculated out of 240 and not 300. This is quite obvious yet many present percentages with a wrong denominator.

Linkage of hypertension with A, B, O blood groups has not been investigated much. If the distribution in a sample of hypertensives is A, 15%; B, 40%; AB, 25%; and O, 20%; it would be naive to conclude that hypertension is predominant in B group. These percentages should be compared with the blood group distribution in the target population for any such conclusion.

Summary measures such as mean and proportion, when based on the aggregated data, can be deceptive. They could mask or aggregate the variation present in subgroups. We illustrate this for percentages in Example 1.

Example 1: Percentages based on aggregated data can mask subgroup variations
Consider case-fatality in cancer patients in a general hospital and a cancer hospital:

	General Hospital			Cancer Hospital		
Stage of cancer	*n*	Deaths	Case-fatality(%)	*n*	Deaths	Case-fatality(%)
Stages I & II	150	30	20.0	50	10	20.0
Stages III & IV	50	30	60.0	250	80	32.0
Total	200	60	30.0	300	90	30.0

Both the hospitals have same case-fatality in aggregate but actually cancer hospital is receiving patients predominantly in advanced stages. In them, its performance is markedly better: 32 percent case-fatality in cancer hospital against 60 percent in a general hospital. If only the aggregate percentage is reported, this distinction is lost. It is for such discrepancies that standardization is advocated.

Another fallacy occurs in stating too many decimal places. Percentage of 3 out of 35 can be stated to as many decimal places as one wishes to. The tendency in medical literature is to use excessive decimals. The rule is as follows. Use one decimal place if $n < 100$, two if $100 \leq n < 1000$, three if $1000 \leq n < 10{,}000$, etc. Mean and SD should be stated one decimal more than the accuracy of original values. This retains the accuracy without sounding too precarious. If uric acid is measured in mg/dl as 3.3, 4.5, etc., the mean and the SD should be stated with two decimals. For mean and SD of total lipids one decimal is enough since the measurement is in integers. Correlation coefficient is conventionally stated to two decimal places.

There are exceptions to these rules. Sometimes it is considered neat to state all results with same number of decimals, even when they are based on different *n*, or even when different measurements have different accuracies. Another concept is of **significant digits**. Zeroes immediately after decimal are not counted. Thus 0.0032 has two significant digits and 0.32 also has two significant digits. It would be unfair to state 0.0032 as 0.00 when the system of two decimal places is religiously followed for uniformity.

INADEQUATE INTERPRETATION

If you happen to see 3 children wearing specs watching TV for long hours, can you conclude that wearing specs in children is associated with watching TV for long hours? The key is how many others watching TV for long hours do not wear specs. Counter-evidence is as important

as evidence. You will find many such instances in the popular media, and many people tend to believe them.

Many medical researchers are much too keen to look at the difference or gain in medical parameters after the treatment compared to values before treatment. In their keenness, they forget that a gain of 3 g/dl in Hb level over pretreatment value 8 g/dl has a different meaning than the same gain over the pretreatment value of 11 g/dl. It is relatively easy to affect a rise over lower Hb values than over higher values.

Another common misinterpretation is considering mere association or correlation as an evidence of cause-effect. Incidence of cardiovascular diseases in a developing country is negatively correlated with birth rate but it has no causal implication. Such nonsense and spurious correlations are discussed further in Medical Biostatistics.

MISUSE OF *P*-VALUES

Scientists debate about the validity of the conventional cut-off 0.05 for *P*-values. The opinion is growing to use confidence intervals instead of *P*-values. But in some situations there is no escape and a level has to be fixed to assess statistical significance. Certainly no magic happens at 0.05. A safe rule is to interpret *P* between 0.04 and 0.06 with caution and conclude that further work is needed. This is the same kind of precaution that is always taken for patients with borderline values. In any case, do not take p-values too seriously. They must be complemented by common sense. P-value is only probability of Type I error. Other errors can not be ignored.

Many examples can be cited from medical literature when more than one statistical test is done on the same data, each at level 0.05, without realising that this inflates the error rate. There are statistical methods such as Bonferroni and Tukey that should be used in such cases to control the chance of error to the specified level.

Statistical inference methods are applicable only to random samples. The basic purpose is that the sample is representative of a defined target population and generalisability is intact. Many articles use tests of significance on nonrandom samples with ill-defined target population. Before extrapolating results to a larger group, ensure that the sample in the study is representative of that larger group. Implication of most medical findings is in managing future cases. Thus examine how the conclusion will be applicable to the future cases and of what type. Write the discussion accordingly.

MISUSE OF STATISTICAL TOOLS

Misuse of percentages and of *P*-values have already been discussed. Figure 1 (a and b) illustrate how a graph can be manipulated to show a small gradient as steep and vice-versa.

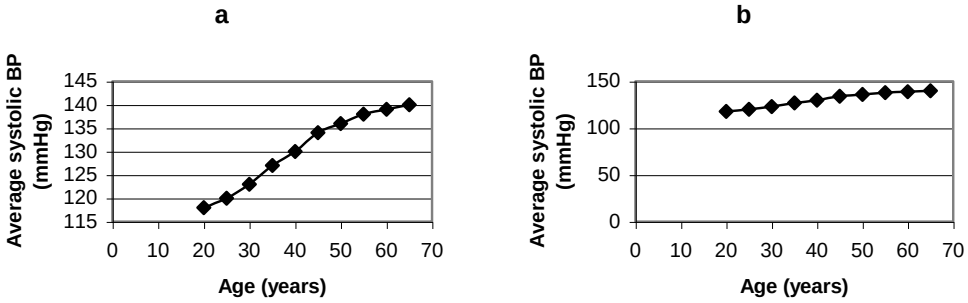

Figure 1: Same gradient shown as steep in Figure a and mild in Figure b

Most graphs do not fully represent the sample size. A mean or percentage based on n=3 is shown the same way as the one based on n=50. Even if n is stated in the graph, perhaps the perception and cognition received from the graph still remains the same.

Looking for linearity when the relationship is clearly nonlinear can damage the results. Ignoring assumptions such as Gaussian form of distribution, independence of observations, and uniform variance, can produce results of doubtful quality.

Statistical packages also are misused. Data are over analysed to investigate aspects that were not part of the protocol. Such analysis is not prohibited but fallacy arise when they are packaged as original investigation, suppressing that these are incidental findings. The results of such analysis should be stated as tentative for the purpose of generating hypothesis rather than for testing it.

All round availability of computers and software has made it easier to reanalyze the data after dropping some inconvenient observations. This misuse is called **data dredging**. Editors and reviewers would rarely be able to detect it because a finished report may not contain any such evidence.

Computer technology has surged ahead at a fast rate but the understanding of statistical methods among medical professionals has not kept that pace. As a result, some researchers use **black-box approach** to analyse the data. They just grind it through a statistical package without worrying about applicability of those methods to their data set. They may use regression where analysis of covariance is needed, use nonparametric methods where parametric methods are required, use nonlinear approach where linear is required, develop a model with eight variables where only four were enough, etc. Our advice is to involve a biostatistician right from the beginning and use his expertise where needed. He helps in improving both the design and analysis but can not be expected to resurrect a badly designed study (Hall 1996). Beware though that the biostatistician may not be able to fully comprehend the medical implications of the results and thus may not be able to provide ideal advice. A frequent interaction between the biostatistician and the researcher helps to understand each other better. Also realize that good biostatisticians are rare. Ensure that he has adequate knowledge of statistical tests and other tools and knows which software is best for a particular application. If time permits, cross-validate your analysis by two different softwares.