WLODZIMIERZ KLONOWSKI

How to lie with statistics or How to extract data from information

A well-wrapped statistics is better than Hitler's „big lie";
it misleads, yet it cannot be pinned on you...
There is terror in numbers.
-Darrell Huff *How to Lie with Statistics*

Logic is a systematic method for getting the wrong conclusion...
with confidence;
Statistics is a systematic method for getting the wrong conclusion...
with 95% confidence.
-Anonymous

Abstract In this paper we try to demonstrate how to distinguish good statistics from bad statistics and to emphasize the points one should watch for while reading statistics. We also give examples of manipulating the data, especially medical data, with statistics.

1. Introduction

The first part of the title is borrowed from Darrell Huff's book *How to Lie with Statistics* [1]. The second comes from a joke a friend of mine told me recently:

- *What's the difference between information and data?*
- *?!*
- *If one knows that an organization or an institution has money to spend on research, that's information, isn't it? If one has that money already in the pocket or in his/her bank account, that's data...*
- *You know* - I told - *if this is the difference, then the main purpose of the majority of research ourdays, and of statistics in particular, is to extract data from information...*

The purpose of **bad statistics** is to extract data from information in the above mentioned sense... Unfortunately, reading scientific papers shows that much too often it is really the case.

One can lie with statistics, because numbers can be manipulated to support any argument. If one wants to demonstrate that the population is not starving, one adjusts the threshold where starvation sets in. If the numbers run up on one group don't look so good, pick another group. If the average is too low or high, go for the median and arrange data to discard the high or low end. Statistics, done honestly, can make a statement like no other, but done dishonestly are deeply deceptive because the readership believes the numbers have been run up honestly. In an era of increasing distress, governments want the statistics on the homeless, the unemployed, and the uninsured to appear healthy. Likewise, corporations wishing to lie to consumers or to their stockholders discard the unpleasant from the computation and hope no one looks too closely.

Medical statistics is not any exception. John C. Bailar III, Chair, Department of Health Studies University of Chicago wrote on an Internet discussion list Scifraud, Discussion of Fraud in Science: „I was the Statistical Consultant for *The New England Journal of Medicine* for eleven years, and reviewed about 4000 submitted papers during that time. These were very nearly all passed as potentially publishable by the peer reviewers expert in the subject matter. About half had serious statistical problems, and a majority of those were not remediable. I do not like to think about the proportion of problems in the incoming stream of submissions, before the regular peer reviewers did some sorting out of them. Nor do I like to think about what gets through journals with less rigorous criteria. (I read enough of them to get an upset stomach)".

In this paper we try to demonstrate shortly how to distinguish good statistics from bad statistics. We also give some examples to illustrate given points, taken from my own experience, the Internet, Darrell Huff's book, and other sources.

2. How to read statistics

While reading statistics one should ask the questions like those listed in Table 1 and watch out for important things listed in Table 2, especially to look for conscious biases as listed in Table 3.

By applying different statistical approach to the same set of data one may demonstrate three different, mutually exclusive trends in changing of the mean value, i.e. that the mean increases, decreases, or does not change at all. Such "techniques" are often applied purposely by governments to lie to people.

Table 1.

QUESTIONS TO ASK WHILE READING STATISTICS

- who says so?
- how does he/she know?
- what's missing?
- does somebody change the subject?
- does it make sense?

Table 2

THINGS TO WATCH OUT FOR WHILE READING STATISTICS

- that the question being asked is relevant
- that the data come from reliable sources
- that all the data are reported, not just the best (or the worst)
- that the data are presented in context
- that the data have been interpreted correctly

Table 3.

CONSCIOUS BIASES TO LOOK FOR WHILE READING STATISTICS

- selection of favourable data and suppression of unfavourable
- the sample not large enough to permit any reliable conclusion
- using of unqualified word „average”
- not stating deviations from the mean value
- correlations given without a measure of reliability (stadard errors)
- deceptive use of percentage, percetage points, and percentiles
- using unjustified extrapolation of trends „everything else being equal”
- large-scale falsifying at source by those wanting to get benefits
  (saying and doing may not be the same thing at all)

3. How to tell bad statistics from good statistics

Good statistical analyses have several components, listed in Table 4. It follows that bad statistical analyses violate one or more of these precepts (Table 5). To challenge somebody’s statistics (to ”talk back to a statistics”, cf. [1]) ask questions like these in Table 6. Because when statistical thinking is merely a license for loose talk, then it is bad statistics. There is a lot of it around.

Statistical studies depend upon how many are polled, how the info is gathered, how much time the study covers, and a couple zillion other things [1].

To be meaningful, statistical thinking should include what’s listed in Table 7. This is clearly very far from the views and practices of a lot of armchair "statisticians", who see statictics as the calculation of p-values and confidence limits. Next time you see a statistical analysis that you think is flawed, see whether it was prepared by a fully trained professional statistician or by someone (possibly with other expertise) who has learned some of the language and how to run some computer package. The latter are dangerous.

The other side of the story, however, is that much of nature and our observation of it, is inherently statistical. Ignoring that merely gives an illusion of certainty. *Statistics is a wonder of Nature* - says Polish poetess A.Osiecka.

Table 4.

GOOD STATISTICAL ANALYSIS

A. Asks meaningful questions
B. Gets good data
C. Truthfully represents the data
D. Truthfully interprets the data

Table 5.

BAD STATISTICAL ANALYSIS

A. Asks meaningless questions
B. Gets Bad Data

1. Selection bias
2. Nonresponse bias
3. Biased questions
4. Excluded data

C. Misrepresents the Data

1. Truncates a histogram to maximize a difference
2. Expands a histogram to maximize a difference
3. Presents one dimensional data multidimensionally
4. Ignores some factors (e.g. population growth)
5. Takes the data out of context
6. Compares dissimilar groups
7. Uses different measures for different groups

D. Misinterprets the data

Table 6.

QUESTIONS TO ASK TO CHALLENGE A STATISTICIAN

What are you looking for?
What kind of questions you want to ask?
How your data aquisition and data analysis is planned?
What is the relevant information you try to collect?
What variables are relevant?
What is a statistical unit? (one kidney? one patient? one procedure?)
What was included in the data?
How did you arrive at these figures?
What method of randomization was used?
What method of normalization was used?
Do you know what the computer programme you use really does?

Table 7.

WHAT STATISTICAL THINKING SHOULD INCLUDE

1. Choosing and phrasing the question in a way that can be answered by data that one can collect;
2. Specifying methods of analysis as well as principal hypotheses before the first look at any data (to avoid using one test after another until one gets the desired answer);
3. Disclosing to readers all of the soft spots in the analysis and the data, careful attention to the protocol to avoid or reduce the likelihood of bias;
4. Continuing attention to the quality of the data and their improvement;
5. Unbiased procedures to record, process, and reduce the data to usable forms such as tables, graphs, and summary measures;
6. Careful attention to the limits of generalizing results.

4. The ways to use statistics to deceive... the crooks already know these tricks. Honest men must learn them in self-defense [1]

4.1. The sample with the built-in bias

A river cannot rise above its source - the result of a sampling study is no better than the sample it is based on. For good statistical analysis one needs a sample which is random. The test of the random sample is this: Does every member of the whole group (the ”universe”) have an equal chance to be in the sample? Another possibility is to use stratified random sampling - first one divides the universe into several group in proportion to their known prevalence.

Before random sampling has been invented by persons like famous dr George H. Gallup, to predict a result of an event like presidential election in USA, sometimes millions of persons were polled by magazines like *Literary Digest*. In 1936 Gallup correctly predicted Franklin Delano Roosevelt’s victory over Alf Landon by polling only a randomly chosen sample of only 5,000 persons while the *Digest* which polled over 20 millions (!) of its readers uncorrectly predicted Landon’s victory. The reason of this was simple - readers of the *Digest* were not representative for the whole US population.

Also any questionnaire is only a sample (another level) of the possible question, and the answer to the question is no more than a sample (third level) of the respondent attitudes and experiences on each question.

The answer often depends also on the interviewer, not only on the respondent - a tendency that must always be allowed for in reading poll results it’s a desire to give an answer which pleases the interviewer. The different groups of interviewers choose different kinds of people to talk to. The polls are usually biased toward the person with more money, more education, better appearance etc. than the average in the population he/she is chosen to represent.

4.2. Manipulating the data

In Table 8 we give examples of manipulating the data with statistics and with words.

Table 8.

MANIPULATING THE DATA:
SAME DATA - COMPLETELY DIFFERENT ANSWERS

THE DATA: FRUITS' PRICES

	1995	1996
APPLES	$ 1.00	$ 2.00
ORANGES	$ 2.00	$ 1.00

Q. DOES THE MEAN PRICE OF FRUITS:

i. REMAINED UNCHANGED?
ii. INCREASED?
iii. DROPPED?

	1995	1996
i. ARITHMETICAL AVERAGE OF THE PRICES		
	$ (1.00+2.00)/2	$ (2.00+1.00)/2
Mean price	$ 1.50	$ 1.50

A. unchanged

ii. ARITHMETICAL AVERAGE OF PERCENTAGES, FIRST PERIOD = 100%

APPLES	100%	200%
ORANGES	100%	50%
Mean price	100%	125%

A. 25% increase

iii. ARITHMETICAL AVERAGE OF PERCENTAGES, SECOND PERIOD = 100%

APPLES	50%	100%
ORANGES	200%	100%
Mean price	125% -->100%	100%-->80%

A. 20% decrease

iv. GEOMETRICAL AVERAGE OF PERCENTAGES USING EITHER PERIOD

$$\sqrt{(50\% \cdot 200\%)} = \sqrt{(200\% \cdot 50\%)} = 100\%$$

A. unchanged

4.3. „Normality”, averages, and statistical errors

The word *average* has a very loose meaning. In English, the *average* may denote the *mode* (i.e. the value which occurs most frequently), the *median* (i.e. the value which is in the middle of the distribution, with 50% below and 50% over it), and the *mean (arithmetical average).* For normal (Gaussian) distribution mean, median, and mode fall at the same point. For a skewed distribution the mean could be quite a distance from the mode. Also the notion of *standard deviation* is closely related to the normal distribution. Those are the reasons why normal distribution is so often used without necessary basis - it frees from careful consideration of the real meaning of the words one uses.

Often the kind of average is carefully unspecified and the errors are not given. Unqualified ”average” is virtually meaningless. When you see an ”average”, ask the questions like in Table 9.

Especially in clinical medicine mean values are of little or no value [2] because standard deviation is often of the same order as the corresponding mean. For example, normal physiological values of intracranial pressure (IP) differ as much as from 1.47 kPa to 1.96 kPa in adults, and from 0.44 kPa to 0.88 kPa in children. So, what for one person may be considered as a pathologically high IP, for another person is completely normal, and oppositely. If one would like to cause IP to decrease for all the patients with high IP and to increase for all those with low IP, in order to attain some mean value taken for a *”standard IP”*, one would cause death of many patients

One must always keep in mind the deviation from the mean value (assumed to be the ”normal” value for the investigated population) EVEN or ESPECIALLY if they are NOT stated. The same concerns statistical standard errors.

Table 9.

„AVERAGE” QUESTIONS

1. Average of what?
2. Who’s/what’s included?
3. What kind of average this is?
4. How accurate the figure is?
5. Who says so and how he/she knows?

In statistical mechanics one shows that:

THE AVERAGE OVER STATISTICAL ENSEMBLE =
= THE AVERAGE OVER TIME FOR A MEMBER OF THE ENSEMBLE

In the medical statistics the question arises: *Does the risk of 5 patients for 1 year equals the risk of 1 patients for 5 years?* My answer is „NO” - we are not molecules in a volume of an ideal gas, we are not members of a normally distributed ”specimen”.

4.4. Fallacious correlations

The cause-and-effect nature of the correlation is often only a matter of speculation - When „B" always follows „A", it does not mean that „A" causes „B".

A co-variation may be real, but it may not be possible to be sure which of the variables is the cause and which the effect. Or neither of the variables has any effect at all on the other, despite there is a correlation. The primitive people of the New Hebrides found by observation over the centuries that people in good health usually had lice and sick people often did not; their conclusion: *„Lice make a man healthy"*; scientific explanation - when somebody get a fever (possibly brought by the lice) he/she became too hot for comfortable habitation and the lice left.

Correlation does not prove cause and effect, but only parallelism. Moreover, a single event or a single person is not a statistic, and statistical conclusions are often of little relevance to a single person.

For example, one reads: *„Each year of post-secondary education puts some more income into pocket"*. But if persons with post-secondary education earn more money it does not mean that it is so because they are college graduates; and it does not mean that if your child attend college he/she will earn more money than otherwise. A positive correlation may quickly become a negative one when some threshold is oversurpassed. Persons with post-secondary education may earn more but those **"overqualified",** with Ph.D. degrees, often become scientists and so do not become members of the highest income groups...

5. Meanders of statistics

Application of statistics has many *meanders*. We list below the most important ones (cf. [1]):

Changing the subject between the data and conclusion:
Sometimes the conclusion is based on another model than the model presented. E.g. the presented model is based on the theory of branching processes but the used computer program and so the data obtained are based on percolation theory, which may often be applied to the same problems but fundamentally differs from the theory of branching processes.

The variation of something *with* something else is presented as *because of*:
The meaning of a correlations is very often misunderstood.

The data might be true but the conclusion doesn't follow:
One reads: „*More people die in own bed than in any other place*". What to do, don't go to bed?! Better go to somebody's else bed... In a report on a possible influence of the electromagnetic fields around a radio-broadcasting antenna on the health of those living in the close neighbourhood, the authors give many data which demonstrate that a negative influence may be a real danger, and then in the *Summing Up* they formulate the conclusion for politicians who will decide about antenna's future: *Electromagnetic fields emitted by the antenna did not set any threat for the health of those living in the close neighbourhood* [3].

Irrelevant numbers:
One reads: *„This mark is 19.6% better than any other brand as proved by an independent laboratory test"*. Such statements like this are suspiciously precise!

Inconsistent reporting at the source:
Contradictory figures about tuberculosis, influenza, polio, malaria, abortions, STDs (by the definition I like to use, *life is a sexually transmitted terminal disease*) - the ”facts” mean often nothing else but that in some periods or in some places there were far more cases reported, due to the increased consciousness, better diagnosis etc. E.g. cancer is often listed now where ”causes unknown” was formerly used and the total numbers may be greater if used instead of appropriate rates, just because there are more people now than there used to be.

Manipulating figures' scale:
Suppose you wish to win an argument, shock a reader, move him into action, sell him something. Chop off the bottom. The figures are the same and so is the curve ....nothing has been falsified - except the impression that it gives...Why stop with truncating? Simply change the proportion between the ordinate and the abscissa. Or apply logarithmic, or even better double-logarithmic scale.

Many ways of expressing the same:
For example, exactly the same fact may be called:

- a one-million profit;
- a decrease in profits of sixty per cent from the previous year;
- an increase in profits of twenty per cents compared with the last decade average;
- a two per cent return on sales;
- a ninety per cents return on investment.

Different ways of stating the same:
When the sharp-tongued Benjamin Disraeli, so the story goes, was ordered in the last century to withdraw his declaration that half of the cabinet were asses. *”Mr. Speaker, I withdraw,”* was Disraeli's response, *”Half the cabinet are not asses.”*

6. The abuse of logic

Not only statistcs is abused in medical papers, but quite often also pure logic. Here are two examples.

The no-threshold - regression contradiction -
cholesterol and coronary heart disease (CHD):

It has been universally accepted among epidemiologists that the relationship between blood cholesterol and CHD is continous, graded, and without a threshold [4]. In fact, it is also recognized that atherosclerosis begins in newborns whose cholesterol levels are at their lowest. CHD progression seems to occur at every cholesterol level and the best that can be obtained from cholesterol lowering is *a reduction in the rate of CHD progression*.

However, some authors of angiographic studies mantain that not only does atherosclerosis progression cease with cholesterol-lowering, it also reverses - they report *regression* of atherosclerotic plaque (cf. e.g. [5]). This regression concept is accepted and supported by National Heart, Lung, and Blood Institute (NHLBI), a part of NIH.

The no-treshold, regression contradiction has been overlooked often within the same report even by the most prominent epidemiologists [6]. If one accepts the concept that there is a cholesterol threshold *and* the regression concept, then one must necessarily accept the concominant concept that cholesterol is degenerative above and regenerative below the threshold. Moreover, regression has been purported to occur at almost any level of cholesterol, so one must also accept the premise that the threshold varies accross almost the entire range of cholesterol level as well [7].

The well established fact that there are two kinds of cholesterol - High-Density-Lipoprotein (HDL) and Low-Density-Lipoprotein (LDL) Cholesterol, of which the former is considered to be „good" and the latter to be „bad", does not take off the no-threshold - regression contradiction, as some doctors state. There are actually many reasons to doubt the validity of the angiographic studies, not the least of which is the near impossibility of obtaining a picture of the exact same spot in an artery at the exact same angle at different points in time - it seems that the angiographic studies have been designed and conducted, and their results have been analysed in a systematically biased manner [7].

Left-handed people have shorter life spans than righties:

Coren [8] claims that adjustments to a right-handed world ultimately kill many lefies; he also claims that lefties are five time as likely to die in accidents and injuries, as the right-handed world will not accomodate them. An article in *Smithsonian* [9] claims that lefties suffer from a higher incidence of specific health problems, including learning disabilities, depression, migraine, allergies and autoimmune disorders such as rheumatoid arthritus and ulcerative colitus.

In 1988 how Coren and Diane Halpern, a psychologist at California State University in San Bernadino, analyzed the life spans of 2,271 baseball players, from "The Baseball Encyclopedia". They found that, on average, right-handers live eight months longer than lefties, a small but notable difference. However, it was their 1991 study, which was reported in a five-paragraph letter to the *New England Journal of Medicine*, that made Coren and Halpern anathema among left-handers. They polled relatives of 2,000 people in Southern California who had died recently, asking if the deceased was left-handed. The researchers tabulated a mean age of death for right-handers at 75; for left-handers at 66 -- a difference of nine years! ... "Don't wait for Lefty," a *New York Times* article announced, "He's dead." Other researchers dispute this. Marcel Salive, an epidemiologist at the National Institute on Aging came to the conclusion that death rates were almost the same for lefties and righties. The lack of aged lefties, Salive speculates, may well be due to switching in the early part of the century.

Why does right-handnes prevail? Neurobiologist William H.Calvin suggested that early human mothers carried their babies in their left arms, close to the soothing rhythms of the heart, leaving their right hands free for throwing rocks at rabbits... Conclusions? Maybe there is truth here, but I don't trust the statistics, or the social aspects that influence the results leading to such conclusions. I'm stopping to write this paper to go throw some rocks at a few rabbits...

7. Statistics and chaos

Statistics has been considered to be applicable to the processes and phenomena governed by stochasticity, i.e. a mechanism of chance [10]. But a signal which seems to be stochastic may be as well produced by a completely deterministic mechanism [11]. The phenomena like this are now called *deterministic chaos*.

Brownian motion is a good example of what was considered by physicists to be a stochastic process. In 1931 Kappler made an experiment to verify Smoluchowski's theory of Browniam motion. Mark Kac demonstrated [11] that the graph of a simple deterministic function (e.g. the sum of cosines with different „frequencies") can not be distinguished from the stochastic signal (like that obtained by Kappler) by any analytical technique used in statistics (after [12]).

The fact that the body sometimes appears to have random processes is because the numerous variables are involved, and we have not even begun to understand all of the interactions that most certainly have logical meaning - illogic is a creation of the mind [7].

One may say that chance concerns order in disorder, while chaos - disorder in order. In this meaning both chance and chaos are two opposite but complementary "wings" of statistics, and both are applicable in clinical practice.

8. Concluding remarks

From what was told, these matters are the core of training of any medical statistician, and should form the basis of his or her contributions to the progress of science:

1. Ask questions in ways that allow for some real advance in understanding, not in ways to support a predetermined conclusion;
2. Draw up the protocols honestly, and do not depart from them unless you tell your readers;
3. Be very attentive to the quality of data, not just to get the best you can but to measure the irreducible residual uncertainty from either bias or random variation; for complex multidisciplinary projects, one cannot possibly ever comprehend the quality of all the data. We must operate on the hope that others have collected good data;
4. Use the right "statistical" procedures in the right way;
5. Present results in ways and with explanations that will help readers to understand the real strengths and limitations of what you have done.

Medical statistics is one of the mostly misunderstood fields of medical sciences. It is misunderstood by many doctors. Like that surgeon, saying to a patient just before the operation: *Oh, you are a happy man! You know, 90% of patients who are undergone this operation dies in the operating room. And imagine, 9 patients I have most recently operated did die. Oh, you are really a very happy man!*

Carl Friedrich Gauss (who gave us the famous bell-like Gauss curve) once said: *It has been a long time since I have got my results, but I still do not know how to obtain these results* (cf. [12]). And, most probably, he applied statistics...

References

[1] Huff D.: How to Lie with Statistics, W.W.Norton & Co. Inc, New York, 1954.

[2] Tanur J.M. et al., (eds.): Statistics: A Guide to Biological and Health Sciences, Holden Day, San Francisco, 1977.

[3] Grzesik J, Jonderko G., Langauer-Lewowicka H.: Evaluation of the influence of electromagnetic radiation emitted by the Konstantynow broadcast center on the sanitary condition of the population living in the II protective zone on the basis of the results of epidemiological examinations of that population, Instytut Medycyny Pracy i Zdrowia Srodowiskowego, Sosnowiec, 1993, (in Polish).

[4] Stamler J., Wentworth D., Neaton J.D.: *JAMA* 1986, 256: 2823-2828.

[5] Arntzenius A.A. et al.: *NEJM* 1985, 312: 805-811.

[6] Glueck C.J.: *J.Lab.Clin.Med.* 1990, 115: 263-264.

[7] Smith R.L.: *Amer.Lab.* March 1993, 28-29.

[8] Coren S.: The Left-Hander Syndrome: The Causes and Consequences of Left-Handedness.

[9] Life for lefties: from annoying to downright risky, *Smithsonian* December 1994.

[10] Mallows C.L., Tukey J.W.: An overview of the techniques of data analysis, emphasizing its exploratory aspects, in: Some Recent Advances in Statistics, pp. 113-172, Academic Press, New York, 1982.

[11] Kac M.: Enigmas of Chance, University of California Press, 1987.

[12] Rao C.R.: Statistics and Truth. Putting Chance to Work, Second edition, Council of Scientific and Industrial Research, New Delhi, 1992.