Research and Publication Ethics

Some issues have been already discussed in the context of scientific writing, but there are more. Note that "research" does not only refer to academic research, but most points apply to research in industry, too.

- First and foremost, cheating in research, in whatever form, is unacceptable. Cheating hurts science in many ways: wrong results are produced and other researchers build on them in good faith, the wrong people get reputation and funding, quality decreases, general trust in science is damaged, and so on.
- Peer reviewing is the process of evaluating scientific manuscripts for possible publication, or evaluating project proposals for possible funding. It is done by anonymous peers (other scientists in the same field). This is the main tool for ensuring quality of research. The idea is that only experts can give informed and motivated judgements. But peer reviewing can also have unwanted effects. How can they be avoided? Possible issues include: several kinds od bias, conflict of interest, lack of competence, sloppiness, unfairness.
- Papers received for peer reviewing are confidential. Moreover, it is not allowed to use the ideas learned there for one's own research. But it can be difficult to draw the line, if an idea was "in the air" and the reviewer was already close to it independently.
- Who should be authors of an article? Widely accepted rules are: Authors should be all persons that have substantially contributed to the work. All authors must have approved the publication, and all authors are responsible for the whole content. Persons that had only minor or auxiliary roles should be mentioned in an acknowledgement section. Authorship for any other than scientific reasons is not appropriate.

- Publication is a business and a "currency" for scientists. The careers of scientists crucially depend on their publication records. This pressure can lead to various temptations and forms of misconduct, in the worst case even criminal activities. How can fraud be avoided?

- A publication may not be really objective: A subject, a solution, a method or a conclusion may be pushed for other than scientific reasons. Authors may deliberately promise too much ("overselling"), perhaps in order to attract more funding.

- Research must be free, in order to facilitate both open discussion and production of new ideas. Researchers themselves are the most competent persons to decide what research subject are worth considering. On the other hand, research is costly, and society has an interest in a good usage of these investments. Some research topics are clearly more urgent than others because of societal needs. What is the right extent of external steering of research? What is a good balance between basic and applied research? Freedom of research also entails some responsibility of researchers to find the most promising research subject that will be important in the future. – One should be aware that the greatest technological breakthroughs are based on fundamental results in the natural sciences rather than on activities that were genuinely applied research from the beginning. The most striking example is quantum physics. It is the basis of virtually all modern technology but started as an endeavor to understand the inner structure of matter, without having any applications in mind. A myopic way of supporting only immediately useful research would miss such great opportunities.

Fallacies of Statistics

Wrong use of statistics means wrong conclusions and "knowledge", which can even be dangerous, especially in medicine and justice. Usually it does not origin from evil motives. Apparently our brains are not well suited for working with probabilities, and even mathematically educated professionals may have a poor understanding of probability.
Computer scientists should be aware of the most common fallacies of statistics, since also computer programs for data analysis and inference are based on statistical reasoning after all. Statistical methods that we implement

should at least be proper methods. Mathematicians, computer scientists, etc., are those people who could be consultants for others when it comes to statistical inference.

Interpretation Mistakes

We list some common mistakes, some of them being variations of the same underlying misconception.

- To start with a very elementary mistake: The median (the value in the middle) and the mean (average, expectation) are often confused or wrongly communicated.
- Pretended accuracy: If a study claims that 71.43% of persons were satisfied with some product, one should be suspicious: This number is very close to 5/7. Looking into the study one may find that in fact only 7 people have been asked! One must give a confidence interval, and for a very small sample this interval is broad, such that no reliable conclusion can be drawn at all, even if the sample was representative and homogeneous.
- A hypothesis is accepted based on a small sample that gives no evidence against the hypothesis, thereby forgetting that the sample is also too small to support the hypothesis. In the simplest case, a general claim is made ("all objects O have property P") and no counterexample is found among a few examples. This is not enough to conclude the hypothesis is true. (This mistake is also made if one believes that some algorithm is correct because it worked well on a few test instances, and a laborious correctness proof is not needed...)
- Data are oversimplified. Records consist of only a few key numbers, ignoring the complexity and multidimensionality of the subject, nevertheless they form the basis of an analysis.
- Conclusions are oversimplified. For instance, we may find that the expected loss (of property or even lives) in a certain type of disaster is rather low, and conclude that this is an acceptable low risk. But an expected value is only an average! It could be the product of a very small probability of the event and an unbearable high loss *if* the event occurs. Shouldn't one invest in prevention, although the probability of disaster is indeed very low?

- Another misconception regarding low probabilities: An event may have a low probability for each individual, but in a large enough population it will happen to *some* individuals, for no special reason. (Think of a lottery.) From low probability alone we cannot conclude that "this cannot have happened by chance, there must be something behind it". In fact, this type of fallacy has led to very bad cases of miscarriage of justice.

- Cherry picking: Unfavorable data are omitted from a single data set, a whole study with undesired results remains unpublished, etc. Here is a toy example: 10,000 people are asked to predict the results of 10 coin tosses. The probability to be always right is about 0.001. That means that still about 10 people are always right. We publish only their results and declare them prophets. We also explain that, "of course, prophecy is rare, but these few people have clearly demonstrated their prophetic skills". – Here the nonsense is obvious, but there are more subtle scenarios prone to this type of error. An interesting proposal to avoid suppression of unwanted results is that publication of *all* statistical studies, e.g., in medicine, should be obligatory, and scientific quality should be evaluated based on their methodology rather than on their positive findings.

- A related fallacy is data dredging. Patterns are seen in data without a previously formulated hypothesis to be tested. This is of no value because any large enough amount of data will exhibit some spurious patterns by chance. (Some patterns even appear with necessity, for combinatorial reasons. For instance, Ramsey's Theorem says that for every k, every large enough graph contains a clique or co-clique with k nodes.) A statistical finding must be tested on a second, independently chosen data set. However, just deriving a hypothesis from data is a correct procedure, if one does not claim immediately the truth of the hypothesis.

- A statistical result is carried over to another or a more general population the sample is not representative of, or the sample is biased towards a subgroup. (By catching fish with a net with meshes of a certain width one will only find fish of at least that size, and one should not conclude that all fish species have this minimum size.) A sample can be biased because gathering data is more costly for some subgroups, or because people can decide whether they want to participate in a study or not.

- Correlations are taken for causal dependencies without thinking. But two events may also show correlation, for instance, because both are caused by a common third factor. Especially in complex big data analysis this is easy to forget. On the positive side, such relationships can be accurately modelled and analyzed by probabilistic networks: graphical models, Bayesian networks, etc., which is an important subject in Machine Learning.

- In medical studies, false causal relations can be avoided by comparing the effect of a treatment with a control group. Then the statistical fallacy is avoided, bu an ethical dilemma is incurred instead: Can we give people in a control group treatment, or prevent them from necessary treatment, knowing that this can harm those people, but also knowing that the scientific results will be important for many more other people? What are the ethical values steering such decisions?

- A more subtle phenomenon is Simpson's paradox: Conclusions from data may depend on the grouping of individuals. There exist many variations, for instance: Some trend appears within each group, but the whole population exhibits the opposite trend. A method X performs better than a method Y in every subgroup, but in the whole population the result is the other way round. (Numerical examples can demonstrate the sources of this, at first glance, incredible effect.) It can be tempting to choose a partitioning that supports a desired result. Or the analyst can be honest but unaware of the fallacy. It also arises the question what the "correct" grouping of a population is. Here we cannot discuss approaches; the point was mainly to draw attention to this paradox.

Misleading Graphs

We can gather much information instantly through our eyes. In the era of big data the importance of data visualization has only increased. Visualization is a nontrivial field in its own right, including psychological aspects but also algorithmic questions. (How can we draw a graph in the best way, according to some optimality criteria?)
But graphics can also be misleading, either on purpose or unintendedly, even if they are based on true data. A few "methods" to construct misleading graphs are:

- Comparable quantities get different scaling, to make some of them appear larger than they are. Distorted perspective or additional dimensions can have similar effects.

- An axis starts at some positive value rather than at zero, thus differences and changes appear larger than they are. Truncation is justified when small differences need to be displayed, but then the truncation should be clearly pointed out. In scatterplots, truncation can also cut away outliers without clearly stating this fact.