Why the Empirical Sciences Need Statistics So Desperately

Olle Häggström*

Abstract. Science can be described as a systematic attempt to extract reliable information about the world. The cognitive capacities of *homo sapiens* come with various biases, such as our tendencies (a) to detect patterns in what is actually just noise, and (b) to be overly confident in our conclusions. Thus, the scientific method needs to involve safeguards against drawing incorrect conclusions due to such biases. A crucial part of the necessary toolbox is the theory of statistical inference.

There exists a large and well-developed (but of course incomplete) body of such theory, which, however, researchers across practically all of the empirical sciences do not have sufficient access to. The lack of statistical knowledge therefore forms a serious bottleneck in the quest for reliable scientific advances. As has been observed by several authors in recent years, statistical malpractice is widespread across a broad spectrum of disciplines, including (but not limited to) medicine, cognitive sciences, Earth sciences and social sciences.

Here I will first try to describe the overall situation and provide some concrete examples, and then move on to discuss the more difficult issue of what can and needs to be done.

2010 Mathematics Subject Classification. Primary 62A01.

Keywords. Science, statistical inference.

1. Introduction

What is science? Despite what some adherents of Popperian falsificationism [26] may claim, it seems unlikely that we can find a single short definition of science that captures all important aspects. See, e.g., Haack [14] for a sensible discussion on some of its many facets. The complexity notwithstanding, I hope most of us can agree on the somewhat vague statement that science consists of systematic attempts by us humans to extract reliable information about the world around us.

Since science is carried out by humans, it is in practice dependent on our cognitive capacities. Evolution has equipped us with impressive abilities to observe and draw conclusions about the world around us, necessary for finding food and sexual partners and to avoid predators. On the other hand, since Darwinian evolution by natural selection is not a perfect optimization algorithm, it should not come as a huge surprise that we have some striking cognitive biases. Some of these form

*Work supported by the Swedish Research Council

serious obstacles to the scientific endeavor. In particular the following two spring to mind.

(a) The human pattern recognition machinery is often too trigger-happy, i.e., we tend to see patterns in what is actually just noise. There is a famous experiment where, because of this phenomenon, human subjects typically perform *worse* than pigeons and mice. The subject is faced with two lamps, and are asked repeatedly to predict which lamp will light up next. Unbeknownst to the subject, the lamps are turned on randomly, with a 0.8 probability for the lamp on the left to light up next, versus a 0.2 probability for the one on the right, in an i.i.d. sequence. Human subjects notice the asymmetry, but try to mimic the intricate pattern of lights, predicting the lamp on the left about 80% of the time, and ending up making the right guess $0.8 \cdot 0.8 + 0.2 \cdot 0.2 = 68\%$ of the time. Simpler animals quickly settle for guessing the most frequent lamp every time, getting it right 80% of the time. See, e.g., Hinson and Staddon [17] and Wolford et al. [29].

(b) In many situations, we tend to be overly confident about our conclusions. The following experiment is described by Alpert and Raiffa [2]; see also Yudkowsky [30]. Subjects are asked to estimate some quantity whose exact value they typically do not know, such as the surface area of Lake Michigan or the number of registered cars in Sweden. They are asked not for a single number, but for an upper bound and a lower bound together encompassing an interval with the property that the subject attaches a subjective probability of 98% to the event that the true value lies in the interval. If subjects are well-calibrated in terms of the confidence they attach to their estimates, then one would expect them to hit the true value about 98% of the time. In experiments, they do so less than 60% of the time, indicating that severe overconfidence in estimating unknown quantities is a wide-spread phenomenon.

Because of these cognitive biases and several others, we need, in order to perform good science, to set up various safeguards against our spontaneous tendency towards faulty and overconfident conclusions. Randomized, double blind and placebo-controlled clinical trials is a typical example of a formalized protocol for precisely this purpose. The theory of statistical inference offers plenty of others, including a variety of important techniques for telling pattern from noise and for quantifying the amount of confidence in a given conclusion that a given data set warrants – that is, for circumventing biases (a) and (b) above.

Statistical techniques are indispensable for doing high-quality and trustworthy science. Fortunately, the use of such techniques are wide-spread, to the point of permeating the empirical sciences. Unfortunately, they are often used in erroneous ways and in situations where they simply do not apply, leading to unwarranted conclusions.

In Section 2, I will try to argue the seriousness of the situation by pointing out some indications – some of them quite shocking – about how widespread this misuse is. In Sections 3 and 4 I offer a couple of concrete examples of erroneous application and interpretation of statistical arguments. In an unabashed attempt to catch the readers' attention, I take them from two of the most hotly debated (in public discourse) research areas: climate science and gender studies. Then, in Section 5, I will exemplify how the lack of statistical expertise in many empirical sciences has given room to a population of self-proclaimed and mostly self-taught statistical "experts" giving erroneous advice to their colleagues. Finally, in Section 6, I will offer a few thoughts on how it might be possible to improve the situation in the future.

2. Statistics in the empirical sciences

There is an abundance of anecdotal evidence suggesting the wide-spread malpractice of statistics across the empirical sciences. Bland [7] builds his statistics textbook largely around real-world examples of bad statistical practice. Every university statistician and every statistical consultant is familiar with the phenomenon. I would conjecture (not very boldly) that if we sample at random from the population of researchers in medicine, and interview them concering the meaning of a p-value (the statistical concept that, for better or worse, tradition dictates to be the most important in medical statistics), then at most one in ten would give a satisfactory definition, and more than half of them would commit the so-called *fallacy of the transposed conditional* by confusing $\mathbf{P}(\text{data} \mid \text{null hypothesis})$ and $\mathbf{P}(\text{null hypothesis} \mid \text{data})$. Mutz and Pemantle [23] observe the wide-spread use, particularly in political science, of so-called randomization checks in experimental studies, and they demonstrate that typically such practice is not only pointless but outright harmful, leading to "inferior model choices in the analysis of findings and unjustified interpretation of findings". And so on.

Systematic scientific evidence quantifying the extent of the malpractice is more sparse, but there are a few studies. A recent study is the one of Nieuwenhuis et al. [24] concerning to what extent neurobiological papers in a range of top journals commit a certain error in significance testing discussed by Bland [7, p 142–143], namely the following. If, say, some drug shows a statistically significant effect in some subgroup such as men, but not in some other subgroup such as women, then it may be tempting to draw the (unwarranted) conclusion that there is a statistically significant difference between the drug's effect on men and its effect on women. A separate significance test for the difference is in fact needed. Among the papers scrutinized by Nieuwenhuis et al., 79 jumped immediately to the unwarranted conclusion, whereas 78 used the correct procedure.

To my knowledge, the most impressive systematic treatment of the issue of how wide-spread bad statistical practice is, is the 2008 book by Ziliak and McCloskey [31]; see also the review in [15]. Ziliak and McCloskey discuss the phenomenon across a wide range of disciplines, but their main focus is on their own subject economics, and the core of the book is a couple of chapters in which they discuss their findings concerning statistical (mal-)practice in the prestigious journal *American Economic Review* through the 1980's and 1990's. They looked closely at all of the

369 papers through that period that employed regression analysis, and noted the frequency of various statistical errors. Their findings are discouraging throughout. Here, let me mention just one: what Ziliak and McCloskey call *sizeless science*.

To understand the concept, imagine, as an example, that a new drug for reducing blood pressure is being tested and that the fact of the matter is that the drug does have a positive effect (as compared to placebo) but that the effect is so small that it is of no practical relevance to the patients' health or well-being. If the clinical study involves sufficiently many patients, then the effect will nevertheless with high probability be detected, and statistical significance will be obtained. The lesson to learn from this is that in a medical study, statistical significance is not enough for the result to be interesting – the detected effect also needs to be large enough to be *medically significant*. And similarly, empirical studies in economics (psychology, geology, etc.) need to consider not just statistical significance but also economic (psychological, geological, etc.) significance. It turns out that that many researchers are so obsessed with statistical significance that they neglect to ask themselves whether the detected discrepancies are large enough to be of any subject-matter significance, and this is precisely what Ziliak and McCloskey call sizeless science. Of the 369 *American Economic Review* papers, 276 committed sizeless science – three quarters of the papers!

The precise extent to which statistics is abused in all of the myriad possible ways, and across all of the empirical sciences, remains unknown, but it seems clear that the abuse is wide-spread. This is serious stuff, as suggested by the Ziliak's and McCloskey's [31] somewhat polemical subtitle: *How the Standard Error Costs Us Jobs, Justice and Lives*. In the context of medicine, Bland [7, p 4] writes:

> Bad statistics leads to bad research, and bad research is unethical. Not only may it give misleading results, which can result in good therapies being abandoned and bad ones adopted, but it means that patients may have been exposed to potentially harmful new treatments for no good reason.

3. A climate science example

Some readers may recall how, in the wake of the so-called Climategate incident, where a server at the Climate Research Unit at the University of East Anglia was hacked and thousands of emails and computer files were made public by copying them to various Internets, the debate between climate denialists and climate scientists reached unprecedented intensity in the winter of 2009/2010. In the morning of February 16, 2010, I was brutally awoken by the following message from the news show *Dagens Eko* of Swedish public radio.

> The critical examination of the United Nations climate panel IPCC and their work continues, and now it is one of the leading characters who says in an interview with British BBC that he is no longer so sure about global warming. It is none other than the director of CRU at the

> University of East Anglia, professor Phil Jones, who says that there are no statistically significant proofs of global warming during the last 15 years. [11, my translation]

The following discussion of the background to the news broadcast is mainly taken from Häggström [16]. The BBC interview with Phil Jones is available at [6], and contains the following exchange:

> Q: How confident are you that warming has taken place and that humans are mainly responsible?
>
> A: I'm 100% confident that the climate has warmed. As to the second question, I would go along with IPCC Chapter 9 – there's evidence that most of the warming since the 1950s is due to human activity.

Jones' answer here contrasts sharply to the view that was attributed to him the the *Dagens Eko* report. To understand better how *Dagens Eko* came to write what they did, we must turn to a different exchange from the same interview:

> Q: Do you agree that from 1995 to the present there has been no statistically-significant global warming?
>
> A: Yes, but only just. I also calculated the trend for the period 1995 to 2009. This trend (0.12C per decade) is positive, but not significant at the 95% significance level. [...] Achieving statistical significance in scientific terms is much more likely for longer periods, and much less likely for shorter periods.

The time series Jones is refering to here is the so-called HadleyCRUT3 series, which is one of the main global average temperature time series available [9]. Due to chaotic weather fluctuations, this series shows large fluctuations on top of the long-term trend that we call global warming. The shorter the time interval we choose to focus on, the worse signal-to-noise ratio we get, and the harder it is to detect the trend. If we ignore enough of our data by focusing on a sufficiently small time window, we will almost always "be able to fail" to detect the trend. In this case, the year 1995 was cherry-picked to be the earliest X such that from year X to year 2009 in the annual HadleyCRUT3 data, the trend is not statistically significant at the 95% level. The p-value in this case turned out to be 0.072, which is more than 0.05; hence Jones' agreement that "from 1995 to the present there has been no statistically-significant global warming".

But what can we actually make of this figure $p = 0.072$? It certainly cannot be interpreted as the probability of the null hypothesis – which in this case means zero trend – because, first, it would be to commit the fallacy of the transposed conditional discussed in Section 2, and, second (but relatedly), it would be insane to base such a probability judgement merely on part of a single time series, ignoring all other data and everything we know about, e.g., the greenhouse effect and the rest of physics.

This has been pointed out many times, and indicates that the wide-spread (especially in the blogosphere) juggling with p-values in connection with recent years' global temperature averages can hardly be viewed as a central issue in the science of climate change. But the irrelevance of these p-values is even greater than this. When you push the linear regression button in your favorite statistical software package (as I am sure Phil Jones did), the p-value it produces is based on a null hypothesis which stipulates not only a zero trend, but also that deviations from the trend line consists of Gaussian white noise. This means in particular that the deviations from the trend line at consecutive times are uncorrelated. Even if the trend is zero, a deviation from this white noise assumption may drastically alter the distribution of the p-value statistic.

To illustrate the problem, I decided to reconstruct Jones' linear regression of the 1995–2009 time series, not only with annual data as Jones did, but also with semiannual and monthly data. The estimated trend is almost the same for the three choices of time resolution, but the p-values differ drastically: I got $p = 0.072$ for annual data, $p = 0.017$ for semiannual data, and a seemingly impressive $p = 0.0000004$ for monthly data. But it is clear that the "great smallness" of the last p-value is due to the implicit assumption of uncorrelated noise. When refining from annual to monthly data, the statistical software package thinks it has 12 times as much information to work with, whereas in fact much of this information is, due to strong correlations on monthly scales, only apparent.

Thus, in order for the p-value based on annual data to mean what it is supposed to mean, we need annual deviations from the trend line to be close to uncorrelated. This seems unlikely to be the case, given the variations and oscillations on biennial to decadal time scales that exist both in external forcings (such as the sun spot cycle) and the internal workings of the climate system (such as the El Niño-Southern Oscillation). So to use the annual data in this way, we would need to model the autocorrelation function. This seems to require deep understanding of both climatology and statistical modelling.

In my view, the main moral in this story is the importance, when carrying out a standard statistical procedure, of knowing what are the underlying assumptions that are needed to justify the procedure. I have the impression that failure to do so is a very common shortcoming across many empirical sciences. In climate debate, the mistake that Jones appears to have made – to accept p-values for temperature trends produced by a straightforward linear regression without bothering to address the crucial issue of autocorrelation – is disturbingly common on both sides of the science vs denial fence.

What, then, is my verdict on the key players in this story? The way that *Dagens Eko* chose to represent the BBC interview (sadly representative of much of the reporting of the same thing in blogs and mainstream media) I can only condemn. And indeed, on May 31, 2010, *Dagens Eko*'s broadcast was convicted by the *The Swedish Broadcasting Commission* for its errors and lack of objectivity [11].

I am less certain about how to judge Phil Jones' decision in the BBC interview to play along in the pretty much meaningless game of p-values and statistical significance in recent years' global temperature data. One thing that might be put forth to his defence is to ask why he should be singled out for criticism when so many others have taken part in the same game. Another defence might be to point out that he was not writing a scientific paper, only answering an interview question, so cannot be held to the same scientific standards as in the peer-reviewed literature. I'm not sure, however, about the validity of either of these. Communication of science to the general public is an important matter, and one might argue that it is just as important to get things right in BBC as in *Nature*.

4. A gender studies example

In this section, I will briefly a discuss series of papers by Janet Hyde and coauthors [18, 19, 20] on gender differences in cognitive abilities – papers that suffer from errors and sloppy thinking of a kind that I would like to think had been impossible if the authors had involved a qualified statistician in their work. I have reason to be especially interested in the first two papers [18, 19], because these formed large part of the scientific basis for the new gender policy at my own university Chalmers in 2007; the unwarranted conclusions in [18, 19] were, in typical "Chinese whispers" fashion, magnified in the local document outlining this scientific basis [1].

Hyde [18, p 581] states her theoretical point of departure as follows:

> The gender similarities hypothesis holds that males and females are similar on most, but not all, psychological variables. That is, men and women, as well as boys and girls, are more alike than they are different.

To ask whether two categories are "more alike than they are different" sounds to me as hopelessly vague and dependent on one's point of view as asking whether 100 is a large number or a small one. To her credit, however, Hyde does provide a concrete definition of "similar" (more on that below).

On the other hand, the notion "most psychological variables" remains completely unclear. In the large meta-analysis of studies of gender differences done in [18] (and reported again in [19]), 128 psychological variables, ranging from "reading comprehension" and "spatial visualization" to "talkativeness" and "attitudes about casual sex", were considered. 4 of the 128 were rejected because of too large uncertainties in the estimates, but among the 124 that remained, more than three quarters – 96 out of 124 – exhibited gender differences that were classified as either "close-to-zero" or "small". Superficially, this may seem to support the notion that "most psychological variables" exhibit small gender differences. On second thought, however, one may wonder why the 96 characteristics with "close-to-zero" or "small" shold count for more than those 28 with "moderate", "large" or "very large" differences. In the absence of any explanation, it seems that Hyde takes for granted that, e.g., "spatial visualization" should carry the same weight as "attitudes about casual sex" in an overall assessment of psychological characteristics (surely an example of comparing apples and oranges!). Furthermore, the list of 124 characteristics obviously does not exhaust the space of human psychological characteristics. Hyde offers no argument for why those 124 characteristics should be considered in any way representative of that larger space, and in the absence of such an argument her claims about "most psychological variables" carries about as much weight as if I would walk around knocking doors and interviewing neighbors on my street in Gothenburg, and then report that 94% of all people in the world speak fluent Swedish.

Concerning the notion of gender "similarity" with respect to a given characteristic, Hyde does have (as mentioned above) a concrete definition. Individuals are scored along some dimension, and we compute $d = (M_M - M_F)/s_w$, where is M_M is the average score for males, M_F is the average score for females, and s_w is the pooled within-sex standard deviation. Depending on the value of $|d|$, the difference is classified as "close-to-zero", "small", "moderate", "large" or "very large", with cutoffs at $|d| = 0.10, 0.35, 0.65$ and 1.00.

Hyde's project rests on the silent assumption that a small value of d implies that the distributions of scores for males and for females are close to identical. If we could assume that within-sex distributions are Gaussian with the same standard deviation, then there would be some merit to this view. If, on the other hand, we leave room for the possibility that the two within-sex distributions have different standard deviations (which in some cases there is ample evidence for; see, e.g., [28]) and/or different shape, then the distributions can be vastly different even when $d = 0$. In general, the mean value does not determine a probability distribution.

The following will serve as a summary verdict on the extent to which Hyde [18] succeeds in providing support for the gender similarities hypothesis: For such support to be at all possible, the hypothesis needs to be well-defined. Unfortunately, it is only partly well-defined, and moreover, even for the parts that are well-defined, the evidence in [18] is very weak – especially compared to how the results are marketed in [18] and in secondary publications such as [19] and [1].

Let me now move on to the later paper by Hyde and Mertz [20] on gender and mathematical performance. Here, the authors seem to have picked up on the criticism that the mean does not determine a distribution, because they go on to complement the study of means with a study of the upper tail of the distributions; this may be especially relevant in discussing unbalanced recruitment to higher education or unequal represenation among university faculty. One way to study the (extreme) upper tail is to look at data concerning the proportion of female students in national teams in the International Mathematical Olympiad. Using data from 30 countries during 1978–2008, they find that the percentage of females in the teams for different countries and different decades vary between 0 and 22%, and much more than can be explained by chance variations. This suggests that cultural factors and gender inequality may influence the mathematics achievements of males and females, a conclusion that is supported by a positive correlation between female representation in the IMO teams and the so-called World Economic Forum Gender Gap Index. It must be noted, however, that what the data suggest is only that the between-country variation, between 0 and 22% and with a median around 3 or 4%, can be explained in this way; the much larger discrepancy up to the gender-symmetrical 50% is left unexplained. Yet the authors write, with shocking disregard of what their data actually say:

> Thus, we conclude that gender inequality [...] is the primary reason fewer females than males are identified as excelling at the high and highest levels in most countries. [20, p 8806]

5. When the cat's away the mice will play

In Section 6 I will argue that we statisticians need to increase our presence in the empirical sciences. Inevitably, statistical methods will still be used – and often, as we've seen, abused – even in our (relative) absence. Consequently, the methodological discussions in a given empirical discipline will often involve statistical issues, but without participation of statisticians these discussions risk being uninformed. In this section I will briefly point out two examples of researchers who take on the role of statistical "authorities" in their respective fields, and how their discussions suffer from lack of statistical training and from their unfamiliarity with the statistical literature: Ambaum [3] in climate science, and Kruschke [21] in cognitive science.

In fact, even the book by Ziliak and McCloskey [31], which I praised in Section 2, exemplifies to some extent such shortcomings. Their description of the alarming situation is mostly very good, whereas their discussion of what needs to be done is not on the same level. When, on p 240, they pose the rhetorical question "If null hypothesis significance testing is as idiotic as we and its other critics have so long believed, how on earth has it survived?" (quoted and criticized also in [15]), they go a bit overboard their mission against sizeless science. While they are absolutely right that single-minded focus on statistical significance is bad practice, throwing out the use of significance tests would be a mistake, because it is a crucial tool for concluding with confidence that what we see really is a pattern, as opposed to just noise (cf. item (a) in Section 1). To be able to conclude that we have reasonable evidence in favor of an important deviation from the null hypothesis, we need *both* statistical *and* subject-matter significance.

5.1. Ambaum. Ambaum [3] is also (like Ziliak and McCloskey) highly critical about significance testing, but on different and rather confused grounds. He does warn, rightly, about the fallacy of the transposed conditional (discussed above in Sections 2 and 3), and goes on to provide some illustrative calculation to show that if a Bayesian approach had been taken, the posterior probability $\mathbf{P}$(null hypothesis | data) might have ended up very different from the p-value resulting from the significance test. In particular he shows that, no matter how small the p-value is, it is possible, by putting sufficiently large prior probability on the null hypothesis, to make the prior overwhelm the data and make the posterior probability $\mathbf{P}$(null hypothesis | data) as close to 1 as we wish.

So far so good, but when Ambaum goes on to conclude that (quoting the final sentence of his paper) "significance testing of a single experiment alone cannot be used to provide quantitative evidence", he is just wrong. A posterior probability is not the only way to quantify evidence, and in fact p-value is another such quantification, the logic being as follows. If the p-value p is really small, then we must conclude that either the null hypothesis is false, or something very unlikely has happened. The smaller p is, the greater the "miracle" needed to save the null hypothesis, and the stronger reasons we have to think the null hypothesis is false.

One way to see the defect of Ambaum's argument is that, if it were true, it would show that no data whatsoever (short of logically entailing some scientific hypothesis – a rare phenomenon) can provide evidence for or against any hypothesis, because by carefully selecting a prior distribution we can obtain a posterior that points in the opposite direction from the suggested inference. This borders on radical skepticism, which is wholly useless as a philosophy of science. At first sight, one might be seduced to think that a Baysian procedure solves Ambaum's quest for quantitative evidence, because it produces a value of $\mathbf{P}$(null hypothesis | data), as desired. On second thought, however, we realize that Ambaums's argument has just as much force against this inference as against significance testing, because we *could have* started from a *different* prior producing a radically different value of $\mathbf{P}$(null hypothesis | data). Further criticism of Ambaum's paper is provided by Guttorp and Häggström [13]; see also [4] for a reply.

5.2. Kruschke. The paper by Kruschke [21] on statistics in the cognitive sciences parallels Ambaum [3] in that its main message is a fierce attack upon null hypothesis significance testing. His arguments are, however, different, and somewhat more sophisticated.

Suppose that we wish to test whether a coin is fair, i.e., whether heads (H) and tails (T) have the same probability $\frac{1}{2}$. Suppose furthermore that the data we have is the following sequence of coin tosses:

T T H T T T T T H

It turns out that the standard theory of null hypothesis significance testing yields different p-values depending on whether the experiment performed was "toss the coin ten times" or "toss the coin until H has come up twice". This may seem surprising, but the phenomenon is a standard textbook example in the theory of statistics; see, e.g., [8, p 278–272]. Kruscke [21] considers similar examples, concludes that a fatal flaw of significance testing has been discovered, and informs the reader about this in a series of vigorous proclamations, beginning with:

> This brief article assums that you, dear reader, are a practitioner of *null hypothesis significance testing.* [...] Did you know that the computation of the p value depends crucially on the covert intentions of the analyst, or the analyst's interpretation of the unknowable intentions of the data collector? [21, p 658, italics in original]

Exactly what "unknowable" means is not explained, nor what exactly is so fatal about different experiments leading to different conclusions.

What Kruschke is doing in his paper, without using the term and apparently with no awareness of the many decades of discussion and debate on the topic, is to defend the highly controversial notion which has been known at least since

1962 as *the strong likelihood principle*, and which says that if two experiments yield proportional likelihood functions, then the conclusions drawn from them should be identical. See, for instance, Savage [27], Cox and Hinkley [10], Casella and Berger [8], Barndorff-Nielsen [5], Pawitan [25] and Mayo [22]. Kruscke's paper would have been much more useful and instructive to his cognitive science colleagues if he, rather than inventing his own rhetoric from scratch, had taken into account and built upon this long intellectual tradition.

Kruschke's preferred alternative to significance testing and related frequentist procedures is Bayesian statistics. To advocate Bayesian statistics is certainly a legitimate mission, but by refusing to seriously discuss the Achilles heel of Bayesian statistics – the choice of prior distribution – Kruschke does his readers a major disfavor.

6. What to do?

Identifying a problem is one thing. Providing a remedy is quite another, and – in this case – rather more difficult. What can be done in order to improve statistical practice in the empirical sciences? I have no miracle solution, and can offer only a few fairly obvious thoughts.

The first solution that comes to mind is to teach mathematical statistics to future researchers in all empirical sciences, on undergraduate as well as graduate level. Of course this is to a large extent already being done, but (as is clear from what has been said in earlier sections) these teaching efforts are by no means sufficient. So we could – and should – intensify them. This is by no means a novel idea. University statisticians have been attempting to do this for many decades, sometimes with local success, but global progress appears modest.

Despite the difficulties, we need to persist in these attempts. Perhaps the greatest obstacle is the PR image of mathematical statistics. To equip, say, future biologists with a reasonable amount of statistical competence, we need, first, to convince those in charge of the various biology programs at our universities that solid knowledge of statistical theory and method is important to biologists. Once that is done, we have formal access to the biology students, but not automatically to their whole-hearted attention, which is needed for effective learning. A typical biology student is at the university because she is interested in biology, and is likely to view courses in statistics as an annoying distraction from what she considers to be the real thing. This is a barrier that we need to overcome by showing students how essential statistics is to their chosen discipline – a pedagogical challenge that we must find better ways to handle.

The problematic situation outlined in this paper will, however, not be solved via undergraduate and graduate education in the empirical sciences alone. We cannot realistically hope to convey more than one or two basic courses in statistics to the bulk of future researchers in the empirical sciences. Their understanding of statistics will therefore, in most cases, still be too superficial to ensure high-quality treatment of the statistical parts of their research projects.

I therefore see no way around the conclusion that university statisticians need to get much more involved in the research of the empirical sciences. Of course, many of my statistician colleagues already do an excellent job in precisely such collaboration, but this activity needs to be scaled up drastically. The ideal, in my opinion, is that every research group working with real-world data should interact regularly with some highly qualified statistician. I realize that this vision is somewhat utopian and cannot be realized in a year or even a decade, but this is what we need to strive for.

Also for this kind of expansion of statistics as a university discipline – towards increased participation in interdisciplinary research – the PR image of our subject is of vital importance. Many (perhaps most) researchers think of statistics as a p-value calculation procedure that is basically independent of the rest of the research, and as a simple add-on feature to be applied when all the real work is done. What they overlook is the need for careful and often quite deep thinking about statistical modelling and assumptions needed for this or that statistical procedure to be applicable – a task that typically requires expert knowledge both of mathematical statistics and of the empirical science at hand, and which therefore calls for cross-disciplinary collaboration. Gill [12] describes this kind of underestimation of what statistics is all about well:

> Real world problems are often brought to a statistician because the person with the question, for some reason or other, thinks the statistician must be able to help them. The client has often already left out some complicating factors, or made some simplifications, which he thinks that the statistician doesn't need to know. The first job of the consulting statistician is to find out what the real question is with which the client is struggling, which may often be very different from the imaginary statistical problem that the client thinks he has. The first job of the statistical consultant is to undo the pre-processing of the question which his client has done for him.

To me it is obvious that we need to teach our colleagues in other disciplines that high-quality statistical inference requires the kind of broader scope of statistics where modelling issues are taken seriously. As long as they think of statistics departments as "department for calculating mean-square deviations and p-values", they will not realize that they need our help, because they have computer programs that can quickly do this calculation for them.

There are of course other measures, besides this huge PR and pedagogical task, that can and should be taken in parallell. For instance, if we could convince our universities to earmark substantial amounts of money for researchers in empirical sciences to consult statisticians, a lot would be gained.

None of this will work on any large scale, however, unless the supply of statisticians is increased substantially. The number of statisticians we educate, especially on graduate level, needs to be several times larger than today in order to meet the demand that my utopian vision will imply. And for this we need to attract good students, so the PR image of our discipline turns out to be crucial on yet another level. I hope that this paper may serve as a small step towards improving this image, or that at least it doesn't make it any worse.

Acknowledgement. The set of people who have inspired and influenced me significantly in topics discussed here is too lagre to allow giving a full list. Let me nevertheless mention Johanna Andersson, Christian Azar, Kimmo Eriksson, Peter Guttorp, Jacques de Maré, Staffan Nilsson, Peter Olofsson, Marita Olsson and Bo Rothstein as particularly helpful and influential for this text. Needless to say, none of them (not even my dear Marita) can be held responsible for any of my opinions or errors.

References

[1] J. Almer, G. Berndes, L. Brink, P. Gluch, U. Jarfelt, M. McKlevey and H. Rootzén, Jämställdhet i teori och praktik på Chalmers, 2007.
`https://www.chalmers.se/insidan/SV/om-chalmers/moten/fakultetsradet/ovriga-dokument_1/downloadFile/attachedFile_2_f0/Fakultetsradsrapport-jamstalldhet.pdf?nocache=1250842068.11`

[2] M. Alpert and H. Raiffa, A progress report on the training of probability assessors, in *Judgment under Uncertainty: Heuristics and Biases* (eds D. Kahneman, P. Slovic and A. Tversky), Cambridge University Press, Cambridge, 1982, 294–305.

[3] M. H. P. Ambaum, Significance tests in climate science, *J. Climate* **23** (2010), 5927–5932.

[4] M. H. P. Ambaum, Reply, unpublished manuscript, 2011
`http://www.met.reading.ac.uk/~sws97mha/Publications/reply_ambaum.pdf`

[5] O. E. Barndorff-Nielsen, Diversity of evidence and Birnbaum's theorem, *Scand. J. Stat.* **22** (1995), 513–515.

[6] BBC News, Q&A with Professor Phil Jones, February 13, 2010.
`http://news.bbc.co.uk/2/hi/science/nature/8511670.stm`

[7] M. Bland, *An Introduction to Medical Statistics*, 3rd ed., Oxford University Press, Oxford, 2000.

[8] G. Casella and R. L. Berger, *Statistical Inference*, Duxburry Press, Belmont, CA, 1990.

[9] Climate Research Unit temperature data page, downloaded in October 2011
`http://www.cru.uea.ac.uk/cru/data/temperature/`

[10] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*, Chapman and Hall, London, 1974.

[11] T. Fredriksson, Klimatskandalens huvudperson tvekar, *Sveriges Radio*, February 15, 2010.
`http://sverigesradio.se/sida/artikel.aspx?programid=83&artikel=3445547`

[12] R. Gill, The Monty Hall problem is not a probability puzzle (It's a challenge in mathematical modelling), *Stat. Neerl.* **65** (2011), 58–71.

[13] P. Guttorp and O. Häggström, Sense and nonsense about significance testing, unpublished manuscript, 2011
`http://www.nrcse.washington.edu/NordicNetwork/reports/Significance.pdf`

[14] S. Haack, *Defending Science – Within Reason: Between Scientism and Cynicism*, Prometheus Books, Amherst, NY, 2003.

[15] O. Häggström, Book review: The Cult of Statistical Significance, *Not. Am. Math. Soc.* **57** (2010), 1129–1130.

[16] O. Häggström, Global uppvärmning och statistisk signifikans, *Qvintensen* 1/2011.

[17] J. M. Hinson and J. E. R. Staddon, Matching, maximizing and hill-climbing, *J. Exp. Anal. Behav.* **40** (1983), 321–331.

[18] J. S. Hyde, The gender similarities hypothesis, *Am. Psychol.* **60** (2005), 581–592.

[19] J. S. Hyde and M. C. Linn, Gender similarities in mathematics and science, *Science* **314** (2006), 599–600.

[20] J. S. Hyde and J. E. Mertz, Gender, culture and mathematics preformance, *P. Natl A. Sci. USA* **106** (2009), 8801–8807.

[21] J. K. Kruschke, Bayesian data analysis, *WIREs Cogn. Sci.* **1** (2010), 658–676.

[22] D. G. Mayo, An error in the argument from conditionality and sufficiency to the likelihood principle, in *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science* (D. G. Mayo and A. Spanos eds.), Cambridge University Press, Cambridge, 2010, 305–314.

[23] D. Mutz and R. Pemantle, The perils of randomization checks in the analysis of experiments, preprint, 2011
`http://www.math.upenn.edu/~pemantle/papers/Preprints/diana.pdf`

[24] S. Nieuwenhuis, B. U. Forstmann and E.-J. Wagenmakers, Erroneous analyses of interactions in neuroscience: a problem of significance, *Nat. Neurosci.* **14** (2011) 1105–1107.

[25] Y. Pawitan, *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford, 2001.

[26] K. R. Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge* Routledge, London, 1963.

[27] L. J. Savage, *The Foundations of Statistical Inference*, Methuen, London, 1962.

[28] S. Strand, I. J. Deary and P. Smith, Sex differences in cognitive abilities test scores: a UK national picture, *Brit. J. Educ. Psychol.* **76**, 463–480.

[29] G. Wolford, M. B. Miller and M. Gazzaniga, The left hemisphere's role in hypothesis formation, *J. Neurosci.* **20** (2000), RC64.

[30] E. Yudkowsky, Cognitive biases potentially affecting judgement of global risks, in *Global Catastrophic Risks* (eds N. Bostrom and M. Cirkovic), Oxford University Press, Oxford, 2008, 91–119.

[31] S. T. Ziliak and D. McCloskey, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*, The University of Michigan Press, Ann Arbor, MI, 2008.

Olle Häggström, Mathematical Sciences, Chalmers University of Technology, 412 96 Göteborg, Sweden

E-mail: olleh@chalmers.se