

Randomness and the Games of Science

Jelle J. Goeman

Abstract Recently it has become clear that too many findings reported in the scientific literature are irreproducible. We study the causes of this phenomenon from a statistical perspective. Although a certain amount of irreproducible research is unavoidable due to the randomness inherent to scientific observation, two related phenomena conspire to increase the proportion of such findings: publication bias, i.e. the custom that negative findings are usually not published, and confirmation bias, i.e. the human inclination to interpret observations in a way that confirms prior beliefs. Both biases are poorly held in check in the current scientific publication model in which there is no explicit role for the views of a critic, i.e. a scientist with opposing theoretical views. We argue that if researchers are able to play the critic's role imaginatively, they will publish science of higher methodological quality that is not only more reproducible, but also more relevant for theory. To allow for this, we must promote a different view on statistical methodology, seeing statistics not as the gatekeeper of scientific evidence, but as a language scientists may use to discuss uncertainty when they talk about the implications of observations for theory.

1 Introduction

In 2009, a highly remarkable scientific experiment was performed by Bennett, Baird, Miller and Wolford, four American brain researchers. They used functional magnetic resonance imaging (fMRI), a brain imaging technique, to determine which brain areas respond to emotional stimuli in a test subject. The subject was shown

This text is based on my inaugural lecture “Toevalstreffers” (in Dutch), held on June 20, 2014 at Radboud University Nijmegen.

J.J. Goeman (✉)

Faculty of Medical Sciences, Radboud University, Nijmegen, The Netherlands
e-mail: jelle.goeman@radboudumc.nl

several emotionally laden pictures and was asked to verbalize the emotion shown. The display of pictures was alternated with rest, and by comparing the brain readings between exposure and rest, the researchers were able to clearly identify a brain area that showed a response to the stimulus offered (Bennett et al. 2011).

What was so remarkable about this experiment? Certainly not the idea of measuring brain response to pictures using fMRI; this had been done countless times by other researchers in the past. Also not the statistical methods used to find the relevant brain regions by comparing exposure and rest states; the same techniques had been used in many influential publications in brain imaging before. The originality of the study lay in the choice of the test subject. This was not, as usual, a human, but an Atlantic salmon. Moreover, the salmon was stone dead, having been bought in the local supermarket on the very morning of the experiment.

The paper describing the experiment, when finally published, created quite a storm among brain imaging researchers, and was credited with the Ig Nobel prize in 2012.¹ Apparently, standard imaging techniques with standard analysis methods could produce clearly nonsensical results. In the future, the authors of the salmon experiment argued, more stringent statistical methods should be used in fMRI research that have a smaller risk of false positive results. As a result of this paper, methodological standards in brain imaging have increased substantially in the last few years. However, the salmon experiment not only had implications for future research, but also casts doubt on past results. How many published papers on brain imaging would have used the same methods as the salmon experiment to come to equally wrong conclusions? How reliable, then, is the brain imaging literature?

Other authors in other fields have also raised questions about the reliability of the scientific literature. Prominent among these is the epidemiologist John Ioannidis with his (2005) essay “Why most published research findings are false.” Ioannidis argued quite generally from statistical arguments that a large proportion of the results presented in medical publications can be expected to be wrong. This proportion may differ between subfields of medicine, and depends on several factors, which we will come back to later. He comes to several surprising conclusions, among which one is that ‘hot’ scientific fields, in which many teams work on the same problems, and scientific breakthroughs are eagerly anticipated, are especially prone to produce unreliable findings. Consequently, results in high status journals, such as *Nature* and *Science*, would be especially unreliable.

Ioannidis’ theoretical arguments have been confirmed by researchers that have actually tried to reproduce published scientific results. The results of such attempts have varied greatly. In psychology, where Ioannidis’ arguments can be expected to hold as well, the journal *Social Psychology* published a special issue that reported replications of 13 recent studies (Klein et al. 2015). In 10 out of 13 cases, the effects reported in the original papers were found again, although often with a smaller magnitude. One study was on the borderline, replicating with a very small effect. The other 2 studies (14 %) failed to reach the same conclusions. More dramatic was

¹The Ig Nobel Prizes honor scientific achievements that make people laugh, and then think.

the experience reported by Begley and Ellis (2012), scientists working at Amgen, a pharmaceutical company in California, who tried to replicate the results of many ‘landmark’ papers describing promising drug targets. They failed in no less than 47 out of 53 cases (89 %). Statisticians have tried to quantify the proportion of unreliable results in larger chunks of the scientific literature. Jager and Leek (2014) estimated the proportion of unreliable results in the whole medical literature at 14 %. Statisticians commenting on this effort almost invariably stressed that the percentage is very likely to be an underestimate, and possibly a severe one. The discussion of the reliability of scientific results has also reached the popular media, where regularly a bleak image is sketched of science in crisis. As to the cause and prevention of the unreliability scientific results, different opinions are voiced. Two competing explanations dominate the debate.

According to the first, scientists striving for fame and status deliberately engage in ‘sloppy science’. They make their results look better than they are in order to publish them in higher ranking journals. Results are not fabricated, and ‘sloppy science’ is not the same as downright fraud, but ‘sloppy’ scientists are accused of wilfully neglecting proper checks and validations in order to publish more quickly. In variants of this argument, scientists are the victims rather than the perpetrators, as they are forced into their behavior by external institutional pressures. Because of savage competition between scientists and the demands from universities and funding agencies for ever longer lists of publications, scientists would have no choice but to engage in this type of dubious behavior.

A second explanation does not blame the scientists, but the methods they use. Since statistical methods are supposed to protect scientists against spurious findings, a high incidence of unreliable scientific results clearly indicates a design error in these methods. The type of statistical method most commonly denounced is the hypothesis test and the p -value, which, as critics point out, are frequently misunderstood and often used in a wrong way. Some authors argue that these methods should be banned altogether, a policy recently implemented by the journal *Basic and Applied Psychology* (Trafimow and Marks 2015). Some commentators advocate different statistical methods instead, e.g. Bayesian statistics. Others such as the editors of *Basic and Applied Psychology* simply advise against all advanced statistical methods, advocating simple descriptive statistics instead.

Interestingly, these two explanations suggest radically different solutions to the problem of unreliable results in science. If ‘sloppy science’ is the problem, scientists should be forced to adhere more strictly to proper statistical methodology. They should be kept in check by statisticians, who would then be cast into the role of policing various fields of science. Conversely, if statistics itself is the problem, the solution would be to free scientists from the influence of statisticians as much as possible. Scientists would then either convert to a completely different way of doing statistics, or just report their findings unencumbered by any need to demonstrate statistical significance.

More statistics or less? Which is better for the advancement of science? Which of the two explanations for the current flood of irreproducible research is the right one? Discussing the second explanation first, we will first review where

randomness and irreproducibility in science come from, and discuss the way statistical methods deal with this. We will explain that randomness is inherent to scientific observation, and that statistics provides scientists with a way to discuss the implications of this randomness on their experiments. Next, to shed light on the first explanation, we discuss several models for the way scientists interact with each other. We emphasize the important role of critics with different theoretical views in scientific inquiry, arguing that statistical reasoning is an essential part of the dialogue between scientist and critic. Finally, we look at the current publication model for reporting scientific results, and how it encourages a different, much more mechanical view of statistics. In this view statistics is seen as an arbiter of truth rather than as a language for discussing uncertainty. Rejecting both of the explanations given above, we will argue that it is primarily this distorted view of statistical methods that explains the current reproducibility crisis in science.

2 Randomness in Science

“Everything changes and nothing remains still; you cannot step twice into the same stream” said the Greek philosopher Heraclitus, stressing the ever-changing nature of reality. This truism applies very much to research, where no two experiments ever return exactly the same result: different subjects respond differently to treatment, and measurements are always variable. Randomness is inherent to scientific observation.

Randomness, moreover, is bound to produce flukes. Since scientific observation is subject to variability, seemingly meaningful patterns that the researcher observes may well be one-time events rather than repeatable ones. For example, the patients in a treated group may happen to recover very well, while the patients in the untreated group do poorly, all because of their own particular reasons not related to the treatment. To the researcher this may suggest a strong effect for a treatment that is in reality not effective. When the experiment is subsequently replicated by the same group of scientists or by a different one, the spurious patterns are very likely not observed again. Irreproducible results, therefore, are a fundamental consequence of randomness in scientific observation, and are unavoidable even in the most meticulous and honest scientific practice. We can, however, try to limit the frequency of the occurrence of such result. This is what statistics tries to do.²

Statistical theory makes an explicit distinction between the *sample*, i.e. the concrete observations the researcher has in hand, and the *population*, i.e. a larger pool that these observations were drawn from. For example in a preselection poll, the

²This statistical view on (lack of) reproducibility is a limited one. There are of course many other ways in which research can be irreproducible, for example because of systematic measurement error, such as when the CERN-OPERA group in 2011 reported neutrino's that traveled faster than light, or downright fraud, such as for example with the Dutch social psychologist Diederik Stapel. See Baggerly and Coombes (2009) for an shocking account of how wrong things can go.

sample consists of the voters that have been interviewed by the pollsters, whereas the population is the much larger group of all voters. In many cases the ‘population’ is more abstract, such as in a lab experiment, where the sample might consist of a number of measurements the scientist has made, and the population we assume they have been drawn from is then the abstract collection of all possible measurement outcomes.³

The distinction between sample and population allows for an explicit definition of what replication of scientific experiments means. From a statistical perspective replication of an experiment means taking a new sample from the same population. Each sample is similar to the population it is drawn from, but deviates from the population in its own random way. Irreproducible findings then are statements that hold for a particular sample, but not for the underlying population, so that they do not typically occur again in other samples.

The central tenet of statistics is that we are not generally interested in the capricious sample, but only in the stable population behind it. Descriptive statistics describing the sample are therefore of limited use. We use the sample only as a means to learn about the population, a type of reverse engineering that we call *statistical inference*.⁴ To do this in a quantitative way we must make an additional assumption on the manner in which the sample was obtained from the population, typically that it was drawn randomly. This assumption makes the powerful mathematical instrument of probability theory available that describes exactly how much the sample and the population are likely to differ, which in turn allows us to quantify the reliability of inferences about the population.

In particular, we can quantify the probability of drawing a wrong conclusion about the population from a sample. If we assume that a researcher has set out to find a certain relationship or pattern, i.e. to make a scientific discovery, then we can distinguish two possible erroneous conclusions. In the first place, the pattern can be visible in the sample, but not in the population. We call this a *false positive* or a *false discovery*. Secondly, the pattern can be present in the population, but obscured in the sample, called a *false negative*. While both types of errors are harmful, false discoveries are generally considered the more serious of the two. Where a false negative represents a waste of resources because a scientific experiment fails to produce a result, a false positive typically initiates an even greater waste of resources, as it will often be a trigger for misguided follow-up research. In terms of scientific progress, a false negative is a failure to take a step forward, but a false discovery is a step in the wrong direction.

With limited resources it is impossible to prevent both false positive and false negative results completely. A researcher could be very restrained, only publishing a result if there is ample evidence. Such a researcher will incur many false negative results while avoiding false positives. Conversely, an audacious researcher

³In such situations statistics is very explicitly platonic in its philosophy. It supposes that the unobservable abstract population really exists and is of more interest than the observable sample.

⁴As opposed to descriptive statistics, which describe the sample.

publishing results on precarious evidence can expect to have many false positives and few false negatives. Both researchers, however, risk both false positive and false negative results. The only way to avoid false positives completely is never to publish, and the only way to avoid false negatives completely is to always to publish, regardless of the evidence. The inherent randomness of scientific inquiry causes it to have elements of a game of chance. Even the best designed experiment may, by sheer bad luck, produce a sample that is different from the underlying population in crucial aspects and that therefore suggests a wrong conclusion.

False positive and false negative results are an inevitable consequence of the randomness of scientific data. They are not caused by statistical thinking, or inherent to any particular statistical method. Rather, by making the distinction between sample and population explicit statistics provides a language to discuss randomness of empirical data. Avoiding inferential statistics as *Applied Social Psychology* proposed, mostly ignores the problem. Switching to a different statistical framework, such as the Bayesian, merely rephrases it. Wrong conclusions will result from empirical research whatever methods we use, and this fact must be somehow taken into account.

3 The Likelihood of Irreproducible Research

The outcome of the experiment is never fully under the researcher's control, but the probability with which an adverse outcome occurs can be. One way to take randomness into account is to control the probability of an adverse outcome (a false positive or a false negative result). To avoid large differences between researchers regarding the reliability of the evidence they present, in most scientific fields the acceptable risk of a false positive result is pre-specified for all researchers. It is conventionally set to 5 %, which implies that 19 out of 20 times that a researcher performs an experiment the result should not be a false positive, and should therefore be reproducible at least in the limited statistical sense.

This may seem to imply that 19 out of 20 published scientific results are reliable. Ioannidis, however, argued that this is not the case. This ratio of 19 out of 20 represents the perspective of the researcher, but is not immediately relevant from the perspective of the readers of the scientific literature. Even if 95 % of the time researchers produce results that are not false positives, this does not mean that 95 % of all scientific publications are not false positives. This is because negative results, being less newsworthy, are seldom published. Looking only at published results, the proportion of false positives is likely to be much higher than 5 %.

The argument follows from Bayes' rule. It is most conveniently illustrated with a table. Suppose that 200 experiments have been carried out by researchers in a certain field of science in a certain period of time. Sometimes the conjecture the researchers set out to prove was correct, sometimes it was not. For some experiments the researchers accumulated enough evidence to prove the conjecture; for others they were not. Based on these two dichotomies we can summarize these 200 experiments

in a 2×2 contingency table. If we suppose that half of the conjectures that researchers try to prove are in fact true, then we have 100 experiments on true and false conjectures each. If 5 % false positive results are allowed, then 5 out of 100 experiments on false conjectures regardlessly accumulate enough evidence lead to a publication. Conversely, researchers typically accept a 20 % chance of false negative results, so that 80 out of the other 100 experiments lead to a publication. These numbers are summarized in Table 1. As readers of the scientific literature we only see the 85 published results, not the 115 experiments in which the researchers failed to demonstrate their point. The percentage of false positive results among the publications is $5/85 = 6\%$, clearly more than 5 %, but not dramatically so.

This changes if we think of a field in which researchers try much more ambitious conjectures. Let us suppose that instead of 50 %, only 10 % of the conjectures that the researchers attempt are in fact true. In this case we can create a similar table, which will look like the one in Table 2. Now the researchers have to work a lot harder for their publications, and only 25 publications result from their 200 experiments. More importantly, the percentage of irreproducible findings soars to $9/25 = 36\%$.

The percentage of irreproducible results can also be high if many of the experiments on true conjectures are underpowered, i.e. if researchers have a small probability of finding evidence for a conjecture even if it is true. If we would have 50 % true conjectures as in Table 1, but only for 30 out of 100 true conjectures enough evidence would be accumulated, then the proportion of false positive would be as high as $5/35 = 14\%$, as we can see in Table 3. In general, even when the percentage of false positive results per experiment is at most 5 %, the percentage of false positive, i.e. irreproducible results will be large if most of the conjectures researchers set out to prove are false, or if the probability of accumulating enough evidence for publication of a true result is low.

It is interesting to note that in both Tables 2 and 3 we see that the percentage of experiments that leads to a publication is relatively low: 12.5 and 17.5 %, respectively.

Table 1 Illustration of Ioannidis’ argument with 50 % true conjectures

	True conjecture	False conjecture	Total
Evidence for conjecture	80	5	85
No evidence for conjecture	20	95	115
Total	100	100	200

Table 2 Illustration of Ioannidis’ argument with 10 % true conjectures

	True conjecture	False conjecture	Total
Evidence for conjecture	16	9	25
No evidence for conjecture	4	171	175
Total	20	180	200

Table 3 Illustration of Ioannidis' argument: underpowered studies

	True conjecture	False conjecture	Total
Evidence for conjecture	30	5	35
No evidence for conjecture	70	95	165
Total	100	100	200

respectively. One of the things that is crucial for judging the viability of scientific findings is therefore the success rate, i.e. the proportion of failed experiments for every successful one. This success rate is typically hidden from the view of the reader of the scientific literature, who only gets to see the successful experiments. The resulting selection bias, also known as *publication bias*, is inherent to the publication model that is currently dominant in science. Here, the initiative for performing experiments and publishing about them lies with the researchers. The experiment has clearly defined positive and negative outcomes, with positive outcomes being the only ones of real interest. The scientific readership has an exclusively passive role, only taking note of the experiment at a late stage after an apparent positive result has been obtained. Even the reviewers and editors who judge the manuscript are limited to retrospective checking of quality and plausibility. In this model no one except the researchers themselves can see the success rate. No one except the researchers themselves can therefore judge the probability that published results are false positives.

A third way, however, in which the proportion of false positive results in the literature may be high is when there is a large probability that evidence is seemingly found for a conjecture that is wrong. This probability is supposed to be at most 5 %, but it can be much larger because of the well-known psychological mechanism of *confirmation bias*. This is a natural tendency to look for evidence that supports our initial views, and to discard evidence that seems to counter those. Confirmation bias is a very strong force in human thinking, and one which is very difficult to counter. In research, confirmation bias works in rather the same way as publication bias, but at an earlier stage.

Confirmation bias in science may arise for example when there are multiple ways to perform an experiment, a number of statistical models and tests that can be used, or a number of ways to pre-process the data prior to that analysis. Some of these methods are better than others, but which ones those are is often not clear. If an experiment does not give the result that the researcher expected, this may therefore be due to several reasons. Of course the researcher's theory may be false, but it is also likely that something just went wrong in the experiment or that the right analysis method has not been chosen. It is perfectly reasonable, then, and scientifically sensible, to redo the experiment or the analysis. If a second experiment or a reanalysis now turns out to support the scientist's views, a natural explanation will be that there was an error in the first experiment or analysis, which has been corrected by the second.

In practice, researchers therefore do not usually perform one single analysis, but perform several, selecting relatively favorable ones by their confirmation bias. Even if every individual experiment yields a false positive result only once every 20 times, a series of experiments like this may easily have a much larger probability a false positive result, because a researcher trying to demonstrate something that is not true will make several attempts, each of which again has a probability of a seemingly favorable result. When the existence of confirmation bias is taken into account in Ioannidis' argument, it is easy to see that it will result in an even larger proportion of false positive results in the scientific literature.

Ioannidis' simple reasoning can be used to pinpoint areas in science in which we would expect false positive rates to be exceptionally high. These are for example areas with small studies that have low power, areas with exploratory studies where error control is lacking, areas in which statistical methods are not well standardized so that many will be tried out, areas with cheap but difficult experiments in which it is accepted that many experiments fail. However, these are especially those areas in which the scientific conjectures are a long shot, so that most of them are actually false. The resulting findings, paradoxically, are typically the most newsworthy ones which tend to get the attention of the high profile journals. As a rule of thumb, according to Ioannidis' analysis, the more excitement surrounding a scientific result, the greater the probability that it is a false positive. As an extreme case, Ioannidis also describes the existence of *null fields*, areas of research based on false prepositions, in which all researchers are working on research conjectures that are not true. From the reader's perspective, it is difficult to unmask such a field, because the failed experiments remain under the waterline, and a steady trickle of promising results will still be published, especially if many researchers are working in the area. Note that null fields are often sparked by an initial false positive result.

Confirmation and publication bias work together to increase the number of irreproducible results in the scientific literature. The argument we have given here is reminiscent of the 'sloppy science' argument for explaining irreproducible research described in the introduction, but subtly and importantly different. The 'sloppy science' argument implies wilful neglect of proper checks on scientific quality by scientists eager to publish, either because of their own ambition or because they are forced by external pressures. The argument implicitly assumes that if there would be no sloppy science (i.e. if scientists would adhere to statistical rules) there would not be many false positive results. Although it is true, of course, that 'sloppy science', when practiced, would increase confirmation biases and lead to irreproducible results, not all confirmation bias arises from 'sloppy science'. It is also clear from Ioannidis' arguments that large proportions of false positive findings would still arise if 'sloppy science' would cease to exist. Both publication and confirmation bias are inherent to the publication model used to disseminate results in science. We will discuss that model later in more detail, but first look at alternatives.

4 The Dialogue with the Critic

The discussion so far carried an implicit assumption about the way scientists communicate with each other. We take it for granted that they do so via scientific publications, which are well-prepared solitary efforts by a single research group, made public after extensive quality checking by editors and reviewers. This is the current dominant model for science, but it is not the only possible model. To see how other models might function, it is helpful to look back into the history of science. Current science has an amazing productivity in terms of sheer volume of knowledge, but early pre-20th century science has an even more surprising productivity if we take into account the relatively small number of scientists active at the time. In this period, when the foundations of many modern fields were laid out, how did science progress?

Let us illustrate this with an example. In the eighteenth century two Italian scientists were interested in electricity and its relationship to life. It was known that application of static electricity to the limbs of dead animals could cause them to jerk in movements similar to those a living creature would make. Surely, therefore, there was a relationship between electricity and life. This was at least the opinion of Luigi Galvani, a researcher from Bologna. He believed that electricity was an essential life force in animals. According to him, static electricity was sent to the muscles, where it was stored and used as energy for movement. By applying external electricity to the limbs, the researcher released the reservoir of ‘animal electricity’, thus causing the movement that was observed. Not everyone agreed with his views, however. Alessandro Volta, from Pavia, did not agree with Galvani’s views. He did not believe in reservoirs of animal electricity, but held the opinion that it was the externally applied electricity alone that caused the movements.

Galvani and Volta corresponded extensively on this issue, each trying to convince the other. In 1781 Galvani performed what he thought was the definitive experiment. He hung a dead frog on an iron wire on which he had also attached a copper wire. When he touched the frog’s leg with the copper, it jerked in the same way as when he applied static electricity to the frog’s leg. The interpretation, to Galvani, was obvious. No outside static electricity had been applied, and still the frog’s leg had moved. The electricity for the movement must have come from inside the frog. Volta replicated the experiment, getting exactly the same result. However, he remained unconvinced, while Galvani set out his grand theory of animal electricity in a large monograph entitled *De Viribus Electricitatis*.⁵

Volta still maintained that the electricity that caused the frog’s movement must be external, but for a long time he stood alone in his opinion. Only many years later, in 1800, was he able to show that contact between two different metals, such as the copper and iron used by Galvani, may generate a minute electrical current, and that

⁵Although Galvani’s theory turned out to be wrong, this is not irreproducible research in the statistical sense. All experiments the theory was based on were reproducible. Reproducibility is necessary but not sufficient for good theory.

this current was sufficient to cause the jerking of the frog's leg. The electricity was external after all. The exchange between Galvani and Volta has been of crucial importance both for physiology and for physics, as Volta's insights eventually led him to develop the first battery.

It is helpful to look more closely into the dialogue between these two scientists, which represented a type of scientific interaction quite typical for their time. We see a hefty competition that is fueled by irreconcilable theoretical views. Despite, or perhaps because of their differences the two researchers remain in frequent contact. Each tries to challenge his opponent by designing and performing an experiment of which he expects that the result will be in concordance with his own theory while at odds with his opponent's. In this 'duel', it is natural for each of the scientists to immediately try to replicate any crucial experiments in order to try to understand the results and to dismiss them should they turn out to be irreproducible. Volta never believed the results of Galvani's experiment until he had seen them with his own eyes. When he did see them, he still had his own explanation for the result, of course. Reproducibility of an experiment is not enough; in the end it is the implications of the experiment for theory that matter.

A competitive collaboration between scientists with diametrically opposed theoretical ideas can lead to research of high methodological quality, as we can see in the example of Galvani and Volta. For Galvani's experiments, Volta functions as a professional critic, always alert to false assumptions, wrongly designed experiments or hasty conclusions. Galvani could count on Volta immediately replicating every crucial experiment, attacking any weak spots in the design. Irreproducible research would be immediately exposed by him. Moreover, the competition with Volta gave focus to Galvani's experiments. It was not enough if his experiments lent support to his own theory, but they had to simultaneously discredit Volta's. Only experiments for which Volta and Galvani would expect a different result would be relevant to their argument.

The insight that collaboration between scientists with different views can be highly productive motivated the psychologist Willem Hofstee to advocate a 'wager model' for scientific research.⁶ In this model, a scientist who wants to conduct an experiment first tries to find a scientist with different theoretical views and who, on the basis of these views, expects different findings from the experiment than the researcher him or herself. Let us call this scientist the critic. He will play a similar role as Volta in Galvani's experiments. If a critic cannot be found it is not necessary to perform the experiment, since no one would be surprised by the results. Such experiments apparently have no implications for theory. Once a critic is found, the researcher and the critic should sit together to discuss the details of the way the experiment will be performed, making sure that methodological biases do not favor the researcher or the critic. The experiment can proceed when both scientists agree on its validity, and it should possibly be executed in duplicate in both labs to

⁶'Weddenshapsemodel' in Dutch (Hofstee 1980). My translation.

prevent confirmation bias. An experiment set up in this way will have scientific merit whether the outcome is positive or negative for the researcher, and the researchers should commit themselves to publication whatever the outcome. From their competing theoretical views, it is likely that the two researchers will disagree on the final interpretation, with the 'losing' side trying to salvage their theory by alternative explanations.

The name of wager model has been appropriately chosen for two reasons. Firstly, because it suggests a clear investment of both parties into the experiment, with a commitment for each party to 'pay up' and proceed with the publication even in case of an adverse outcome. Secondly, because the word wager invokes the image of betting, suggesting that an element of chance plays a role. In fact, this is usually the case. As we have described above, the competing researchers will have to draw their conclusions on the basis of a sample, while their theoretical dispute is about the underlying population. Since the sample is variable, the risk is that the experiment favors the researcher although the critic's theory is right, or vice versa. This risk the contestants should be prepared to take.

Statistics can help to even the odds for both parties. In fact, the original framework of statistical hypothesis testing as proposed by Neyman and Pearson is highly suitable for the wager model. It uses a 'null hypothesis' representing the critic's view and an 'alternative hypothesis' representing the researchers view, and treats them symmetrically. The famous lemma of Neyman and Pearson tells us how to summarize the data most effectively in order to discriminate between these two hypotheses. The probabilities of a false conclusion favoring either the researcher or the critic can easily be calculated. Using this information a decision boundary can be set in such a way that the wager is a fair one, and the investment can be calculated that is needed to make the probability of both erroneous conclusions acceptably small. The statistician, therefore, has all the tools to stand as a natural arbiter between the researcher and the critic.

Like with the exchange between Galvani and Volta, close attention to methodology is naturally built into the wager model. The crucial element in both cases is the influential presence of a critic. The critic will insist on publication in those cases in which the researcher may not want to publish, thus countering publication bias. The critic will not share the confirmation bias of the researcher because of his competing theoretical views, and will thus be vigilant to counter it. The wager model thus avoids both confirmation and publication bias in a natural way. Since Ioannidis' causes for the large number of false positive results in the literature do not apply, we could expect far fewer irreproducible results if this model would be widely adopted. Sadly, this model is hardly ever used in practice, for various historical, psychological, practical and institutional reasons that we will not explore here.

The value of the wager model here is that provides a very useful ideal that can be used to study the current publication model of science, which we can see as an approximation to the wager model. This perspective will help to understand the methodology better, and also the extent to which this methodology is appropriate.

5 Publishing

Current research practice almost never involves an explicit critic. In contrast to the wager model we can refer to the dominant scientific model as a ‘betting model’. It differs from the wager model mostly by the fact that the critic is abstracted and impersonal.

How does this work? Let us first review an example in which the model works very well.

A group of nutrition researchers from Amsterdam led by Martijn Katan wanted to demonstrate that the consumption of sugar through soft drinks makes children gain weight. This may seem obvious, but other researchers (and soft drink companies) maintained that children would automatically compensate for their sugar intake by being more active or eating less of other foods, negating the weight gain of the sugar intake. To prove their point, Katan’s group enrolled 650 children in several schools and randomly allocated them into two groups. The first group was handed out a daily sugared soft drink. The second group received a daily sugar-free version. The two drinks tasted the same and the children and their parents were kept in the dark as to which child received which drink. After 1.5 year the researchers measured the weight gain of each of the children. They found that on average the children who drank the sugared drink gained one kilo more weight than the children who drank the sugar-free version. They submitted a description of the experiment and their conclusions to the *New England Journal of Medicine*, writing that consumption of sugar via soft drinks does indeed cause substantial weight gain in children. His manuscript was judged and commented on by an editor and two or more anonymous referees, and found acceptable for publication (De Ruyter et al. 2012b).

Before the study was started, the precise design of the study was laid down in a study protocol published separately (De Ruyter et al. 2012a).⁷ This protocol stipulated exactly how the study would be executed, what measurements would be taken at what time, what statistical analyses would be performed and what would be done with the data (or the absence of data) of children who did not follow the study to the end. The protocol also motivates the number of participating children. This was chosen in such a way that if Katan’s theory was right and children would indeed gain weight as a result of drinking soft drinks, Katan would have 80 % chance of demonstrating it with this trial.

If we compare the approach that Katan followed with the wager model of Hofstee, then we can easily see a number of parallels. Katan investigated an issue about which there was clear disagreement in the field. Katan did not explicitly involve a scientist of a different opinion on the matter at stake, but if we imagine that he would have, the design of the experiment would probably have been very similar. He built in many of the methodological checks that would have resulted from negotiation with a critic and which make the experiment impartial to either

⁷This is usual in clinical trials but not in nutrition research.

outcome, such as the blinding of the children and their parents for the type of drink received. In fact, Katan even put himself at a disadvantage when he accepted a 20 % chance of not being able to demonstrate his case even if he was right, against only 5 % for the absent critic. The protocol thus serves as a strong protection of the interest of the critic.

We call the model that Katan uses a betting model, since it is similar to the wager model, except that the researcher plays the game essentially against himself. Katan also played the role of the critic. Other critical scientists, namely Katan's peers, did come into play, but only at the peer review stage after the experiment had been conducted and reported. Like the critic in the wager model, they judged whether they were convinced by Katan's experiment. However, their role was in many respects very different from the role of the critic. They became involved only at a very late stage, and their power to influence the experiment was therefore extremely limited. Moreover, they had the power to influence whether the experiment would be published, a power that the critic in the wager model does not have. Reviewers do not themselves play the role of the critic, they can only judge whether Katan himself played that role convincingly.

The statistical framework that Katan used to analyze the outcome of his experiment, i.e. Fisher's approach to hypothesis testing, clearly reflects the characteristics of the betting model. In contrast to the symmetric framework of Neyman and Pearson that was suitable for the wager model, Fisher's approach is asymmetric. The null hypothesis, which represents the critic's opinion, becomes more formalized, and assumes a greater importance than the alternative hypothesis. Central to Fisher's approach is the concept of a p -value. This value between 0 and 1 is a measure of how extreme the outcome of the experiment would be from the critic's point of view. High values indicate outcomes that conform to the critic's theory. Low values indicate outcomes that are difficult to reconcile with it, but which would more easily fit the researcher's perspective. The p -value can therefore be seen as a quantitative measure that describes to what extent the absent critic is convinced by the outcome of the experiment. Numerically, the p -value is calibrated to take small values below 0.05 only 5 % of the time if the null hypothesis is true, i.e. the critic is right. Conventionally, this five percent is the threshold below which the critic will be convinced. With a p -value below this cut-off, the researcher may claim to have a convincing (in statistical parlance: 'significant') result.

We can see that the absent critic's role and opinions have been completely formalized in this approach. Katan found that children who drink a daily sugared beverage gained about kilo of weight in a year. He also maintained that these results were very difficult ($p = 0.001$) to reconcile with the theoretical view that it does not matter for children's weight whether or not they drink sugar. How convincing this latter statement is crucially depends on how well Katan represented this theoretical view that he did not himself support. We have seen that Katan built in all kinds of safeguards into his experimental design, such as the blinding and the protocol, to protect the experiment from his own biases. Essentially, these measures limit his own freedom in analyzing his results, evening out the odds between him and the critic, and by doing that making the outcome more convincing.

Not all research is as well designed. Headlines in newspapers in 1995 announced that eating tomatoes would dramatically decrease the risk of prostate cancer. Surprisingly, the beneficial effect was not found in fresh tomatoes, but rather in tomato concentrate in the form of ketchup, pizza, tomato soup and even potato crisps with ketchup flavor. The source of the news was a publication by a group led by Edward Giovannucci from Harvard (Giovannucci et al. 1995). According to him, the substance lycopene, found abundantly in tomato concentrate, eliminated the free radicals which caused the cancer. Giovannucci's article has had a major impact, with over a 1000 citations in the scientific literature over the last twenty years. How did Giovannucci come to his conclusion? He asked a large group of health professionals to fill out food intake questionnaires, focusing on intake of 46 vegetables and fruits. Next, he followed his subjects in time to see who would develop prostate cancer, to check whether people who ate more or less of certain foodstuffs would on average develop prostate cancer more frequently. In only 4 of the 46 food types he investigated was he able to find the relationship he was looking for, supported by p -values smaller than 0.05. Upon closer examination, those four were all related to industrially processed tomatoes. A plausible explanation was found in the lycopene theory, and this was the result that was highlighted in the publication.

How convincing is the result? To answer this question it is helpful to imagine how the investigation would have turned out if Giovannucci would have involved a critic. We have to remember that Giovannucci did not yet have his theory about lycopene when he started his study, so that at the moment he contacted a critic, he would have only had a relatively vague theory that the risk of prostate cancer might be influenced by diet. We can therefore suppose that such a critic would be skeptical about this idea, maintaining that the risk of prostate cancer might depend on all manner of things, such as genetic and lifestyle factors, but that food intake did not matter. To settle this difference of opinion, it would be unethical and impractical to use a clinical trial design such as the one that Katan followed, and Giovannucci and his critic would have quickly decided to study observational data. This is a methodological quagmire because it is difficult to distinguish the effects of different factors. For example, people who eat more vegetables typically also exercise more and are more highly educated. If we find that people who eat more vegetables have less prostate cancer, is that due to the vegetables or due to the exercise? Still, discussing these issues at length, it is conceivable that Giovannucci and an open-minded critic might have come to a wager. Would that wager have taken the form described as the evidence in the eventual paper?

Giovannucci investigated 46 different foodstuffs separately, calculating a separate p -value for each of them. In 4 out of these 46 did he find a p -value smaller than 0.05. In terms of the betting model with which we can interpret the meaning of these p -values, this is equivalent to betting against the critic 46 times, of which he lost 42 times and won only 4. If a real critic would be present, it is likely that he or she would claim victory over Giovannucci rather than the other way around. If we remember that p -values are calculated in such a way that the critic will lose the bet about one out of twenty times even when the critic is right, we can expect

Giovannucci to win about 2.3 times out of 46 even when there is no relationship between diet and prostate cancer. Winning at least 4 times in this situation is not an unlikely event, with an occurrence almost 20 %.⁸ Under a wager model, therefore, the conclusion of the study would most likely have been support of the critic's view that diet and prostate cancer are unrelated. If the four foodstuffs for which a relationship is suggested may be the product of chance, it is especially unlikely that a critic would be convinced by the mechanistic explanation about lycopene, made up only after the experiment. The critic may have wondered what explanations Giovannucci might have come up with had four other foodstuffs come out.

The difference between Giovannucci and Katan does not lie in the statistical methods they used. These are broadly the same. The difference is in the way they realized the meaning of the methods they used. Katan took great care to look at his own experiment from the perspective of a critic, taking that point of view into account in every aspect of the study. Giovannucci seems to have done this to a much lesser extent. He applies the rules of the statistical methods he uses, but he does not seem to realize that the results he presents are not as convincing as they have to be. Interestingly, also the reviewers who deemed his work suitable for publication did not notice this.

It is of course the reviewer's job to check a manuscript's quality before advising publication. We could expect that reviewer's take the same perspective as the critic, checking manuscripts meticulously for methodological errors, vigilantly aware of possible confirmation bias on the side of the researcher. In practice, sadly, this is not the rule. Since the reviewers come into play at a late stage, after the experiment has been carried out and reported, many the important problems resulting from confirmation bias remain invisible to them.⁹ For example, they cannot see how many other analysis methods the researcher tried, or what the original hypothesis was that the experiment was designed for. Moreover, reviewers tend to focus much more on the conclusions of the papers than on the methods. This was demonstrated in 1998 by Fiona Godlee, editor of the *British Medical Journal*. She sent an article with 8 deliberate serious methodological errors to more than 200 regular reviewers of her journal. On average, each reviewer only observed 2 of the 8 errors. Of the reviewers, 33 % suggested to accept the article with only minor changes, while only 30 % advised to reject it (Godlee et al. 1998). Reviewers naturally bring their own confirmation bias. When they disagree with the conclusions they will study the methods much more critically than when they agree with them.

The betting model used for scientific publication can best be described as a watered-down version of the wager model. It calls for the scientist to win a bet against a critic of his or her own making, and it is completely up to him or her how

⁸Calculated under the assumption of independence. If—as is likely here—the *p*-values are dependent, this probability will typically be even larger.

⁹This is not the case for the paper of Giovannucci, who (to his credit) makes his confirmation bias very explicit in the description of the experiment and the analysis. The reviewers should have protested and demanded a proper multiple testing correction here.

formidable an opponent the critic is. Some researchers play the role of the critic very convincingly, others just set up a straw man. The statistical methods used are the same in both cases.

6 Speaking About Uncertainty

We now return to the original question about the role of statistics in creating or preventing irreproducible science. We have seen that the randomness of scientific observation makes it impossible to forestall irreproducible results completely, but that two types of bias may dramatically increase the proportion of such findings in the scientific literature: publication bias and confirmation bias. Both of these are tied closely to the current model we use for communication of scientific results via publications, a model I have called the betting model.

At first sight the role of statistics in this betting model seems a rather mechanical one, emphasizing the calculations that have to be done and the cut-offs that have to be exceeded 'to get the statistics right' and to achieve the necessary statistical proof needed for publication. This is often how statistics is taught, as a cookbook full of prescriptions that researchers have to follow in order to analyze their data in a correct way. This mechanical view underrates the role that statistics can play in scientific discourse. In the mechanical view, statistics is seen as an arbiter of truth. This is something it cannot be. Statistics is just a language researchers can use to speak about chance and uncertainty.

To be relevant for scientific progress, experiments must be designed and analyzed in such a way that they make a difference, changing at least some people's opinions about theory. To be convincing requires to be empathic, studying the other side's arguments and taking them seriously. The betting model, as we have seen, only works well if the researcher is prepared to take a critical point of view throughout the design and analysis of his experiment, while maintaining focus on the theoretical issues at stake. A scientific experiment is only valuable if it furthers theoretical discussion in some way.

The scientific attitude necessary for this is under pressure in many countries due to the demands on scientists to publish and acquire grants. In this rat race publications are often viewed as personal achievements of scientists, and as end products rather than as arguments in an ongoing scientific discussion. Regarding a publication as a personal achievement emphasizes competition between scientists for honors, instead of their collaboration on furthering theory. It is based on the misconception that the essence of science is competition between individuals rather than between theories. Regarding publications as end products promotes the idea that the publication should present definite proof. This, in turn, encourages the mechanical perspective on methodology and statistics.

It may be clear that throwing inferential statistics out of the window represents a step back, leaving us with no language to even discuss the problem of irreproducible research. However, having statisticians police scientists is equally pointless

if these checks are only executed at the final stage when the experiment has already been performed. At this stage, much of the confirmation bias is not visible anymore, and should any clear mistakes be found, there is no way to mend them. In the words of the famous statistician Ronald Fisher ‘To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of’ (Fisher 1938). Moreover, involving statisticians in the role of arbiters only serves to emphasize the mechanical view of statistics. This will hamper the discussion between scientists about uncertainty more than it will stimulate it.

Reduction of the proportion of irreproducible research findings calls for a renewed interest in methodology. The mechanistic view of statistical analysis has made many scientists see methodology and statistics as a necessary evil. Better understanding of methodology might help scientists to think about statistics in terms of convincing rather than in terms of proof, and to see how statistical language is a necessary element of the dialogue between researchers with opposing views. The wager model, even if not practical, may help as a thought experiment for researchers setting up an experiment, and may help to create awareness of confirmation biases, and to design more imaginative experiments. To facilitate this thought experiment in the absence of a critic with opposing theoretical views, collaboration with a neutral methodologist may be a good alternative.

Open Access This chapter is distributed under the terms of the Creative Commons Attribution-Noncommercial 2.5 License (<http://creativecommons.org/licenses/by-nc/2.5/>) which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited. The images or other third party material in this chapter are included in the work’s Creative Commons license, unless indicated otherwise in the credit line; if such material is not included in the work’s Creative Commons license and the respective action is not permitted by statutory regulation, users will need to obtain permission from the license holder to duplicate, adapt or reproduce the material.

References

- Baggerly, K. A., & Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, 3, 1309–1334.
- Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391), 531–533.
- Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2011). Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: An argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results*, 1, 1–5.
- De Ruyter, J. C., Olthof, M. R., Kuijper, L. D. J., & Katan, M. B. (2012a). Effect of sugar-sweetened beverages on body weight in children: Design and baseline characteristics of the double-blind, randomized intervention study in kids. *Contemporary Clinical Trials*, 33(1), 247–257.

- De Ruyter, J. C., Olthof, M. R., Seidell, J. C., & Katan, M. B. (2012b). A trial of sugar-free or sugar-sweetened beverages and body weight in children. *New England Journal of Medicine*, 367(15), 1397–1406.
- Fisher, R. A. (1938). Presidential address. *Sankhya: The Indian Journal of Statistics*, 4, 14–17.
- Giovannucci, E., Ascherio, A., Rimm, E. B., Stampfer, M. J., Colditz, G. A., & Willett, W. C. (1995). Intake of carotenoids and retino in relation to risk of prostate cancer. *Journal of the National Cancer Institute*, 87(23), 1767–1776.
- Godlee, F., Gale, C. R., & Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: A randomized controlled trial. *JAMA*, 280(3), 237–240.
- Hofstee, W. (1980). *De empirische discussie: Theorie van het sociaal-wetenschappelijk onderzoek*. Amsterdam: Boom Koninklijke Uitgevers.
- Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Jager, L. R., & Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1), 1–12.
- Klein, R.A., Ratliff, K.A., Vianello, M., Adams Jr., R.B., Bahnkl, S., Bernstein, M.J., et al. (2015). Investigating variation in replicability. *Social Psychology*.
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic Applied Social Psychology*, 37, 1–2.