A Constructivist View of the Statistical Quantification of Evidence

Christian Hennig

Department of Statistical Science, University College London <chrish@stats.ucl.ac.uk>

Abstract

Paper type – perspective

Background(s) – statistical

Perspective – constructivism

Problem– Evidence is quantified by statistical methods such as p-values and Bayesian posterior probabilities in a routine way despite the fact that there is no consensus about the meanings and implications of these approaches. A high level of confusion about these methods can be observed among students, researchers and even professional statisticians. How can a constructivist view of mathematical models and reality help to resolve the confusion?

Method – Considerations about the foundations of statistics and probability are revisited with a constructivist attitude that explores which ways of thinking about the modelled phenomena are implied by different approaches to probability modelling.

Results – The understanding of the implications of probability modelling for the quantification of evidence can be strongly improved by accepting that whether models are "true" or not cannot be checked from the data, and the use of the models should rather be justified and critically discussed in terms of their implications for the thinking and communication of the researchers.

Implications – Some useful questions that researchers can use as guidelines when deciding about which approach and which model to choose are listed in the paper, along with some implications of using frequentist p-values or Bayesian posterior probability, which can help to address the questions. It is the – far too often ignored – responsibility of the researchers to decide what model is chosen and what the evidence suggests rather than letting the results decide themselves in a "objective way."

Constructivist content: A constructivist attitude to formal modelling in science is applied.

Key Words – mathematical modelling, foundations of probability, p-values, frequentism, Bayesian subjectivism, objective Bayes, reality

1. Introduction

This paper is about the principles statisticians apply to quantify the strength of the evidence provided by statistical data in favour or against certain hypotheses. Most of these quantifications belong to the framework of statistical hypothesis tests, such as p-values and Bayesian posterior probabilities.

Here is an example. Note that examples are presented in a very simplified way here in order to not distract the reader too much from the focus of the paper; but see the remarks given in Sections 2 and 3 about to what extent such simplifications are needed on one hand, but to what extent, on the other hand, they suppress potentially relevant details.

Assume that there are two different species of Acacia trees, which I will call species A and B. Biologists were interested in finding out whether colonies of ants have any preference for one of the species. They cleared 28 trees first (15 of species A and 13 of species B), and then they placed 16 ant colonies in approximately equal distance to all trees. Each of these colonies then invaded a tree (apparently it can be assumed that it does not happen that more than one colony invades the same tree). The experiment resulted in the data shown in Table 1.

Acacia species	**invaded**	**Not invaded**	**Total**
A	2	13	15
B	10	3	13
Total	12	16	28

Table 1: Ants data (from Sokal and Rohlf 1981:740)

Obviously, not all ant colonies chose the same species, but many more colonies chose a tree of species B than one of species A. The number 16 of colonies does not look like a large sample size, so one may wonder how clear an indication this experiment gives that species B is generally preferred. A statistician can quantify the strength of evidence, but unfortunately for the biologists, most of whom would like to have a simple number with a clear interpretation, the statisticians have several different methods to do this, which may lead to different results and interpretations, and all methods are riddled with conceptual difficulties.

A constructivist may even wonder whether it makes sense to postulate that there is any (objective) truth regarding the general Acacia preferences of ant colonies, and therefore whether it is a sensible question at all to ask how strong the evidence in the data is about any conceivable truth. The quantification of evidence as a general problem, as well as the problem of assuming a probability model on which a statistical analysis can be based, are instances of the mathematical modelling of reality.

Statistical quantification of evidence is applied in a wide variety of situations. Here is a collection of further more or less typical applications:

- Does better street lightning reduce crime?
- Does Potassium make a breakfast cereal taste better?
- Do the products of a company satisfy an industrial standard?
- Can it be a coincidence that many patients died when a particular nurse was at work? (The Dutch nurse Lucia de Berk was convicted of murder, but the decision

was later revised, both strongly based on statistical arguments, see Derksen 2007.)
- Does homeopathy work against allergies?
- Is a new teaching method/therapy/fertilizer better than the old one?
- Is the spectrum of a certain celestial object compatible with a standard star type?
- How strongly should evidence from DNA analysis be weighed in court?

In Hennig (2009), I have outlined a constructivist perspective of mathematical modelling, based on the idea that mathematical modelling can be seen as a tool to arrive at an agreement about certain aspects of reality, and therefore to construct a stable and consensual reality. In Section 2, I give a brief summary of the ideas in Hennig (2009), including my personal version of constructivism, on which the present paper will be based as well. Even though I do not claim any particular originality for "my" constructivism, which is strongly influenced by authors such as Ernst von Glasersfeld, Heinz von Foerster, Ludwik Fleck, Niklas Luhmann, and Kenneth Gergen, I regard it as a main consequence of constructivism that every constructivist constructs his or her version of constructivism. So it cannot be taken for granted that a concept such as "radical constructivism" has an objective or at least a clearly defined meaning within a given community. (My use of terms like "objective" and "realist" is explained in Section 2.)

Even among those scientists, statisticians and philosophers of science who more or less adhere to realism, the principles of statistical hypothesis testing are highly controversial. Furthermore, the standard approaches of presenting and teaching statistics leave many intelligent and critical beginners confused and frustrated. Apparent paradoxes such as the observation that most professional statisticians on one hand do not believe that, except in the most elementary and exceptional situations, the statistical models are "really true", but on the other hand insist that any statistical method is only valid if its model assumptions "hold", are nowhere discussed in the standard literature in a satisfactory way.

A constructivist view, as opposed to a naive realist one, shifts the focus of attention away from the question of "truth of the models". Instead, the models and quantifications are regarded as items of personal and social construction of perception that may be adopted only temporarily in order to make and communicate systematic observations, without forgetting that there may be more relevant aspects, in personal or social reality, that are ignored in the model but may still enter the discussion of the implications and results of modelling and quantification. It also highlights that and where personal or consensual subjective decisions have to be made about how to perceive and conceptualise reality in order to arrive at helpful quantifications. Quantifications of evidence are produced by such decisions, and can therefore never be fully objective. Accepting this instead of looking for the "best method" and the "correct number" leads, in my opinion, to a much clearer and less problematic view of what statistics can and cannot do, what is required from the scientists in order to arrive at meaningful results and what the price of quantification is. In this sense, I hope that a constructivist view of statistical quantification of evidence also has to offer something helpful to the critical realist who struggles, for good reason, with the confusing way in which the foundations behind the statistical methodology are usually presented.

Based on the general ideas given in Section 2, Section 3 comes back to the example above and introduces the problem of quantifying evidence in some more detail. Sections 4 and 5 are devoted to the two major statistical approaches to quantify evidence, namely p-values connected to the framework of frequentist hypothesis testing, and Bayesian posterior probabilities.[1] Frequentism, subjectivist and objectivist Bayesianism (as explained in Sections 4 and 5) are the major schools in the more than 100 years old controversy about the interpretation of probability and the foundations of statistics. There is still no agreement between these schools. Section 6 compares the schools from a constructivist point of view, focusing on the question which way

[1] There are also further probability-based approaches not treated here such as interval probabilities (Walley 1991) or non-Bayesian versions of the likelihood principle (Edwards 1972).

of perceiving and constructing the world they (and the additional model assumptions that are required for a statistical analysis) entail. This means that the choice between these approaches is not a question of optimality or correctness, but should be guided by decisions of how the scientist wants to think about reality in a given situation, based on personal and social perception of the subject matter prior to the data analysis and the research aims. Section 7 concludes the paper by some general considerations about the role of model assumptions and by listing some questions that may serve to guide researchers when deciding about how to quantify evidence in a given situation.

2. Mathematical models and reality – a summary

Domains of reality

Before turning to mathematical models, I will give a general overview of my personal interpretation of constructivism as applied here. In Hennig (2009) I distinguished different domains of reality, namely "observer-independent reality", "personal reality" and "social reality". *Personal reality* is the reality experienced by an individual. There is a personal reality for every individual. It comprises all sensual perceptions, thoughts and conceptions about the world. The term "constructivism" refers to the idea that personal reality is constructed by the individual, interpreted as a self-organising system. As a constructivist, I do not interpret the personal reality as a reflection or representation of an observer-independent reality outside the individual, but as a result of the self-organising activity of the person (see Foerster 1984 for a condensed overview). Construction is not necessarily meant to be explicit and conscious here; a construct can for example be regarded as made up of behaviour implying tacit assumptions etc.

Social reality is the reality made up (constructed) by all acts of communication. This establishes social reality as something between communicating individuals, separated from the personal reality within each individual. This is similar to, and inspired by, the distinction in Luhmann (1995) between the psychological and the social self-organised systems. Having in mind that there is a heated debated among constructivists about the relation between personal and social construction (see, e.g., Glasersfeld 2008 and the many open peer commentaries of that paper), some more reflections may be in place. Obviously, following the conception above, this idea of social reality (as all the ideas I want to express in this text) is my personal construct. It may be shared or partly shared by other individuals. The text itself, as something that is meant to convey a meaning, is part of social reality. As part of my personal reality, I distinguish between my personal perception of the text (and of acts of communication in general) and the text itself by means of the idea that the text may be outside of myself in a way that it is possible that other individuals may have a different perception of what I perceive to be the same text. This does not claim that the text belongs to any observer-independent reality and "exists" outside any personal reality, but it does claim that the idea that there is something outside myself and that the text and some potential readers belong to it is part of my personal reality. It does also mean that the idea is part of my personal reality that other individuals have potentially different perspectives on what I see as social reality. I write this text hoping that some readers will make some sense of it, and this probably requires some related personal constructs on their side.

Social reality can be seen as a personal construct, but once the idea of social reality is part of a personal reality, the idea of personal reality can be seen as a social construct (a construct in social reality, made up by communication) as well. Communication (more precisely, my perception of it) is the origin of me having language. It inspired me to all the ideas that I outline here, and to the very concepts of personal and social reality, with which I was confronted in perhaps not identical but closely related form, before I made them explicitly my own personal constructs. My whole attempt to make my ideas more precise here is communication. It is intended to contribute to social reality, to have an effect outside my own personal reality. Therefore at least among those individuals for whom something like my social reality is part of their personal reality, it makes sense to refer to and analyse social reality and social constructs in their own right. By regarding personal and social reality as separated domains of reality and

by distinguishing personal and social construction as taking place in these separate domains, I think that it is possible to embed radical constructivism (according to, e.g., Glasersfeld 1995, focusing on personal construction) and social constructivism (according to, e.g., Gergen 2000, focusing on social construction) in a common constructivist framework (some in my view related ideas are for example outlined by Raskin, Krippendorff and Baecker in the peer discussion of Glasersfeld 2008).

Various different social realities can be defined as belonging to different social systems, but the separation of these systems and realities is usually not as clear cut as the separation between personal realities of different individuals. Different social realities are not necessarily disjoint (an exchange of letters discussing scientific ideas may be seen as belonging to a friendship and to science at the same time). Note that these constructivist concepts are themselves constructs that belong to certain social and personal realities. So how and how strongly certain different social systems and social realities are separated from each other depends on the degree individuals perceive and communicate them as separated concepts. Whereas social and personal reality are conceptually separate domains, there exists strong feedback between them. Individuals try to communicate personal constructs and these communications (to be distinguished from the personal constructs themselves) enter social reality. On the other hand, the personal perception of social interaction and communication is a very influential part of personal reality, and many personal constructs can be interpreted as personal adaptations of social constructs (this again refers to the distinction between personal perception of social reality and the idea of social reality itself that is potentially perceived in a different way by different individuals, both of which are part of my personal reality and of the personal realities of those who make such a distinction themselves).

The *observer-independent reality* outside the individual observers (called "world outside" later on) can be said to exist at least as a personal construct of those individuals who construct it, and as a social construct in the social realities in which it is communicated. It is not directly accessible. Regarding the ontological existence of a unique observer-independent reality, constructivism (by which I generally mean my personal interpretation of it) takes an agnostic position.

This means that it is not incompatible with constructivism to believe that an observer-independent reality exists and that my personal constructs and the social constructs I am aware of have something to do with it. In this sense, constructivism is compatible with a quite weak form of realism. Constructivists may want to built up stable and reliable personal and social ideas about the world outside, if they subscribe to this "world outside" construct (which as constructivists they may or may not, and which also may or may not include the belief in the ontological existence of a unique observer-independent reality[2]). I interpret the "scientific method" as aiming at an agreement about social constructs, which can also lead to stable and reliable personal constructs. With this interpretation in mind, constructivists can take part in science as well as realists and objectivists. It distinguishes constructivists from (my interpretation of) objectivists that from a constructivist point of view, science can reasonably be only about personal reality, and personal constructs of social reality and the world outside, whereas for objectivists it is about the observer-independent reality (usually calling something "objective" means "observer-independent"[3]), and for them social agreement is merely a vehicle to achieve knowledge about it. Constructivists do not think that knowing anything objectively is possible in the sense above, and to constructivist scientists, arguments that refer to the uniqueness of (or any other objectivist assumption about) the observer-independent reality should not be acceptable in order to enforce agreement.

---

2 I personally believe that many constructivists can find traces of realism in their own personal realities, but that contradicts constructivism by no means.

3 Weaker definitions of the term "objective" exist, for example related to observability and reproducibility, which may be connected to social and personal realities.

Mathematical reality and mathematical modelling

Mathematical reality is a particular and quite well delimited social reality, made up by formal communication involving mathematical objects. According to the formalist philosophy of mathematics (see, e.g., Hilbert 2004[4]), abstract mathematical objects, constructed formally by axioms and definitions, evolved historically from the use of fingers and notches to count and the use of idealised geometrical shapes to think and communicate about reality (see Hennig 2009 for more details and references). The emergence of abstract and well defined mathematical objects can be attributed to the desire to construct a domain that enables absolute agreement. This is explicitly apparent, for example, in Euclid's axiomatic system. So the idea of absolute truth in mathematics can be explained by a historical process of construction that made binding agreement the essential aim of mathematical communication. In order to make such an agreement possible, mathematical objects needed to be defined in an abstract way, which is devoid of traces of individually different personal perception. As with other social constructs, there is a strong feedback loop between the social mathematical reality and its personal counterparts. There are strong personal adaptations of mathematical reality (strong in the sense that the individuals holding them ascribe high authority to them), and on the other hand the strength of social mathematical agreement is based on individual contributions of intuition, doubt, and convincing arguments. This process was very successful in the sense that almost all people either agree with proved mathematical truths or regard themselves as unqualified. Within mathematics, "truth" can be interpreted as a formal construct in the sense that (according to the formalist philosophy) truth values are established through axioms and transmitted by transparent and formalised logic and proof techniques. Therefore, the mathematical concept of truth is much less problematic, in terms of social agreement, than informal objectivist truth claims concerning the observer-independent reality.

Mathematical modelling consists of assigning mathematical objects to (personally or socially) real entities[5]. Usually it is applied in order to interpret mathematical truths about the mathematical objects as information about the modelled real entities (a standard example is the use of the mathematically derived physical laws in engineering). *Quantification of evidence* is an instance of this; it assigns numbers to the social or personal construct of "the strength of evidence that certain observations carry in favour or against something unobserved that may or may not be true". "Truth" here is informal and therefore potentially controversial, and may refer to constructs of "existing aspects of the world outside unknown to the observer", "uncertain implications of a system of personal or social constructs", or "ideas which to hold will be useful in the future".

But how do we arrive at mathematical objects in the first place? This is the basic problem of mathematical modelling, i.e., the impossibility to formally analyse the assignment of non-abstract personal or social constructs to mathematical objects. Non-abstract constructs are, by virtue of being non-abstract, essentially different from mathematical objects. Furthermore, it is inherent in the process of abstraction that some properties of the constructs to be abstracted have to be cleared. Many realists hope that mathematical models allow insights into observer-independent reality, because mathematical truths seem to be observer-independent. But the strong agreement about mathematical truths can be explained as a result of the construction process of mathematics. Formal truth about mathematical objects is only informative about the modelled non-abstract objects to the extent that individuals and social systems treat the modelled objects in an abstract way. This involves suppressing all aspects of personal and social

---

4 I do not regard formalism as the "correct" philosophy of mathematics, but as the currently still strongest element in the – not necessarily consistent – social construct of the essence of mathematics among mathematicians. Formalism may be legitimately criticised for example for not giving a proper account of the intuitive aspects of mathematics, which according to the present terminology would be located outside the clearly delimited social mathematical reality, but inside many personal mathematical realities.

5 Note that this use of the term "model" is different from the one in mathematical model theory, see, e.g, Manzano (1999). The latter one is fully formal and therefore belongs, according to the terminology here, fully to mathematical reality.

reality that are lost in abstraction. In other words, mathematical modelling is a way of thinking about, and operating on, reality, which here may mean quite general aspects of personal and social reality including the world outside, but it is not a way of getting closer to the observer-independent reality. Quite often individuals and social systems attach more authority to the results of mathematical modelling than these results deserve. The idea that science aims at agreement and stability makes the use of mathematical modelling as part of the scientific method seem natural. However, general science is not restricted to what can be formalised, and therefore it cannot rely on mathematical truth but has to deal with the basic problem of modelling as well, which requires negotiation about and exchange of differences between personal and social realities.

Regarding the history of mathematical modelling, the following pattern has been observed in Hennig (2009): Abstract mathematics evolved from re-organising some practical operations. First, mathematics and the practice from which it arose were not considered to be separated. The Greeks started to consider mathematical objects as a different, more ideal domain of reality. Only much later, starting with modern science, already existing abstract mathematics was deliberately used to model objects and operations that historically had not been connected to mathematics. Mathematical modelling may therefore be considered as thinking about modelled non-abstract constructs in terms of those operations that gave rise to the mathematical structures. In this sense, it is metaphorical thinking.

Most discussions of mathematical modelling in scientific contexts focus on the question whether the models "really hold". From a constructivist point of view, this can never be decided independent of the observing system. Hence a different set of questions becomes relevant.

- What is the pragmatic aim of the model, and is it constructed in order to achieve this aim? In Hennig (2009), I list several aims and potential benefits of mathematical modelling apart from "approximating objective reality" such as improving mutual understanding, stimulating creativity, and decision support.

- Which aspects of the respective realities of involved individuals and social systems are not captured by the model? What implications does this have? Asking this question does not mean that all aspects should ever be included because this is generally impossible due to the nature of abstraction and the limited complexity that can be handled mathematically. My impression is that a major problem with the objectivist way in which mathematical models are usually discussed in scientific practice is that differences between model and modelled reality tend to be swept under the carpet because the existence of such differences is usually regarded as a weakness of the model.

- What are the implications of thinking about and communicating the concerned reality in the way implied by the mathematical model? The feedback of mathematical modelling to the personal and social realities of those involved should be acknowledged. Is this desirable in the given situation?

3. Quantification of evidence

Quantification of evidence is an instance of mathematical modelling. It connects direct perceptions to (more or less) general statements or theories such as "ant colonies do not prefer any particular Acacia species". Quantitative or categorical data are the perceptions with which statistics deals. Measurements are general operations of transformation of socially real items into mathematical objects. They produce data. For example, a count is a (rather basic) measurement.

The term "direct perception" requires some clarification. I treat it as a widely held social

construct that it is possible to distinguish between what is directly, "materially" observable and what is not directly observable, but may still be real (in a certain domain of reality, see above). This construct is based on the (usually) strong agreement about what is directly observable in a given situation. It depends, however, on observer-dependent constructs such as the delimitation of Acacia species. The data in Table 1 were directly observable at the time of the experiment, assuming a clear definition of the categories. In some instances, not only the definition of the measured values is observer-dependent in a non-trivial way, but also how they are related to the issue of interest. For example, there are various personal and social constructs of intelligence around, and depending on these constructs the IQ may be seen as a measurement of "general intelligence", certain aspects of intelligence, or inappropriate for any reasonable measurement of intelligence (but possibly still appropriate for measuring another potentially interesting property of the test person such as fitness for certain jobs).

When quantifying evidence, the extent to which data support a statement like "ant colonies do not prefer any particular Acacia species", which does not refer to direct perceptions, is expressed by a hopefully easily interpretable and comparable number.[6] This obviously assumes that it makes sense to believe, or at least to act as if, ant colonies either do or do not prefer a particular Acacia species, so it assumes a construct of a "not directly observable reality, about which observations are informative though not necessarily conclusive". I do not see problems with this from a constructivist point of view, but it is probably helpful to acknowledge it, because constructivism does not take the objective existence of such a reality for granted – personally and socially, it exists only if it is constructed.

In many cases, the pragmatic aim of the quantification of evidence is decision support, as illustrated in the following example. Table 2 shows the results of a study on coffee consumption and coronary heart disease in men aged 40-55 employed by the Western Electric Co., Chicago, after eight years of follow-up. One obviously observer-dependent element of these counts is the class definition of "heavy coffee drinking" with a cutoff value of 100 cups/month. The reader may wonder whether more precise information, namely the number of cups/month for every single observed person, should be used, but it is conceivable that the actual measurement procedure for this would seem to be much less reliable than the given assignment to just one of two classes.

	Coffee> = 100cups/month	**Coffee<100 cups/month**	**Total**
CHD	38	39	77
Non-CHD	752	889	1641
Total	790	928	1718

Table 2: coffee consume and coronary heart disease. Data taken from Greenland & Mickey (1988:338), original study by Paul (1968).

The question of interest is whether, speaking in constructivist terms, it is sensible to construct coffee consumption as a cause for coronary heart disease.[7] The results of this study may for example be used for deciding whether people should be advised to limit their coffee consume or whether it makes sense to try to produce less harmful coffee. But it is not only relevant whether people should limit their coffee consumption or not, but also to have an idea of how conclusive

---

6 Note that in some cases evidence is evaluated concerning statements that refer to events that are constructed as "directly observable in principle, but not actually observed by those who evaluate the evidence" such as "Mrs B is a murderer".

7 Note that much more comprehensive evidence concerning this question is available in the literature, which I omit here to keep things simple; furthermore I disregard the question of "effect size" that would be relevant in practice.

the evidence actually is. Should more observations be made before any recommendation is given? If further observations are made, what weight should the present study have compared to others? In order to address these practical questions transparently and to support general agreement, a number quantifying the *strength of evidence* would be helpful.

The ant preferences data do not seem to address an immediate practical decision problem. A number quantifying the strength of evidence could be used for communicating the "message" of the table regarding the question of interest more efficiently, contributing to a larger body of scientifically agreed knowledge in the field that at some point can be used for decision support or other aims. There is a clear difference between "constructivistically valid" aims, concerning decisions, behaviour, and personal and social construction processes and the aim to "find out whether the statement is really true".

"Strength of evidence" is an abstract construct. It is not directly connected to operations that initiated mathematical objects, and therefore it cannot be expected to be quantifiable in a straightforward way. A possible starting point is to analyse how people assess the strength of evidence informally. If considerations can be restricted to 2*2 contingency tables, it will be possible to start with some – for reasons of simplicity somewhat imprecise – axioms such as "assuming that the marginal totals remain constant, evidence against independence of rows and columns is the stronger, the more the row-wise relative frequencies for the columns deviate from the marginal relative frequencies for the columns" or "assuming that the row-wise relative frequencies for columns deviate from the marginal relative frequencies for the columns, and all of these are constant, evidence against independence is the stronger, the larger the overall number of observations is." Using such axioms is an attempt to formalise the way how "rational people" think about a construct. In some cases they may lead to a exhaustively specified mathematical model. (As will be illustrated later, a scientific discussion about what "rational" denotes in this respect is necessary as it is by no means straightforward.)

Statistical evidence in general is concerned with all kinds of different types of data (for simplicity, the examples in the present paper only concern 2x2 tables), and therefore a direct axiomatic approach would be quite cumbersome. However, the problem of quantifying the tendency of certain events to happen under uncertainty is closely related, and this is modelled by the probability calculus, which itself is based on axiomatic considerations about relative frequencies and has been applied to very general types of events and data. Therefore, most approaches of the quantification of evidence use probabilities. However, there is no scientific agreement about the interpretation of the probability calculus, and it can be used in different ways to apparently rationally formalise the strength of evidence (as discussed in Sections 4–6).

Most of the remainder of the paper deals with the evaluation of evidence from data based on probability models[8]. In this regard, it is important to keep in mind that by virtue of being mathematical objects, the data as well as the probability models differ from the personal and social constructs that are really of interest. Knowledge about how precisely the ant preference experiment was carried out, for example whether and how interaction between ant colonies was prevented, is relevant to decide whether the data could be seen as a reliable source of evidence. For the coronary heart disease data, one problem is that it cannot be taken for granted that statistical dependence between rows and columns can be interpreted as indicating causality. It is for example conceivable that an unobserved confounding factor causes the desire to drink a lot of coffee and susceptibility to coronary heart disease. There are scientific principles to back up a scientist's decision to regard some problems as irrelevant, such as controlled randomised trials for confounding factors. However, differences between model and personal and social realities can never be completely removed (data from a controlled randomised trial are obviously observed under circumstances that deviate from the uncontrolled realities to which the results are to be applied) and are ignored by the statistical method. This is an important factor when

8 Generally, probability models are defined as [0,1]-valued functions on certain systems of sets, interpreted as "events", that obey Kolmogorov's (1933) axioms such as additivity.

interpreting statistical results.

An essential aspect of the statistical approaches discussed here is that they attempt to quantify evidence in a unified way regardless of the subject matter. Computations are only based on the data, cleaned of their meaning, and the probability models under examination. This kind of unification may be seen as an aim in itself, and it may also support communication between disciplines. However, it is not clear, and a matter of case-wise negotiation, whether it is appropriate to treat very different subject matters in the same way. For example, the application of probability models may be assessed differently depending on whether situations are treated that can be interpreted as repetitions of more or less identical conditions such as routine tests of the quality of products produced under very similar conditions, or singular situations such as somebody being suspected of murder. An implication of the unification is that it tempts researchers to separate the statistical work from the subject matter expertise. Very often, when statisticians collaborate with subject matter researchers (or when the researchers use statistical software), subject matter expertise is used for deciding whether the data properly reflect the issue of interest, setting up the model, and deciding whether the model assumptions are regarded as sufficiently fulfilled (see Section 7 about a constructivist view of the role of model assumptions), but apart from this, statistical calculations are carried out in an abstract way without making reference to the meaning of the data, and the researcher does not worry about not understanding them. Statistics is often expected to come up with some kind of "objective result" for which the researcher does not have to assume responsibility. But the actual way data are statistically processed implies certain ways of thinking about the subject matter and therefore reveals certain ways of constructing it. If computations are treated as separated from meaning, it remains opaque what they imply and how they transform meaning. Therefore it seems to be desirable that the statistician and the subject matter researcher attempt to have a joint understanding by using knowledge of both areas.[9] The constructivist way of discussing the quantification of evidence may even be useful for realists because it explicitly emphasizes where observer-dependent decisions have to be made. Its focus on the question of agreement between observers makes it more transparent where and why stronger or weaker agreement can be expected (related to what is constructed as "directly observable"), what has to be negotiated, and to what extent disagreement cannot be expected to disappear. I think that one does not need to be a constructivist to see the benefit of this. Furthermore I also think that the realist focus on objectivity encourages researchers to ignore the problems of observer-dependence and differences between personal and social realities, and that this ignorance is responsible for much of the confusion about statistics.

4. Frequentist p-values

How p-values work

In order to quantify the evidence against an independence hypothesis in a 2x2 table – such as the hypothesis in Table 1 that ants do not prefer any of the two Acacia species –statisticians use a standard probability-based method, namely a p-value from Fisher's exact test of independence (Fisher 1935).

The general idea behind p-values and statistical significance tests is to address the question "could the given data have occurred by chance?" Depending on the problem at hand, "by chance" may have different meanings. When testing independence in a 2×2 table, "by chance" refers to a situation in which the row and column variable (Acacia species and ant invasion) are independent. In other situations it may mean: "application of homeopathy does, on average, not change an allergy indicator" or "all nurses have the same probability to see patients dying".

---

9 Since, however, this paper focuses on statistical aspects, much potentially relevant subject matter knowledge about the example data will have to be ignored here.

More complex constructs are also possible: “all nurses’ probabilities to see patients dying depend only, and in the same way, on how their work shifts are organised.”

It is then required to set up a probability model for “by chance”, so that the probability can be evaluated that something that is as “far away” as the given data from what would be expected under the model could have occurred in case that this model holds. Under the not unproblematic but standard additional assumption that what the ant colonies do is independent of each other, it is straightforward to set up such a model for the situation in Table 1. Table 3 shows the expected frequencies under such a model, given the marginal totals.[10]

Acacia species	**invaded**	**Not invaded**	**Total**
A	6.4	8.6	15
B	5.6	7.4	13
Total	12	16	28

Table 3: Expected frequencies given the marginals under the independence model

Given the fixed marginal totals, Table 1 can be entirely reconstructed from the value of a single cell. Therefore it suffices to ask whether the fact that only two trees of species A were invaded by ants is compatible with the model that expects, on average, 6.4 in this cell. This leads to the so-called “hypergeometrical distribution,” which was originally developed for urn problems. The situation is equivalent to computing the distribution of the number of black balls when drawing 12 balls (invaded trees) from an urn with 15 black (species A) and 13 white (species B) balls. Therefore, it can be said that using this model amounts to thinking about the ant colonies as if they were balls from such an urn.

The probability that two or fewer black balls are drawn in this situation is 0.001. This is the “one-sided p-value”. It is obviously a very small value and can be interpreted by saying that the observed data are quite incompatible with the model, or, in terms of quantification of evidence, that the data provide strong evidence against the hypothesis that the ants do not prefer any species, and not species B in particular.

The corresponding computation for the data in Table 2 yields a probability of 0.31 for having 38 or more CHD cases among 790 heavy coffee drinkers if these are drawn out of a population of 77 CHD and 1641 non-CHD under independence. $p = 0.31$ is not very small and it is therefore well conceivable that such a distribution is observed if the model holds. In other words, there is no evidence against independence.

Two things are worth being noted. Firstly, the method depends on probabilities for events that

---

10 Fisher’s approach treats the marginal sums as fixed, as opposed to the famous approach of Neyman & Pearson (1933), which in the given situation would make things much more complicated, and will therefore be omitted in the present paper. In general, however, it is based on the same underlying principle, only using an additional optimality criterion.

did *not* actually happen. Not only the probability for 2 invaded A trees is computed, but the sum of the probabilities for 0, 1 and 2 invaded A trees. In the coffee example, the probability for the precise result 38 is 0.07, which is much smaller than the p-value above. But this probability is not useful under the logic discussed here because if there are many possible outcomes the probability for any precise outcome will be small. This in itself, however, cannot be reasonably interpreted as evidence against the model. Therefore, the p-value is the probability that under the model something happens that is as far *or farther* away from what is expected than what was actually observed. Some statisticians find it counter-intuitive that under this conception for example the result of 2 in the ants example would be defined to provide weaker evidence against a model under which 2 had the same probability as before, but the unobserved values 0 and 1 had higher probabilities. A controversial discussion about this is going on among statisticians and philosophers of statistics, and as far as I know, almost all protagonists hold that this intuition is either "correct" or "wrong". From a constructivist point of view, it can be observed that the inclusion of probabilities of unobserved events entails a certain way of looking at the situation, in which rather "the observation is much smaller than expected" than "the observation is 2" counts. It is certainly legitimate to point out that this view may have odd consequences. Unfortunately, the opposite view based on the so-called "likelihood principle", which holds for Bayesian statistics as treated in Section 5, may have similarly odd consequences in other situations (see, for example, Mayo & Kruse 2001, Davies 2008). There is no objective way to decide the issue. It seems much more helpful to accept that different intuitions exist and to explore what they imply in order to negotiate case-wise decisions.

Secondly, a decision is needed whether the p-value should be evaluated in a one-sided or a two-sided way. Above, one-sided probabilities were computed; it was only taken into account whether the number of colonies A trees under the model could have been 2 or smaller. But the outcomes 11 and 12 are farther away than 2 from the expected value of 6.4 as well; only they are not smaller, but larger. Adding these probabilities yields a two-sided $p = 0.0018$, which is still very small, but situations are conceivable where different conclusions would be drawn from one- and two-sided p-values. The decision whether a one- or a two-sided p-value should be used depends on the focus of interest. Is the research question rather whether the ants prefer species B (*one-sided question*). If the data indicate that rather species A is preferred, this would not count as evidence against the model because independence and preference for A are identified in terms of interpretation), or whether they prefer any of the two species (*two-sided question*)? Generally, significance tests and p-values do not only depend on the null ("chance") hypothesis, but also on an alternative hypothesis of interest – even though in practice it may happen that data may be neither compatible with the null hypothesis nor with the alternative.

To summarise, the quantification of evidence based on p-values has four requirements:

- A "null model" formalising "chance" or "no effect",
- An alternative model formalising the direction of deviation of interest from the null model,
- A statistic to measure how far away the observed data are from what is expected under the model in the direction (or directions) of the alternative, and
- The mathematical derivation of the distribution of this statistic under the null model.

The p-value is the probability under this distribution that the statistic is as far or farther away from the expected value under the null model than is observed value. The smaller the p-value, the stronger the evidence against the null model. Large p-values do not provide evidence against

the null model. p-values are “large” if they do not make the observed value of the statistics seem very unlikely; normally any value above 0.1 is interpreted to be “large”.

It is crucial that a large p-value by no means indicates that the null model is true. This does not have anything particular to do with constructivist philosophy. Even assuming that there is a true model, it has to be accepted that data that are perfectly compatible with the null model are also compatible with many other models. p = 0.31 in the coffee example excludes by no means the possibility that strong coffee drinking increases the risk for CHD a little bit, weakly enough that this cannot clearly be seen from the data at hand. Furthermore, p-values may be affected by violations of the model assumptions that do not have to do with the intended interpretation – for example, the dependence between ant colonies or a common unobserved factor behind both coffee consume and CHD. This means that self-critical thinking about conceivable effects of real aspects ignored in the model and a careful experimental design are required.

The concept of p-values is based on an interpretation of probabilities as something that governs the observed phenomena. Probabilities are constructed as modelling an aspect of the “world outside” the observer (which may mean an objective world outside to a realist, or a personally and socially constructed one to a constructivist). The most prominent interpretation of probabilities referring to the world outside is frequentism, and p-values are usually interpreted in a frequentist way. However, as illustrated in Section 5, not all interpretations of probability refer to the “world outside.”

The frequentist interpretation of probability

As all interpretations of probability, frequentism is a way to connect the probability calculus to reality.[11] The basic idea of all frequentist approaches to interpret the probability P(A) of a set A is as follows. Imagine that there is an experiment that can be repeated in an identical and independent way. The outcomes of the experiment are measurements. Let A be a subset of the set of possible outcomes. Imagine that the experiment is carried out n times, n converging to infinity. Imagine further that the frequency of outcomes in A, divided by n, converges to a limit. This limit is interpreted to be P(A).

Obviously this idea is an idealisation. It requires ignoring conceivable sources for dependence and non-identity – actually whenever two executions of the experiment can be distinguished, strictly speaking they cannot be identical. Furthermore, infinitely many repetitions cannot be observed and therefore probabilities cannot be observed. This implies particularly that, even under an objectivist idea of material existence, it is not observable whether probabilities exist. Relative frequencies for finite repetitions of experiments perceived to be sufficiently identical and independent are observable and can be interpreted as “approximations” of probabilities. From a mathematical point of view, however, this is not valid, because a limit point of a mathematical sequence is invariant against arbitrary alterations of any finitely long subsequence. Frequentism has often been criticised for these problems (Finetti 1970, Howson & Urbach 2006). Its defence, usually carried by realists, led to several variants of frequentism, but most arguments about the (approximately observable) existence of frequentist probabilities involve the law of large numbers and are open to charges of circularity. The law of large numbers is a fundamental theorem of the probability calculus and states that, assuming independence and identity of repetition of an experiment, the relative frequency of the observation of A converges in a probabilistic sense to P(A).[12] It even gives bounds for the

---

11 See Section 6 for a historical note on the idea of separating the probability calculus from its interpretation.

12 Actually there is more than a single such law in probability theory but I avoid here the subtleties

difference between P(A) and a relative frequency for fixed n that can only be exceeded with very small probability, so that it can make the connection between probabilities and relative frequencies for finite repetitions precise in some sense. However, the law is itself formulated in terms of probabilitie. “Independence” and “identity” enter in their probability theoretical formal meanings, and thus are not identical to their intuitive meanings, but rather mathematical models of them.[13] As a consequence, its interpretation needs to assume a probability interpretation already.

I maintain that, when attempting to make the connection between mathematical models of relative frequencies of experimental outcomes under uncertainty and reality precise, circularities cannot be avoided, because if the outcome of an experiment is uncertain, it is uncertain as well how close the corresponding relative frequencies under repetition will match any conceivable value of a probability. Whatever is observed cannot prove or disprove any limiting value for relative frequencies, but the implications of the law of large numbers for finite n can be tested – using the probabilistic methods for quantification of evidence.

From a constructivist point of view, the objections against frequentism are not severe, because constructivists are not concerned with the most critical issue, i.e., to establish the observer-independent existence of probabilities. For a constructivist, adopting a frequentist interpretation of probability means to treat a situation – at least temporarily – as if it were a realisation of an experiment that can be repeated infinitely many times in identically independent ways obeying the rules of probability theory. Following von Foerster’s (1984) ideas about stable constructs as eigenvalues of self-referential behaviour, constructivists can accept that they have to live with the kind of circularities encountered above if they want to establish concepts like quantitative values for uncertainty and evidence. Frequentism becomes a particular way to perceive and analyse situations in which uncertainty arises. It can be temporarily adopted but is not right or wrong or good or bad in any objective sense. What needs to be decided is whether it serves the aim of the data analysis in the given situation properly or not. Assuming that the model holds, the constructive power of frequentist models is that it can be mathematically described what pattern of outcome can be expected, and this can always be compared with some observed reality. In order to learn from such models, it is not necessary to assume it to be the “true” one.

Frequentist interpretation of p-values

Interpreting p-values in a frequentist way means that the whole experiment (observing all 28 ant colonies, or all the individuals in the coffee dataset) is constructed as repeatable. In the ant example, the p-value then gives the expected relative frequency, under infinite identical and independent repetition, of observing 2 or fewer invaded A species trees under the null model. It is up to decide for the researcher (and her audience) whether this is a reasonable construct. Such an experiment can certainly be repeated, though it depends on the precise conditions whether it is convincing to model these repetitions as independent and identical. The model-implicit treatment of the ant colonies as independent of each other seems more critically to me, but such things can more convincingly be assessed by the subject matter experts.

A very frequent misinterpretation of p-values is that they give the probability of the null hypothesis to be true (“the probability of the occurrence of CHD to be unaffected by heavy coffee drinking is 0.31”). It has been argued particularly by Bayesians, see Section 5, that the researchers really should be interested in this latter probability, because this would be a direct measurement of whether the model should be believed or not, given the available evidence. The p-value is only a rather indirect indication of the strength of evidence, because the information how likely the observed outcome is under the model does not tell the researcher directly how valid the model is. Under the frequentist interpretation, however, a probability for the model to hold does not make sense except under the rather curious construct of a repeatable development

---

of discussing their differences.

13 In the probably most well known reference for the foundations of frequentism, Mises (1928), avoided the terms “independence” and “identity” as basic concepts for his version of frequentism in order to avoid circularity. However, his own suggestion attempted to formalise the same intuition, was riddled with difficulties as well and did not gain general acceptance.

of the observable world so that the null model is true in a constant limiting relative frequency of cases. The frequentist idea is that the model either holds or not, unknown to the researchers, and that probabilities describe what the model does, but not how the researchers should think about it. As said before, even an arbitrarily high p-value cannot exclude the possibility that many other models are compatible with the observed data as well. This also implies that p-values are not unique as a frequentist way to measure strength of evidence, though they are by far the most popular one.

5. The Bayesian approach

Bayesian interpretations of probability

Bayesian posterior probabilities are the most widespread statistical alternative to p-values. Their adherents claim that they have two major advantages over p-values. Firstly, they deliver a value $P(H_0)$, H_0 being the null hypothesis, which apparently allows direct interpretation as the "probability that the null hypothesis is true", as opposed to p-values. However, to some extent this is a misinterpretation as well, see below. Secondly, their computation does not involve probabilities of unobserved events.

Before carrying out Bayesian computations, it makes sense to discuss how Bayesian probabilities are interpreted. Bayesian statistics is named after Thomas Bayes's (1763) Theorem. Omitting some formal details and assuming that H_0 and H_1 together cover all possibilities, Bayes's Theorem roughly states that

$$P(H_0 \mid data) = \frac{P(data \mid H_0)P(H_0 \mid p.i.)}{P(data \mid H_0)P(H_0 \mid p.i.) + P(data \mid H_1)P(H_1 \mid p.i.)}$$

"p.i." stands for "prior information", "|" stands for "conditionally on". Note that all probabilities are interpreted in Bayesian statistics as conditional probabilities "P(A|some information)" but some conditions are usually omitted by "lazy notation". $P(H_0|data)$ would be more precisely denoted by $P(H_0|data \ \& \ p.i.)$ and P(A) as I will use it below implies conditioning on the state of information in the given situation, whatever it is. In the context of objective Bayes, the case of "no prior information available" is treated later, which in the formula above can be interpreted as a special case of p.i.

Bayes's work was only published after his death, and it is quite brief about the interpretation of probability. Therefore there is no agreement about Bayes's own interpretation, and his name is nowadays used for different interpretations. In this paper, I concentrate on the two most popular ones, often branded "subjectivism" and "objective Bayes". Both of them have in common that, as opposed to frequentism, probabilities do not model a world outside, but a "rational strength of belief" of an individual in the occurrence of a certain event. In the objective Bayes approach the individual is idealised to be unbiased by any prejudice and to have access to all available information, and the resulting probability value should be unique. In subjectivism the probability value is allowed to depend on the individual.

Not knowing whether A will occur (or has occurred) or not, the probability value P(A) can be

interpreted as the fair "betting" rate in a gamble where the individual gets 1 unit back if A occurs but nothing if A does not occur. Assuming that the individual can be forced to bet either on or against A, operationally this means that the individual will bet on A if a rate below P(A) is offered to her, and against A with one minus the offered rate otherwise. In this way, at least the subjectivist Bayesian interpretation can be linked to the individual's behaviour.

The problem with the expression $P(H_0)$ is that such an approach does not allow regarding H_0 as a frequentist probability model that could be true or false in the world outside the individual. $P(H_0)$ therefore cannot exactly be the "probability that H_0 is true" under Bayesian interpretations either (but is often misinterpreted in this way). The most important proponent of an operationally subjectivist interpretation was Bruno de Finetti (1970), who stressed that the expression P(A) only makes sense for events A for which it is possible to decide later, by future observations, whether A has occurred or not. This does not apply for probability hypotheses in the sense discussed above. Finetti's interpretation of $P(H_0)$ is indirect. When assigning probabilities to events A of which the occurrence can be observed in the future, they can be computed by $P(A) = P(A|H_0)\,P(H_0) + P(A|H_1)\,P(H_1)$, so that $P(H_0)$ and $P(H_1)$ become technical devices to compute P(A). A more careful direct interpretation of $P(H_0)$ is that with probability $P(H_0)$ it makes sense to compute Bayesian probabilities for events observable in the future as if they were generated by a frequentist model H_0. This is still not fully operationally understandable, because it still assigns a probability to something unobservable, but many Bayesians do not find operational definitions as important as Finetti.

For most models, the use of $P(H_0)$ to specify probabilities of observable future events involves the concept of "exchangeability", which is the Bayesian formulation of independent (conditionally under a probability model which itself is uncertain) and identical (in terms of probabilities assigned to future events) repetition. Whereas the Bayesians do not assume independence and identity to hold in the reality outside the observer, the Bayesian application of the probability calculus requires the individual to assign probabilities that follow similar assumptions to make learning from experience (inference from past data to future data) possible.

Instead of using the probability calculus for obtaining probabilities from a model that is assumed to be located in the outside world and may be "true" or not, in Bayesian statistics the calculus governs how prior beliefs should be modified in a supposedly rational way in the light of the data. Bayes's Theorem requires the knowledge $P(\text{data}|H_0)$ and $P(\text{data}|H_1)$, which are obtained from the calculus as in the frequentist approach given H_0 and H_1, but it also requires the prior probability $P(H_0|\text{p.i.})$. The *prior probability distribution* is the key ingredient that makes the computation of $P(H_0|\text{data})$ possible in the Bayesian approach, and the dependence of Bayesian inference on the prior distribution is a standard frequentist criticism. It is also the major difference between subjectivists and objective Bayesians. According to the subjectivists, the prior distribution reflects the prior state of belief of the subjective individual, and they allow in principle any distribution as a prior distribution, although there are some suggestions in the literature about which principles to apply when designing it in a given practical situation. According to the objective Bayesians, however, the prior distribution should be unique. In case that there is no prior information, it should be a distribution modelling the absence of any information, and in case of existing prior information, it can itself be a distribution resulting from Bayes's Theorem, starting at some point from absence of information and updating the information by data that had been collected and evaluated before the study for which then a prior distribution is required. However, there is no agreement about how an "objective" non-informative prior distribution should look like, which will be illustrated in the following section; see Kass and Wasserman (1996) for an overview of problems with selecting non-informative priors.

Apart from the selection of prior probabilities, another issue with the Bayesian approach is whether Bayes's Theorem and the probability calculus really provide "rational" updates of probabilities in the light of the data. This is again a question of the connection between mathematical modelling and reality. The Bayesian approach models rationality in particular

way. This is based on the idea of "coherence". Coherent betting in a Bayesian sense means that betting rates (and therefore probabilities) have to be chosen by the betting individual in a way that no opponent can apply a betting system based on the individual's betting rates so that the individual loses money regardless of the outcomes of the statistical experiments. It can be shown mathematically that this demand (properly modelled) entails the axioms of probability theory for betting rates. Here is an illustration of this. Assume that an individual D specifies probabilities for the outcome of rolling a single (not necessarily fair) die. Assume that D specifies P({1})=P({2})=0.2 (there is nothing to stop a subjectivist Bayesian from doing this, and objective Bayesians may do so in certain situations if indicated by prior information). Assume further that D violates the axiom of additivity (for disjoint events) by setting the probability for rolling a 1 *or* a 2, namely P({1,2})=0.3 (instead of 0.4). Assume now that a betting opponent E offers a rate of 0.18<0.2 to I for betting on 1 and 2 separately, and 0.32>0.3 for betting on {1,2}. According to the operational definition of Bayesian probabilities, D will pay twice 0.18 to bet on each 1 and 2, and 1-0.32=0.68 to bet against {(1,2)}. This means that D pays E 1.04 overall, but will only win 1, whatever the outcome of the roll is (either 1 or 2 or any other number, i.e., "non-{1,2}"). So D loses 0.04 in each case. It can be shown that such a situation is not possible if D obeys the probability axioms.

A crucial assumption of this result is that the individual can always be forced to bet either in favour or against an outcome according to her specified betting rates. Several aspects could be controversial (see, e.g., Dawid 1982, Walley 1991):

- Does it make sense to think about any given situation in which evidence should be quantified in terms of bets and betting rates? Even if this is accepted for the situation of interest, it is still not clear that all available prior information can be properly formalised in terms of betting rates /probabilities.

- Should the assumption be accepted that the individual is forced to bet? There is an alternative concept of "imprecise probabilities" in which the individual is allowed to leave some room between the highest rate with which to bet in favour of A and one minus the highest rate with which to bet against A, leading to probabilities that are intervals rather than single numbers (see Walley 1991), in which case the individual would not be assumed to be forced to bet if the offered rate is in the interval.

- Does Bayesian coherence model what is meant by "rationality" in every situation? Exceptional situations involving real betting may be constructed in which the individuals, on average, could be better off even allowing the opponent to win something regardless of the outcome than if they strictly adhere to the probability calculus. Such examples particularly concern situations in which individuals change their opinion about the situation after having observed some data, but where they had specified a prior distribution that does not allow for radical enough changes. Some Bayesians accept that it is sometimes necessary to adapt the prior distribution retrospectively to information that comes in later even if this leads to incoherence (Box 1980, Dawid 1982). Apart from that, it cannot generally be taken for granted that rationality should always be interpreted in terms of gains and losses of money (see Habermas 1984 for a completely different perspective of rationality).

- Should another implicit assumption be accepted, namely that what happens later is independent of the behaviour of the betting indivuduals? This is obviously problematic for setups like the stock market, but even more so from a constructivist point of view that treats the future observations as personal and/or social constructs. It can often be

observed as well in scientific setups that the way experiments are carried out and data are gathered is indeed designed dependent on ealier assessments of evidence.

To these questions again the general remarks about mathematical modelling apply. Idealisations like assuming the individual to be forced to bet, or a formalisation of rationality in terms of money are necessary to set up any formal model in the first place, but there is no objective answer to the questions whether these particular idealisations are the ones to be adopted, and whether the benefits of formal modelling outweigh its problems. A constructivist way to decide in favour or against such idealisations analyses the implications of them on the world view and decides whether they are desired (which includes, if social acceptance is desired, whether they can be convincingly communicated). Here is an example. A university decides about applications for a certain programme and wants to use Bayesian posterior probabilities for later success in the programme as decision criterion. Applicants come from two different regions. The applicants have to carry out a test. Assume that the information that the university has to base its decision on is only the region and the test result of an applicant. Assume that past experience suggests that the probability is higher that applicants from region A are eventually successful in the programme than applicants from region B. It is known that the test result is associated positively with the probability of later success in the programme, modelled by assuming that there is an underlying "true" ability of every applicant of which the distribution of the test results and the probability of later success are monotone functions that do not depend on the region (but abilities distributions are allowed to differ between regions). Applying Bayes's Theorem then yields (without proof here) that the posterior probability of success of an applicant from region A is higher than that of an applicant from region B *with the same test result* .This means that the university, if it selects the applicants according to their success probabilities, commits itself to discriminating against equally qualified applicants from region B. Mathematically there is nothing wrong with this. However, the university cannot pass the responsibility for its discriminating behaviour on to Bayes's Theorem. It is actually a result of the university's decision (implied by the way the model was set up) to reduce the admission problem to a temporary betting rate problem, ignoring completely the effect that the admission policy of the university may have on future abilities (for example by denying potential applicants from region B "role models"; note that the term "underlying ability" is an interpretative wrapper for all kinds of factors that influence the success probability of an applicant, not just pure personal ability) and any possibility that the bad past success rate of region B students may have been caused by some issues in the university's education about which it could actually do something[14].

There are various attempts in the philosophy of statistics to come up with "solutions" for these issues (such as interval probabilities), and they may lead to improvements in certain situations, but they still have to deal with the basic problem of modelling. They come with their own implications, which can be analysed and criticised in a similar way, depending on the situation and the aims of those who model it or are involved..

Bayesians often criticise the frequentists for making supposedly objective assumptions about the world outside that cannot be verified. Constructivists may feel attracted to Bayesian subjectivism in particular, because of the explicit allowance for individual differences, and many frequentists (and some objective Bayesians) certainly appear to be philosophically naive by using this as a major objection against subjectivism. However, in the light of the discussion about mathematical models and reality, the frequentist assumptions about the world outside

[14] I got this example from Deborah G. Mayo by personal communication. Mayo used it to illustrate what she thinks to be a general flaw in Bayesian reasoning, but I rather think that it demonstrates that in some situations the implications of modelling run counter to personal and social constructs, and that the modellers should therefore try to be aware of these implications.

seem to stand on a rather equal footing with the Bayesian ones about rational reasoning. They are idealisations that are not made because they are believed to be objectively true, but that are necessary in order to take advantage of the benefits of mathematical modelling. Only the modelled domains to which they are applied are different for frequentists and Bayesians.

Computation of Bayesian posterior probabilities

The computation of the Bayesian measure of evidence, the posterior probability $P(H_0|data)$ for the null hypothesis that the ants do not prefer any of the Acacia species or for CHD being independent of strong coffee drinking, requires the specification of a prior distribution first. Following the subjectivist approach, the individual researcher (or a group of researchers; it would certainly make sense to involve at least one subject matter expert and one statistician) would need to think carefully about the situation to come up with a quantification of her prior belief, and also with convincing reasons for this to enable others to accept a result that depends on her choices.

An objective Bayesian, or a subjectivist without clear prior information and opinions, needs a prior probability distribution that models the absence of information. 2×2 tables are no standard case for Bayesian statistics and are not treated in every introductory book. I found two surprisingly different approaches how to do this.

The first approach was suggested by Jim Albert (2009:194–196). It is assumed (as in the second approach below) that the behaviour of the ant colonies is exchangeable. Under H_0, the probability for an A tree to be invaded is the same as the probability for a B tree to be invaded. Under H_1, these probabilities are assumed to differ. Non-informativity enters in two ways. Firstly, $P(H_0|p.i.) = P(H_1|p.i.) = 0.5$. This is based on the "principle of insufficient reason" to give any of the hypotheses more probability than the other one. Secondly, in order to compute $P(data|H_1)$, it is necessary to specify a distribution of probabilities for colonising species A and B trees within H_1, which is chosen to be the uniform distribution. Based on these choices, $P(H_0|data) = 0.005$. A straightforward subjectivist way to use this approach of specifying the prior distribution is just to change $P(H_0|p.i.)$. Even if for some reason (which may just be sensitivity analysis), $P(H_0|p.i.) = 0.95$, $P(H_0|data)$ is still as small as 0.09, so that the evidence clearly seems to indicate that the ants prefer species B (it can be computed that almost all the remaining probability is assigned to a preference for species B and almost nothing of it to preference for species A). Note that 0.09 still is a small value for a posterior probability, it still clearly points against H_0, whereas a p-value of 0.09 would rather be a borderline case. For the CHD data, this yields $P(H_0|data) = 0.959$ (non-informative prior) or $P(H_0|data) = 0.528$ for $P(H_0|p.i.)$ as small as 0.05, to illustrate another potentially extreme subjective choice.

Choosing $P(H_0|p.i.) = 0.5$ (or even any number larger than zero) means that positive probability is assigned to the idea of *precise* independence, and it can be argued that it is not realistic to believe in precise independence and the researcher should rather be interested in "practical independence" meaning very weak dependence (but keep in mind that the precise independence discussed here is not frequentist, and therefore it does not entail believing in precise independence in the world outside – though it is often carelessly interpreted in this way). Though assigning a nonzero to precise independence can be interpreted as approximating this in some sense, it can also be formalised in an alternative way, which is suggested by Peter Lee (2009: 152–153). Lee's approach starts by assuming a non-informative distribution for the two probabilities p and q, modelled as independent of each other, for a strong, and a weak or no coffee drinker to get CHD considering the CHD example (uniform distributions could be used for this but Lee suggests the so-called "Haldane prior"). The H0 of "practical independence"

can then be chosen by looking at the "odds ratio" $r = \frac{p(1-p)}{q(1-q)}$, which is close to 1 if q and p are about the same. So H_0 could for example be taken as "0.99 < r < 1/0.99". This yields $P(H_0|data) = 0.029$, which is totally different from the value of 0.959 following Albert's approach with non-informative priors. Note that Lee's approach involves a subjective decision about how close r has to be to 1 in order to speak of "practical independence" (though the posterior distribution as a whole does not depend on subjective decisions apart from the not-so-objective choice of an "objective prior"). If H_0 is taken as "0.8 < r < 1/0.8", $P(H_0|data) = 0.528$, still much lower than Albert's value. Analysing the posterior distribution further, it can be seen that the data still leave a strong uncertainty about p and q with large or at least non-negligible probabilities for both r < 0.8 and r > 1/0.8. Choosing a positive probability such as 0.5 for precise independence in Albert's approach has the effect that much of this uncertainty collapses into $P(H_0|data)$. Interpreting the Lee prior, it can be seen that prior independence of p and q is a quite strong form of non-informativity, because it entails that a very small rate of CHD cases among strong coffee drinkers is by no means informative about the CHD rate among weak or no coffee drinkers – it is not taken into account that CHD may be a rare disease overall (in Albert's approach, similar rates are more likely a priori through $P(H_0)$). This leaves the researcher with strong uncertainty even after having observed more than 1700 workers. Thinking it over, I realised that p and q should probably be restricted to be quite small and potentially similar a priori, and then the odds ratio approach would probably give more sensible results. It is a quite frequent phenomenon in statistics as well as in science in general that striving for objectivity basically implies that some standard approaches are chosen that cannot take into account the peculiarities of the given situation in a sensible way, whereas subjectivism allows for non-standard choices that may adapt better.

For the ants data, Lee's prior points even stronger against H_0 than Albert's one, so that the practical conclusions would probably be the same.

There is a Bayesian result stating that, under some assumptions, enough data eventually swamp the prior distribution, so that even if starting with different prior distributions, posterior distributions become more and more similar if more data are collected. However, this does not apply to a situation like the one above, where different principles of modelling were applied (nonzero vs. zero prior probability for precise independence). The computations show that different priors can lead to quite different posterior probabilities for H_0, and that it is not easy to understand all the implications of a chosen prior distribution, but this is needed in order to design it in a useful way.

6. Differences and connections between interpretations

Keeping in mind the idea that science aims at agreement, how problematic is it that there are several different approaches around to quantify evidence, sometimes leading to quite different results? Many practitioners are not very happy with this state of affairs, and in the statistical literature there are often attempts to "reconcile" the different approaches (see for example Berger 2003). However, another more or less explicit value of the scientific method (or, rather, my personal rather benevolent construct of it with which the reader may or may not agree) is that agreement should not be enforced artificially, and an agreement that is reached in an arbitrary way just because agreement is desired is not scientifically valid. Such attempts of reconciliation usually suppress some aspects of the original concepts that some people find worthwhile to keep visible, and they therefore rarely satisfy everyone (see for example Mayo's comment in the discussion following Berger's 2003 paper). In such a situation, there may be better chances of agreement accepting that different approaches have different merits and fulfil different aims. Destroying the myth of the "objectivity and unity of statistics" may be more worthwhile than looking for ways to pretend it more efficiently. It becomes more interesting to find guidelines about what to use when, and where irreducible elements of subjective decision

cannot be removed.

As explained before, a major difference between the frequentist and the Bayesian interpretations of probability is that the frequentist interpretation is about modelling mechanisms in the world outside whereas the Bayesian interpretations are about modelling rational reasoning. This alone is not of great help because in many situations researchers are interested in rational reasoning about the world outside. However, I can outline some guidelines for deciding between frequentist and Bayesian approaches:

- The Bayesian approach delivers a probability for H_0 given the data at the price that $P(H_0|p.i.)$ has to be specified first. If there is prior information or prior belief that can be convincingly formalised as such a prior distribution, and the researcher is happy that the outcome will depend on this, the Bayesian approach suggests itself. (Typically it is not "objective Bayes" then, because the prior distribution is actually informative.)

- On the other hand, there are decision problems in which it is explicitly not desirable to have an outcome that depends on prior beliefs, in which case obviously subjectivism cannot be used (though there are still irreducible subjective elements in model and prior specification and selection of cutoff values for p-values and posterior probabilities in the less explicitly subjective approaches). This is for example the case if fair and impartial decisions are desired about issues that affect people with opposing interests such as in court.

- p-values are about assessing the compatibility of data with certain idealised models for relative frequencies under repetition. A small p-value allows a statement of the kind "under the null model it would be almost inconceivable to happen what actually happened". This kind of statement may often be of interest, for example when checking the plausibility of certain scientific hypotheses about the world outside (without actually attempting to make the statement that they are "true") that can be formalised properly as such models[15]. Note that the evidence collected in this way can never be in favour, but only against the hypothesis, though evidence may be observed that is not against the null hypothesis of interest but against certain alternatives, so that at least some competitors of the H_0 could be discarded.

- In some situations the aim is prediction and potential future benefits can be properly quantified, for example if it is about financial issues, or (involving some more complexities) something like drilling for oil. In such situations, the betting rate metaphor for Bayesian probabilities can be seen as quite directly related.

- Generally the social system to which the researchers belong or want to communicate their results plays a role as well, particularly if and how it can be expected that agreement can be reached about the involved choices of a model, prior distributions and decision rules.

Up to now the focus was on the differences between interpretations. On the other hand, they use the same label (probability) and the same calculus, and the connection between them is not only mathematical. The perception of a separation between the two interpretations of probabilities as degrees of belief on one hand and related to randomness in the world outside on the other hand came up around the middle of the 19th century (Gillies 2000:19 traces it back to a remark by Poisson in 1837), more than 100 years after the beginnings of the probability calculus. In response to the formalist ideas that Hilbert started to present around 1900 (see Hilbert 2004),

[15] The connected Neyman-Pearson approach to hypothesis testing (Neyman & Pearson 1933), which does not exactly deliver a "quantification of evidence" but rather a binary decision, utilises the frequentist interpretation to give "error probabilities" for making wrong decisions, which may be of interest as well, based on the model assumptions.

Kolmogorow (1933) axiomatised probability mathematically in a way compatible with both betting rates and relative frequencies, which separated the calculus explicitly from its various interpretations. In the very early works (for example Bernoulli 1713) probabilities were defined as ratios of numbers of favourable and existing events (for example 1/6 for a die to roll a "2"; note that because Bernoulli did not yet separate the mathematical formalism explicitly from its real world interpretation, the word "definition" is chosen here instead of "interpretation"). These were treated as *identical* with fair betting rates and expected relative frequencies under repetition. A difference between these ideas was not yet constructed. Applying the probability calculus to new problems and wider areas (biased dice, death probabilities in age classes, reliability of astronomical observations etc.) required some extensions from which the differentiation of the interpretations emerged. My interpretation of this is that using probability calculus for many phenomena involved treating them in terms of a dice/gambling metaphor, but for some of the phenomena, the "fair betting rate" aspect of this metaphor worked better whereas for others the "relative frequency" aspect became dominant. At some point some people realised that, unless cases such as fair dice are treated, these aspects may have quite different implications, but instead of accepting them as essentially different views both of which have their merits, most probabilists (if they were interested in the issue at all) started to advocate either of these points of view as the "best" or "correct" one. The use of the same word "probability" for them and the general myth of "objectivity of science" suggested that there should only be one correct meaning. To some extent, these views still persist today.

Despite the pluralist position that I take in most parts of the present paper, I find it helpful to acknowledge the deep historical connection between the different approaches in order to understand why they are still so often perceived as directly competing and why there is such a big amount of statistical literature that uses frequentist and Bayesian approaches in a very eclectic way without bothering about differences in interpretation.

There is a vast literature comparing Bayesian and frequentist interpretations of probability including some further interesting aspects that could also be relevant for deciding in a given situation between approaches, see for example Finetti (1970), Fine (1973), Bernardo and Smith (1994), Mayo (1996), Gillies (2000). Most of these authors highlight issues with the interpretations as reasons to attack or defend one of them for use in more or less general situations. A constructivist way of reading these arguments would be different; they can illustrate what kind of "ideal" world view is entailed by these interpretations, and therefore they help with the casewise decision between the approaches, but also with keeping in mind what aspects of reality are suppressed when adopting one of them. Obviously, there is no need for a constructivist to adopt one of the interpretations exclusively. However, given that constructivists aim at understanding how their construction processes work, it seems unsatisfactory as well to follow the kind of eclectic approach that is often found in the scientific literature and that ignores the deep philosophical issues of interpretation. In a given situation, it seems sensible to adopt an interpretation explicitly and to discuss honestly its implications and the restrictions and ignored aspects involved by it, making it clear that this is always a matter of choice, even though good case-dependent arguments may exist.

7. Conclusion

As far as I see it, applying constructivism in statistics does neither necessarily lead to new methodology, nor to discarding old ones. It is rather about considering the methodology and its underlying assumptions in a particular way. For a constructivist, the following questions could be of major importance:

- How do we (the researchers) see our topic, which aspects do we want to model, which aspects do we want to ignore, considering our aim of modelling?

- What a point of view is entailed by the models that we use and our interpretation of them, and how does this relate to what we think about our topic?

- What is the communicative value of the model and the quantification of evidence,

how can it help to support understanding and agreement, to make decisions, draw conclusions, and to give others the chance to disagree in a constructive way?

- Do we really want and/or need to measure evidence by a single value in the given situation?

These questions emphasise the responsibility of the researchers, and are far too often ignored by an attitude that the data should decide "objectively" what the correct model is and what the evidence suggests. Much of the discussion in the previous sections was about exploring the points of view implied by various ways to quantify evidence. Such considerations are hopefully useful when addressing the questions above.

In particular, the role of model assumptions changes when adopting a constructivist point of view. By interpretation, model assumptions translate into ways of thinking about a situation. Frequentist models mean that the researchers think of the modelled phenomena using a metaphor of repetitive result-generating mechanisms. Bayesian models mean that the researchers think about the way how they should rationally learn from data using a metaphor of betting rates. Note that in Bayesian as well as frequentist statistics assumptions interpreted as regarding situations as in some sense identical repetitions are required in order to get the calculus going, regardless of whether it is believed that these are "true" or "rational". Gillies (2000:77-81) demonstrated that Bayesian exchangeability has interpretational implications that are not weaker than frequentist independence. In Hennig (2007) I showed that attempts to test frequentist independence necessarily lead to a paradox. Probability is always about what could have happened apart from what actually happened. In this sense, it always involves what is essentially unobservable. Probability statements, and therefore quantifications of evidence, can never be checked by observation alone

Much more detailed analyses of the models are possible, giving the frequentist for example probabilities of observing low p-values if the H_0 is wrong in certain ways (power analyses), and giving the Bayesians predictive probabilities for all kinds of conceivable future events. To the constructivist, these analyses show what her constructions imply, and therefore they enable a very precise understanding of the implications of the modelled ways of thinking. This is a major benefit of probability modelling. The role of the model assumptions (and to compare their expected implications with the data at hand, which can be interpreted as a constructivist version of "model checking") is then rather to assist the researcher in finding out whether the chosen data analytic method may lead to undesired or misleading results for the given data. Unfortunately this is often totally obscured by the usual way to communicate results from probability modelling such as p-values and posterior probabilities that attempts to hide the responsibility of the researchers. This is particularly obvious in the teaching of statistics, which regularly leaves many intelligent students confused about the contrast between the apparent necessity to check whether model assumptions "really" hold and the striking impossibility to do this in any satisfactory way. The result is that many of them lose any interest in statistics whereas others adapt to the "usual scientific rituals" and start to apply the formal calculus in an unreflected and uncritical way.

The subjectivist Bayesian approach could be a positive exception in this respect, but unfortunately many researchers who follow this approach shy away from assuming responsibility openly for their prior choices. From a constructivist perspective, however, all interpretations share the general problems and merits of mathematical modelling, and each of them could be applied with a constructivist attitude.

There are many potential practical implications of the view outlined in the present paper. I have most personal experience with its influence on my approach and my way of communicating as a statistical consultant or collaborator of non-statisticians. Dealing with models in terms of decisions how to see a problem makes the statistical modelling process seem much less mysterious, and it makes the connection of the researcher's decisions to the chosen models and the results much clearer. It also gives the researchers clearer ideas about the existing possibilities to see the problem in a different way.

It seems to me to be more difficult to me, up to now, to give these ideas a stronger influence on my teaching. I try to do so, and they certainly have an influence on some of my students. But in teaching there is always limited time and the constructivist view runs counter to what most students expect and are told in other courses. There is certainly a lot of potential here for innovative teaching ideas. How much constructivist attitude can and should students learn (and how) in their statistics courses?

The use and presentation of statistics and quantification of evidence in the general public is another interesting issue. Although the presented ideas imply a quite optimistic view of the potential of science and the value of statistical modelling, to some extent they undermine the way people perceive the authority of scientific results that are based on such quantifications. Will it be possible to communicate their potential value along with the view that "finding out how the world really is" is not exactly what this value is? Quite often I have experienced that in discussions in which quantifications of evidence played a role less objectivistically (and probably more constructivistically) minded persons were generally very sceptical toward the use of statistics and statistical models. The way these models are often used gives good reasons for such a sceptical attitude. But I think that they can be used in many situations in a constructive and helpful way, if often with a more modest attitude that is more open to criticism than the one that dominates currently.

References

Albert J. (2009) Bayesian computation with R (2nd edition). Springer, New York.

Bayes T. (1763) An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London 53: 370–418.

Berger J.O. (2003) Could Fisher, Jeffreys and Neyman have agreed on testing (with discussion)? Statistical Science 18 (1): 1-32.

Bernardo J. M. & Smith A. F. M. (1994) Bayesian theory. Wiley, Chichester.

Bernoulli J (1713) Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis. Thurneysen, Basel.

Box G. E. P. (1980) Sampling and Bayes inference in scientific modelling and robustness. Journal of the Royal Statistical Society Series A 143: 383-430.

Davies P. L. (2008) Approximating data (with discussion). Journal of the Korean Statistical Society 37(3): 191-211.

Dawid A. P. (1982) The Well-Calibrated Bayesian. Journal of the American Statistical Society 77:605-610.

Derksen T. (2007) Lucia de B. Reconstructie van een gerechtelijke dwaling. Uitgeverij Veen Magazines BV. For English information see http://www.luciadeb.nl/english/derksen-book-1.html

Edwards A. W. F. (1972) Likelihood. Cambridge University Press.

Fine T. L. (1973) Theories of probability. Academic Press, New York.

Finetti B. de (1970) Teoria delle probabilità. Einaudi, Torino. English translation: Finetti, B. de (1974) Theory of probability. Translated by A.F.M. Smith. Wiley, New York.

Fisher, R.A. (1935) The Logic of Inductive Inference. Journal of the Royal Statistical Society, Series A 98: 39-54.

Foerster H. von (1984) On constructing a reality. In: Watzlawick P. (ed.) The invented reality. W. W. Norton, New York: 41–62.

Gergen K. J. (1999) An invitation to social construction. Sage, Thousand Oaks CA.

Gillies D. (2000) Philosophical theories of probability. Routledge, London.

Glasersfeld E. von (1995) Radical constructivism. A way of knowing and learning. Falmer Press, London.

Glasersfeld E. von (2008). Who conceives of society? (With open peer commentaries.) Constructivist Foundations 3 (2): 59-108.

Greenland S. & Mickey R.M. (1988) Closed form and dually consistent methods for inference on strict collapsibility in 2 x 2 x K and 2 x J x K tables. Journal of the Royal Statistical Society. Series C (Applied Statistics) 37 (3): 335-343.

Habermas J (1984) The Theory of Communicative Action Volume 1; Reason and the Rationalization of Society, Cambridge: Polity Press.

Hennig C. (2007) Falsification of propensity models by statistical tests and the goodness-of-fit paradox. Philosophia Mathematica 15: 166-192

Hennig C. (2009) Mathematical models and reality – A constructivist perspective. Research Report no. 304, Department of Statistical Science, UCL To appear in Foundations of Science. Retrieved on 29 September 2009 from http://www.ucl.ac.uk/Stats/research/reports/abs09.html#304

Hilbert D. (2004) David Hilbert's lectures on the foundations of geometry, 1891–1902. Edited by Hallett M. & Majer U.. Springer, Berlin.

Howson C. & Urbach, P. (2006) Scientific reasoning: the Bayesian approach. Open Court, Chicago.

Kass R. E. & Wasserman L. (1996) The selection of prior distributions by formal rules. Journal of the American Statistical Association 91: 1343–1370.

Kolmogorov A. N. (1933). Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin.

Lee P. M. (2009) Bayesian statistics (3rd edition). Wiley, Chichester.

Luhmann N. (1995) Social systems. Stanford University Press, Stanford.

Manzano, M. (1999). Model theory. Oxford University Press, Oxford.

Mayo D. G. (1996) Error and the growth of experimental knowledge. University of Chicago Press.

Mayo D. G. & Kruse M. (2001) Principles of Inference and their Consequences. In: Cornfield D. and Williamson J. Foundations of Bayesianism. Kluwer Academic Publishers, Dordrecht, 381-403.

Mises R. von (1928) Wahrscheinlichkeit, Statistik und Wahrheit, Springer, Berlin. English translation: Mises. R. von (1981) Probability, Statistics and Truth. Dover, New York.

Neyman J. & Pearson E. (1933) On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philosophical Transactions of the Royal Society London, Series *A* 231: 289-337.

Paul O. (1968) Stimulants and coronaries. Postgraduate Medical Journal 44: 196-199.

Sokal R. R. & Rohlf F. J. (1981) Biometry (2nd edition). W. H. Freeman, San Francisco.

Walley P. (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall, London.