JENA ECONOMIC RESEARCH PAPERS

2013 – 025

Confirmation: What's in the evidence?

by

Mitesh Kataria

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Jena Economic Research Papers 2013 - 025

Confirmation: What's in the evidence?

Mitesh Kataria*

Abstract

In this paper, I discuss the difference between accommodated evidence (i.e. when evidence is known first and a hypothesis is the proposed to explain and fit the observations) and predicted evidence (i.e., when evidence verifies the prediction of a hypothesis formulated before observing the evidence) from a behavioral as well as a statistical perspective. Using a factorial survey on a sample of students, I show that predicted evidence is perceived to constitute stronger confirmation than accommodated evidence. This position deviates from the standard Bayesian epistemological theory of confirmation where accommodated and predicted evidence constitute equally strong confirmation. The findings suggests that trusting a model to predict correctly is intrinsically related to trust in the proposers' (i.e., the scientists') level of knowledge, and relatively more subjects are persuaded by a proposer's abilities if the proposer is successful in predicting rather than accommodating evidence. The existence of an indirect relationship between hypothesis and evidence can be considered to impose undesirable subjectivity and arbitrariness on questions of evidential support.

Keywords: Evidence, Prediction, Methodology
JEL classification: C11, C12, C80

*Max Planck Institute of Economics, Kahlaische Str. 10, 07745 Jena, Germany. Tel.: +49 3641 686632; fax: +49 3641 686687. Email addresses: kataria@econ.mpg.de

1 Introduction

Is it relevant for scientific confirmation whether evidence is known first and a hypothesis is then proposed to explain and fit it (henceforth called accommodation), or whether evidence verifies predictions from a hypothesis formulated before observing the evidence?[1] Bayesian logic implies that the difference is irrelevant and that what is important is the logical relationship between the hypothesis and the evidence, not whether evidence is predicted or accommodated. A hypothesis is confirmed if the posterior is greater than the prior, and this occurs if evidence supports the hypothesis independent of the timing of the empirical claim. The difference between accommodation and prediction is only the timing of the empirical claim and therefore irrelevant for scientific confirmation. But in considering evidence to confirm a hypothesis, do people really pay equal attention as to whether the evidence is accommodated or predicted?

Musgrave (1974) discusses the possibility of how to move away from the purely logical approach to confirmation by a detailed review of three different views of the historical approach. As opposed to the purely logical approach, the historical approach to confirmation holds that predicted evidence is more important than accommodatedevidence unless special circumstances prevail. The first, strictly temporal, view of background knowledge holds that facts known before a hypothesis is proposed cannot confirm the hypothesis since it is already part of background knowledge. But for many this view is too conservative. Taking a contemporary example from the social sciences, Fehr and Schmidt (1999) formalized the notion of fairness in a model where people are averse to inequities. Their model was shown to be consistent with accommodated evidence from a number of different economic experiments. Judging by the impact of their paper on the experimental and behavioral economic literature, it seems fair to conjecture that accommodated evidence has some merits.[2]

[1] In the case of prediction the hypothesis is usually partly based on existing observations, however, a prediction requires the empirical claim to be verified by at least some observations that are made after the empirical claim Lipton (2005).

[2] A famous historical example is that Einstein showed that general relativity agrees closely with the observed amount of perihelion shift, which was not the case with Newtonian physics. Although the motion of the perihelion of Mercury was known long before Einstein proposed his theory, the evidence was considered to support the theory and to be a powerful argument motivating the adoption of general relativity.

The second, heuristic, view of background knowledge claims that an old fact can be novel to a new theory, provided the theory has not been constructed to explain the fact but is still in the process of explaining it. Finally, the third view of background knowledge holds that a new theory is independently testable or novel where old facts can confirm the new theory if and only if its prediction is unique such that it cannot be explained by the old theory or contradict it.

Kahn et al. (1996) developed a model that focuses on different scientific methods. It is shown that if the scientist has different abilities to propose truthful theories and can choose either to predict the evidence or construct a theory that accommodates it, an observer will have a stronger belief in the truthfulness of the theory, if the theory is proposed before the evidence has been considered. The observer updates the probability that the consistent theory is proposed by a scientist with greater ability to propose truthful theories if evidence supports the theory, thus providing stronger confirmation. If the scientist constructs a theory that accommodates evidence, however, nothing is learned about the scientist's type, and no updating takes place. While Bayesian epistemology traditionally avoids the relationship between evidence and personal or psychological attributes, which is considered to impose undesirable subjectivity and arbitrariness on questions of evidential support, Kahn et al. (1996) focus on this link.[3] That personal attributes matter is also argued elsewhere (e.g., Hitchcock and Sober 2004; Lipton 2005) by claims that a scientist who accommodates evidence is susceptible to the temptation of over-fitting, i.e., explaining patterns that have appeared by chance. The choice of a scientist to be more worried about over-fitting or under-fitting the data (i.e., the risk of neglecting existing patterns) is, of course, subjective.

In this paper, I first test whether students with various backgrounds take a purely logical stand on confirmation or whether they believe in a research hypothesis that predicts evidence more than in one that accommodates it. Second, I test whether prediction constitutes stronger confirmation than accommodation because the observer infers that the scientist is more knowledgeable when the prediction is correct (e.g., Kahn et al. 1996). As far as I know, there are no empirical studies that address these questions.

[3]Norms in science value universalism which means that a person's attributes and social background is irrelevant to the scientific value of a person's ideas.

The reason might be that it is only recently that philosophy, where these questions have gained most attention, has been opening up to empirical work that takes folk intuitions into account as opposed to the traditional approach of armchair philosophy. While theoretical work can describe how agents ought to behave, empirical work can offer an understanding of the psychological processes underlying such intuitions. Moreover, communication of scientific results to nonscientists is at times not only important but crucial, e.g., regarding the science of climate changes where it is essential to understand how nonscientists judge scientific evidence.[4]

The remainder of the paper is organized as follows. In section 2, the Bayesian and frequentist inference is discussed and hypotheses are stated. In section 3 the experimental design is explained, followed by the results in section 4. In section 5 I discuss some methodological concerns before concluding in section 6.

2 Bayesian and Frequentist inference

The belief a (Bayesian) agent assigns to some (binary) hypothesis (H) given evidence (E) is expressed by Bayes' theorem:

$$P(H|E) = \frac{P(E|H)\cdot P(H)}{P(E)} = \frac{P(E|H)\cdot P(H)}{P(E|H)\cdot P(H)+P(E|notH)\cdot P(notH)}$$

Dividing the numerator and denominator in the above equation with $P(E|H)$ we have $P(H|E) = \frac{P(H)}{P(H)+\frac{P(E|notH)}{P(E|H)}\cdot P(notH)}$ stating that $P(H|E) = f(P(H), \frac{P(E|notH)}{P(E|H)}$ where f is an increasing function of the first argument which is the prior probability and decreasing function of the likelihood ratio $\frac{P(E|notH)}{P(E|H)}$. Hence, for a given value of the likelihood ratio, the posterior probability $P(H|E)$ increases with the prior. Furthermore, for a given value of the prior, the posterior probability of H is greater, the less probable E is relatively to $notH$ compared to H. Hence, the more likely it is that certain observation is made if (and only if) the hypothesis is true the stronger is the confirmation if the observation is actually made. For our purposes the important message is

[4]Though many climate science studies show evidence that a long-term change in the average atmospheric temperature could occur, the public persists in distrusting the results. Whether the climate predictions for the year 2100 are correct or not will be revealed by the end of the century. Whether short-term predictions can reduce the credibility gap between scientists and the public is a question that partly inspired this project.

that the Bayesian formula does not distinguish prediction from accommodation as there is nothing that suggests that the timing of the evidence is important relatively to the timing of the hypothesis. The reason is that the likelihood ratio is unaffected by the timing of the evidence.

We now turn to a simple example of classical frequentist inference. Assume that theories can be divided into four mutually exclusive sets. Constructing theories involves some randomness and is modeled as the selection of a ball from an urn, the balls representing possible theories. Set A consists of true theories and is drawn with probability p, set B consists of false theories that could be falsified in future but are consistent with current observations and is drawn with probability q. Set C consists of theories that are consistent with all past observations but not with the latest observation(s) and is drawn with probability $1-p-q$. Finally, set D consist of inconsistent and false theories. We will assume that the scientists avoid drawing from set D so that the urn contains no type D balls. A scientist who predicts evidence by constructing a theory before learning that the theory is consistent with past and new observations, has constructed a theory which is true with probability $P(A|A\&B) = \frac{p}{(p+q)}$. A scientist who considers the evidence before constructing theory avoids drawing from set C and set D, and the probability that the theory is true is again $P(A|A\&B) = \frac{p}{(p+q)}$. Hence, the first and main result is:

"*... the probability that a theory is true, conditional on its being consistent with all (old and new) observations is independent of the research strategy.*" (Kahn et al., 1996).

Note that the simplistic model fails to specify a decision framework of when to reject a hypothesis from data that is subject to random variation. In a decision framework a hypothesis is a conjecture about the distribution of one or more random variables. A statistical hypothesis defines the rule of when to reject the conjecture. In a single hypothesis test, an acceptable maximum probability of committing a Type 1 error (i.e. rejecting the true null hypothesis) is defined which is known as the testwise alpha. This is often compared to a p-value which states the probability under the null to observe the sampled or more extreme data. If the probability is sufficiently small (i.e. p-value $< \alpha$) the data is considered to be too extreme to be explained as an outcome of chance and therefore viewed as evidence against the null

hypothesis.[5] The first main result can in the decision framework be stated as the probability to observe the sampled or more extreme data under the null is independent of the scientists' behavior such as if the scientist considers the data or not before proposing the hypothesis. As I will discuss in section 5, not everyone agrees with this view.

Finally, let us consider an indirect model of confirmation presupposing a link between evidence and personal attributes. Kahn et al. (1996) shows that if there are different types of researchers, where one type is more likely to construct true and consistent theories independent of the research strategy chosen, then the outcome of the prediction conveys information to an observer about the type of researcher. Instead of the selection of a ball from one urn, this could be thought of as a draw from one of two types of urns, one containing more true and consistent theories and the other containing more inconsistent theories. Assume the following for the two types i, and $j = 1 - i$: $P_i(A) + P_i(B) = p + q > P_j(A) + P_j(B) = r, +s$. As before we assume that the inconsistent theories are drawn with probability $1 - P_i(A) - P_i(B)$ for scientists type i and $1 - P_j(A) - P_j(B)$ for scientists type j. The posterior probability that the theory is suggested by the high ability type i if the scientist theorizes first and the theory is consistent is found using Bayes' theorem:

$$P(i|A\&B) = i' = \frac{P(i)P(A\&B|i)}{P(A\&B)} = i\frac{p+q}{i(p+q)+j(r+s)} > i$$

Assume that the scientist who is more likely to construct true and consistent theories is also more likely to suggest true theories given they are consistent i.e. $\frac{p}{p+q} > \frac{r}{r+s}$. From an observer's' perspective, the probability γ' that the theory is true if the scientist theorize first and the theory survives the test is higher than the probability γ that the theory is true if the scientist constructs the theory after having considered the data since in this case no updating takes place.

$$\gamma' = i'\frac{p}{p+q} + j\frac{r}{r+s} > \gamma = i\frac{p}{p+q} + j\frac{r}{r+s}$$

[5]Bayesian inference is based on the probability that a hypothesis is true given data i.e. $P(H|E)$, while the frequentist inference is the inverse probability which is $P(E|H)$. Notably they are not the same e.g., assuming that the probability to die after falling from a high building is 1, does not mean that a person who's dead has fallen from a high building.

We can now state the second main result:

"Assume that the scientist's type is unknown. Then the probability that the theory is true, conditional on its being consistent with all (old and new) observations, is higher if the scientist theorized first than if he looked first." (Kahn et al., 1996).

The first result is neutral in the sense that it is assumed that an additional observation before suggesting a fitting hypothesis makes the scientist neither better nor worse off. If, on the other hand, prediction constitutes stronger confirmation, it implies that if there are two scientists who propose identical theories, the only difference being that one scientist is more careful in observing a full sample of events while the other, less observant scientist neglects a subsample of the events that are predicted, the more observant scientist will be penalized for being knowledgeable about the occurred events and made relatively worse off (see, e.g., Musgrave 1974). The mistrust toward the scientist who observes more could be rooted in the belief that more observations increase the risk of over-fitting the data.[6]

Our survey experiment is designed to test two hypotheses. The first null hypothesis is:

H0: *People believe that the probability that a theory is true is independent of whether evidence is predicted or accommodated, i.e., the research strategy of the scientist.*

In case the first null hypothesis is rejected, we want to test if people trust scientists who theorize first to be more knowledgeable. The second null hypothesis is:

H0: *People do not believe that a scientist who makes a correct prediction is more knowledgeable than a scientist who accommodates evidence.*

[6]One can argue that the literature takes an asymmetric stand on this issue as it seldom discusses the risk of under-fitting the data (i.e. the practice of missing robust structural relationships by insisting on a-priorism) while the phenomena of over-fitting have gained much more attention.

3 Factioral Survey Design

The data for this study was collected using an online factorial survey with three treatments conducted in Jena, Germany, in May 2013. The subjects were students with various educational backgrounds and were recruited using the Online Recruitment System for Economic Experiments (ORSEE). Participation was incentivized by a lottery procedure with a 10 percent probability to win 25 euros. In total, 243 (=81x3) subjects were recruited.

In the baseline treatment, T1, subjects are presented with a scenario where a research team accommodates data before constructing the theory. The model of this scenario in the baseline treatment will henceforth be referred as the *A-mode*l. In the second treatment, T2, subjects are presented with a scenario where a research team accommodates data before constructing the model, which results in five predictions that are subsequently shown to be correct. The model of this scenario will henceforth be referred as the *P-model*. Both the A-model and the P-model share the feature that they are partly based on existing observations. However, the main difference is that the P-model is verified by at least some observations that are made after constructing the model, which is essential for the conceptual difference between accommodation and prediction (Lipton, 2005). In both treatments, subjects are asked about how likely they think it is that the model will be correct in future predictions. In the third treatment, T3, subjects are presented with a scenario where two teams of scientists each employ one of the two research strategies to "accommodate"or "predict" and subjects are asked whether they believe that either of the two teams is more knowledgeable and about how likely they think it is that the predictions of the two teams will be correct in future outcomes. The treatments are summarized in Table 1, followed by the scenarios in T1 and T2. Unique features of the scenario in T1 are marked with curly brackets { } while unique features of T2 are marked with square brackets []. The scenario in T3 is presented in the Appendix.

Table 1: Treatments

Treatment	No. Obs	Description
T1	81	A scenario with no predictions
T2	81	A scenario with five successful predictions
T3	81	A joint scenario

El Nino is characterized of abnormal warm ocean water temperatures that occasionally develops off the western coast of South America and can cause climatic changes across the Pacific Ocean. El Nino events happen irregularly. The El Nino phenomenon drastically affects the weather in many parts of the world. Developing countries which are dependent on agriculture and fishing, particularly those bordering the Pacific Ocean, are the most affected. If forecasts could provide warnings before an El Nino episode, human suffering and economic losses could be reduced. Please consider the partly fictional scenario below and answer the question.

A team of scientists have recently constructed a new hybrid model, where an ocean model is coupled to a statistical atmospheric model that accommodates {25} [20] of the known El Nino events of the twentieth century (i.e., $1901 - 2000$). The model is constructed to fit observations that have already been made. Using old data (i.e., the known El Nino events), the model is rigorously tested and able to detect El Nino events 12 months before it starts in {20} [15] of the {25} [20] cases without causing any false alarms. Without any prior knowledge, the chances to detect El Nino before it starts is only 5 percent. A model is considered to be good, if it detects El Nino events 12 months before they start with a probability of 80 percent. {The model has never been tested regarding how well it predicts future El Nino events but has rigorously been tested using old data.} [The model has recently also been tested on how well it predicts future El Nino events. More specifically, after the model was developed, the El Nino event has occurred 5 times and the model successfully predicted these events 12 months before they started in all of these 5 cases without causing any false alarms. In total that would imply 15 correct predictions using old data and 5 correct predictions using new data, i.e., a total of 20 correct predictions out of 25.]

At this stage, a few remarks are necessary about the scenarios before we discuss the results.[7] First of all, note that the performance of the two models in the two scenarios is the same in terms of the ability to detect the El Nio events with a probability of 80 percent. Also note that the total number of events is kept the same in the two scenarios to ensure comparability. The main difference between the two scenarios is the amount of evidence that is accommodated and predicted. In T1, the scenario consist more confirmation of accommodated evidence while in T2 the scenario consist more of predicted evidence. Moreover, in our scenarios a scientific hypothesis is

[7]For more information about El Nio see e.g. Cane et al. (1997).

represented by a model. While a model can be used to represent a scientific hypothesis, these terms should in general not be used interchangeably. Furthermore, while a hypothesis can be true or false, this terminology seems less appropriate for discussing models, and we will talk about models that are good or bad. In our experiment, the task of the subjects is to state their beliefs that the model(s) in the scenarios will make correct prediction(s) in the future, given the evidence they have about the performance of the model(s) and how the model(s) were developed. In particular, we are interested if their beliefs differ depending on whether evidence is accommodated or predicted. More specifically, to interpret the responses to the scenarios in a Bayesian framework I assume that the subjects have prior beliefs whether a certain model is good or bad. Given the parameterization of our scenarios and assuming the existence of two types of models, a good model is defined to make 80 percent correct predictions and a bad model to make 5 percent correct predictions. Subjects are presented with evidence of whether the model produces correct predictions. Based on the evidence, subjects are assumed to update their posterior beliefs about whether the model is good (positively if confirming evidence) or bad (positively if disconfirming evidence) using Bayes' theorem of binary hypotheses. A subject who is sure that the model is good should infer that the model predicts correctly with a probability of 80 percent. In case the subject is sure that the model is bad, she should infer that the model predicts correctly with a probability of 5 percent. Subjects are asked how likely they think it is that the prediction of the model will be correct in future outcomes, i.e., they are asked as to their beliefs about the performance of the model(s), which in turn depends on whether they believe that the model is good or bad and should be sufficient to address our research question whether accommodated evidence is treated differently than predicted evidence. Alternatively, we could have asked the subjects more directly about their posterior beliefs that the model is good. The reason we asked them indirectly was that it also facilitated a classical frequentist interpretation of the scenarios. We do acknowledge that it would have been naive to expect all subjects to behave as Bayesians, given the information provided in the scenarios. Our aim was to present the subjects with a short but content-rich and meaningful scenario without nudging them toward applying either the classical frequentist or Bayesian framework.[8]

[8]For example, I chose not to emphasize the existence of only two types of models, i.e.,

Let us now turn to the question of how a frequentist might react to the scenarios. For a frequentist the probability for the El Nio event to be detected in the future equals the relative frequency of detection to occurrence of the event in the past. Hence, given that subjects weight accommodated and predicted evidence equally, the probability that the prediction of the two models will be correct in future outcomes should be the same in the two scenarios.

4 Results

In this section the three main results will be presented followed by the statistical support. The first result utilizes the single scenario data to test the difference between accommodated and predicted evidence, the second result compares the result of the single scenario data with the joint scenarios data, and finally the third result focuses on the joint scenarios data to test for subjective (psychological) links between evidence and the proposer of the evidence.

Result 1: *A model that predicts evidence is assessed to be more correct than a model that accommodates evidence.*

The probability that model-P will be correct in future prediction is appreciated to 77 percent while model-A is appreciated to make correct prediction with a probability of 65 percent. The difference is statistical significant for any conventional significance level using a two-sided Mann-Whitney-Wilcoxon test (p-value: 0,030).

Result 2: *Trust in a model that accommodates evidence increases when the model is (side-by-side) compared to a model that predicts evidence.*

The probability that model-P will be correct in future prediction is appreciated to 76 percent in the joint scenario which is not significant different to 77 percent which is assessed in the single scenario analysis. Model-A, however, is appreciated to make correct prediction with a probability of 72 percent in the future in the joint scenario. The increase from the 65 percent in

the good and bad type, while I was still giving information so that such an interpretation can have been made.

the single scenario analysis is statistical significant using a two-sided Mann-Whitney-Wilcoxon test. Hence, people take more of a logical approach when both models are compared side-by-side.[9]

The increase notwithstanding, a model that predicts evidence is still assed to be more correct than a model that accommodates evidence. The difference is statistical significant for any conventional significance level using a two-sided signed-rank test. Hence the effect that prediction constitutes stronger confirmation is robust to the framing of the problem.

Result 3: *Trust in a model's capacity to predict seems to be intrinsically (but undesirably) related to trust in the scientists' level of knowledge.*

Note that most subjects (60%) do not believe that a scientist who predicts evidence is more knowledgeable than a scientist who accommodates evidence. However, a substantial share (32%) of subjects believes this is the case. Only a small share of subjects (8%) believes that a scientist who accommodates evidence is more knowledgeable.

Subjects that believe that a scientist who accommodates evidence is equally knowledgeable to a scientist who predicts evidence, trust the P-model (which uses predicted evidence) to be correct in future predictions with a probability of 77 percent and the A-model (which uses accommodated evidence) to be correct with a probability of 74 percent. This small difference is not statistically significant for any conventional significance level using the Wilcoxon signed-rank test.

Subjects that believe that a scientist who predicts evidence is more knowledgeable than a scientist who accommodates evidence, trust the P-model to be correct in future predictions with a probability of 77 percent and the A-model to be correct with a probability of 65 percent. This difference is statistically significant for any conventional significance level using the Wilcoxon signed-rank test.

The third result offers interesting insights. In the philosophical literature many attempts has been done to revise the purely logical approach to confirmation and to show that predictions constitute stronger confirmation than accommodation. The aim has been to formulate a theory without

[9]Forty-nine (49)% of the subjects stated that the A-model and P-model will be equally correct in future prediction(s), while 37% stated a higher probability for the P-model and 14% stated a higher probability for the A-model.

introducing undesirable subjectivity in the relationship between hypothesis and evidence. However, no theory has to date been widely accepted. Interestingly, our results show that the intuition that predicted evidence constitute stronger confirmation is driven by subjects' that not only judge the relationship between the hypothesis and the evidence but that uses the evidence to infer something about the abilities of the scientist which in turn provides stronger confirmation.

5 Discussion

In section 2, I proposed that accommodated and predicted evidence constitute equally strong confirmation. In this section, I discuss plausible objections to such a proposition and in favor of predicted evidence.

Turning back to the scenarios, note that there is a search involved in accommodating the data to find the model that is considered to best explain the data. Practicing econometricians routinely deal with such a search. After a vivid search the performance of a selected model is usually evaluated. But this routine is most likely in conflict with what is taught even in introductory econometric classes. In introductory econometric textbooks of classical statistics students often learn that theory or hypothesis should precede the obtaining of the data (see, e.g., Maddala, 2001).[10] The reason behind such an approach relates to the literature of data mining (Leamer, 1983), which is also known as data snooping (White, 2000) and deals with the problems involved in conducting an extensive specification search of a model. One way to articulate the problem of data mining is that in any specification search there will be a multiple amount of hypotheses tested, while the tools that econometricians traditionally apply are developed for testing a single hypothesis. In a single hypothesis test, an acceptable maximum probability of committing a Type 1 error (i.e., rejecting the true null hypothesis) is defined, which is known as the testwise alpha. If the number of hypotheses that are tested increases, however, so does the probability that

[10] An exemption to this structure is found in a famous book chapter on how to write empirical papers. Bem (2003) provides the following advice to students in psychology: There are two possible articles you can write: (a) the article you planned to write when you designed your study or (b) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (b).". Bem is more concerned with under-fitting the data and missing the chance to discover than with over-fitting the data.

at least some Type 1 errors are made. This means that the probability that a researcher wrongly concludes there is at least one statistically significant effect across a set of tests increases with each additional test. For example, in testing 100 independent true null hypotheses the expected number of false significant tests is 100x0.05=5 at the 5% significance level, and it is almost certain that at least one false significant result will be found. The probability of not making a Type 1 error on the first 100 tests is using the binomial distribution calculated as $(1-\alpha)^m = 0.0059$, where m is the number of tests and α is the threshold value that states the level of significance. The shortcoming of testwise alpha is that it does not say anything about the probability of making at least one Type 1 error in a series of tests.

There are methods to be applied when evaluating the results of trials with multiple comparisons, e.g., the Bonferroni correction applies a family-wise alpha level, which is the testwise alpha scaled down by the number of hypotheses tested. Practically, this implies that lower p-values will be required to reject a null hypothesis and confirm a research hypothesis. Another way to put it is that if a hypothesis is tested after a search process, the "standard" p-values are deflated. Note that the problem of multiple hypotheses is general and while it will occur for accommodated evidence it could occur for predicted evidence too if same sample is used for multiple tests.[11]

White (2000) notes that a method controlling for multiple hypothesis: "...permits data snooping/mining to be undertaken with some degree of confidence that one will not mistake results that could have been generated by chance for genuinely good results." But far from everyone seems to agree. A standard objection to applying the family-wise alpha is that the general null hypothesis is all the null hypotheses are true, which is rarely of interest to testing in the first place. Westfall et al. (1997) notes: "Multiple testing is difficult and controversial on either side of the Bayes'/frequentist fence, with arguments over whether and how multiplicity adjustment should be performed."[12] Another objection is that multiplicity adjustments are too

[11] Another model evaluation approach is cross-validation. The idea is to use part of the data for specification search of a model, and the remaining part to test the performance of the chosen model. Hence, champions of this approach propagates that any claim of predictability should be backed up by out-of-sample performance statistics as any judgment about the predictability of one best model from a specification search will be overrated (e.g. Pickard and Cook, 1984).

[12] Westfall et al. discuss when and how to adjust prior assessments to account for

subjective; whether the correction should be made depending on the number of tests per article or for all tests considered in the scientists' lifetime, or whether each field should correct the number of tests or whether any corrections should be made at all − all of them questions that have puzzled many practicing statisticians. Hoover and Perez (2000) confirm the view that econometricians have highly different attitudes toward data mining, which range from it is to be avoided' to it is inevitable' or even to it is essential.' Similar disagreement over whether prediction constitutes stronger confirmation than accommodation among philosophers of science and statisticians is noted by Lipton (2005).

6 Conclusions

In this paper, I turn to folk intuitions to empirically investigate whether predicted evidence constitutes stronger confirmation than accommodated evidence. Two key findings have emerged. First, I find that predicted evidence constitutes stronger confirmation than accommodated evidence, and the effect is robust to the framing of the problem. Worth noticing, however, is that both, accommodated and predicted evidence, constitute confirmation. This is consistent with the history of science where both accommodated and predicted evidence seem to have had significant merits for the acceptance of theories. It is also notable that most subjects (49%) believe that a model that accommodates evidence is as good as one that has been shown to predict evidence. However, a substantial share (37%) of subjects believe that a model that has been shown to predict evidence is more likely to perform better in future trials. Only a small share of subjects (14%) had a higher trust in the model that accommodated evidence. Hence it seems that people perceive the risk of over-fitting the data as much more severe than under-fitting the data. Second, the findings suggests that trusting a model to predict correctly is intrinsically related to trust in the proposers' (i.e., the scientists') level of knowledge, and relatively more subjects are persuaded by proposers' abilities if they, the proposers, are successful in predicting compared to accommodating evidence. This confirms the conjecture in Kahn et al. (1996) and Lipton (2005) that evidential support is

multiplicity and specify conditions for which the resulting posterior probabilities roughly correspond to Bonferroni adjusted p-values. Berry and Hochberg (1999) also discuss Bayesian attitudes and methods for adjusting inferences for multiplicities.

linked to the proposer of the model. Notably, this link can be argued to be epistemologically unwarranted.

In conclusion, the findings suggest that a persuasive scientist is one who is able to predict and thereby convince others that she is knowledgeable. Granting oneself the flexibility to tell the story after having observed the evidence seems to be the less persuasive strategy, assuming that the scientist is honest about the methodological approach used. Admittedly, persuasiveness and confirmation cannot be the only objectives of the scientists, and for discovery accommodation and explorative research continue to remain important.

References

Berry, D. A., and Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*(1), 215-227.

Bem, D. J. (2003). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III (Eds.), *The complete academic: A career guide* (pp. 171-201). Washington, DC: American Psychological Association.

Cane, M. A., Zebiak, S. E. and Dolan, S. C. (1986). Experimental forecasts of El Nino. *Nature* 321, 827-832.

Fehr, E., Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 1143, 817-868.

Hitchcock, C and E. Sober. (2004). Prediction versus accommodation and the risk of overfitting. *Brit. J. Philos. Sci* 55, 1-34.

Hoover, K. D and Perez, S. J. (2000). Three attitudes towards data mining. *Journal of Economic Methodology* 7:2, 195-210.

Kahn, J.A., Landsburg, S.E., and Stockman, A.C. (1996). "The Positive Economics of Methodology," *Journal of Economic Theory*, 68(1), 64-76.

Leamer, E. (1983). Let's Take the Con Out of Econometrics. *American Economic Review*, 73, 31-43.

Lipton, P. (2005). Testing hypotheses: prediction and prejudice, *Science* 307, 219-221.

Maddala G.S. Introduction to Econometrics. (1992). 2nd ed., Macmillan.

Musgrave, A. (1974). Logical versus historical theories of confirmation, *Brit. J. Philos. Sci*, 22, 1-23.

Westfall, P.H., Johnson, W.O., Utts, J.M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* 84, 419-427.

White, H. (2000). A reality check for data snooping. *Econometrica* 68, 1067-1084.

Appendix

El Nino is characterized of abnormal warm ocean water temperatures that occasionally develops off the western coast of South America and can cause climatic changes across the Pacific Ocean. El Nino events happen irregularly. The El Nino phenomenon drastically affects the weather in many parts of the world. Developing countries which are dependent on agriculture and fishing, particularly those bordering the Pacific Ocean, are the most affected. If forecasts could provide warnings before an El Nino episode, human suffering and economic losses could be reduced. Please consider the partly fictional scenario below and answer the questions.

A team of scientists, henceforth called team A, have recently constructed a new hybrid model, where an ocean model is coupled to a statistical atmospheric model that accommodates 20 of the known El Nino events of the twentieth century (i.e., $1901-2000$). The model is constructed to fit observations that have already been made. Using old data (i.e., the known El Nino events), the model is rigorously tested and able to detect El Nino events 12 months before they start in 15 of the 20 cases without causing any false alarms. Without any knowledge, the chances to detect El Nino before it starts is only 5 percent. A model is considered to be good if it detects El Nino events 12 months before they start with a probability of 80 percent. The model has recently also been tested on how well it predicts future El Nino events. More specifically, after the model was developed, the El Nino event has occurred 5 times, and the model successfully predicted that event 12 months before it started in all of these 5 cases without causing any false alarms. In total, that would imply 20 correct predictions using old data and 5 correct predictions using new data, i.e., in total 20 correct predictions out of 25. Independent of team A's work, but knowing that the latest five El Nino events occurred, another group of scientists, henceforth called team B, developed a different hybrid model (i.e., where an ocean model is coupled to a statistical atmospheric model) which is constructed to accommodate the occurred events. The model is constructed to fit observations that have already been made. Testing the model on old data (including the 5 latest El Nino events), team B's model is able to detect El Nino events 12 months before they start in 20 of the 25 cases without causing any false alarms. Team B's model has never been tested on how well it predicts future El Nino events but has been rigorously tested using old data. The two teams disagree about which model is more correct and will have greater predictive power in the future. Please state a) which of the two teams you consider to be more knowledgeable based on the information you have been given and b) how likely you believe it is that the models will be correct in future predictions.