Hypothesis Space Checking in Intuitive Reasoning

Christopher D. Carroll (cdcarroll@gmail.com)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213

Charles Kemp (ckemp@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213

Abstract

The process of generating a new hypothesis often begins with the recognition that all of the hypotheses currently under consideration are wrong. While this sort of falsification is straightforward when the observations are incompatible with each of the hypotheses, an interesting situation arises when the observations are implausible under the hypotheses but not incompatible with them. We propose a formal account, inspired by statistical model checking, as an explanation for how people reason about these probabilistic falsifications. We contrast this account with approaches such as Bayesian inference that account for hypothesis comparison but do not explain how a reasoner might decide that the hypothesis space needs to be expanded.

Keywords: hypothesis testing; model checking; surprise

Introduction

Many modern scientific disciplines are characterized by strange and unintuitive theories that previous generations of scientists never would have imagined. On a less dramatic scale, people often generate inventive explanations in their everyday lives. The existence of these unintuitive theories and inventive explanations raises an interesting question: how are these new theories and explanations discovered?

In many cases, the process of generating a new hypothesis starts when the reasoner decides that all of the hypotheses currently under consideration are wrong. In some cases, the available evidence is incompatible with every hypothesis under consideration, and this decision is straightforward. In other cases, however, the available evidence is implausible under, but not strictly incompatible with, the hypotheses. In cases like these, a reasoner may engage in *hypothesis space checking* to decide whether the hypothesis space is adequate or needs to be expanded.

Although psychologists have explored many approaches to hypothesis testing, most of these approaches are unable to account for hypothesis space checking. Bayesian accounts, for instance, are able to specify the relative strength of a hypothesis within the hypothesis space, but they do not provide criteria for evaluating the hypothesis space itself.

Statisticians, however, have developed various measures that quantify the extent to which observations are surprising under a given hypothesis or hypothesis space. In this paper, we investigate the possibility that formal measures of this kind can help to explain how people decide that all of the hypotheses in their current hypothesis space are probably wrong.

Hypothesis space checking

Figure 1 illustrates the kind of situation where hypothesis space checking may be required. There is a universe *U* of possible explanations for the given observations, but the hypotheses available to the reasoner fall within a hypothesis space *H* that is a proper subset of *U*. It is possible, of course, that the true explanation is not in *H*; the ability to determine whether this is the case would be useful.

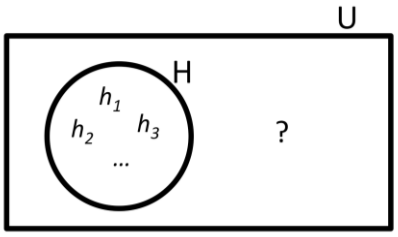

Figure 1: The universe *U* includes all possible hypotheses, and hypothesis *H* is the subset of these hypotheses that are currently available to the reasoner.

In principle, the adequacy of *H* could be evaluated by computing whether the available observations are better explained by hypotheses that lie within or outside *H*. Bayesian inference provides one way to formalize this sort of *comparative hypothesis testing*. Bayes' theorem establishes that given the observed data *d*, the odds that *H* contains the true explanation are:

$$\frac{P(H|d)}{P(\bar{H}|d)} = \frac{P(d|H)P(H)}{P(d|\bar{H})P(\bar{H})}. \quad (1)$$

Equation 1 shows that the probability that *H* contains the true explanation depends on (1) the relative probabilities of the data under *H* and under its complement $\bar{H}$ and (2) the relative prior probabilities of *H* and $\bar{H}$.

Although Equation 1 is appealing in principle, it is impossible to apply. Given that $\bar{H}$ consists of hypotheses that are unavailable to the reasoner, the term $P(d|\bar{H})$ will be impossible to compute (Earman, 1990, Ch. 7; Salmon, 1990). Consider the problem faced by a Newtonian physicist attempting to explain the anomalous precession of Mercury's perihelion. Although the physicist might be able to estimate $P(d|H)$ by considering various Newtonian explanations, estimating $P(d|\bar{H})$ has a paradoxical flavor:

how would the physicist compute probabilities with respect to theories he cannot currently imagine?

The paradox just described applies to any account (Bayesian or otherwise) that uses comparative hypothesis testing to address the problem defined by Figure 1. We therefore propose that this problem can only be addressed by *non-comparative accounts of hypothesis testing* in which the current hypothesis space is evaluated not in relation to specific competitors but on its own merits. In statistical practice, this sort of evaluation is often referred to as model checking or goodness-of-fit testing, and it typically involves comparing the actual observations to the expected distribution of the observations given the current hypothesis space. To the extent that the actual observations seem surprising in this context, there is an incentive to search for new hypotheses.

Comparative and non-comparative hypothesis testing seem to address distinct problems in that comparative hypothesis testing seems most useful for selecting among the hypotheses in *H* and non-comparative hypothesis testing seems most useful for checking *H* itself (for similar proposals, see Bayarri & Berger, 1999; Gelman & Shalizi, 2013; Gillies, 2007). We propose that both kinds of hypothesis testing are represented among people's intuitive inferences, but in this paper we deliberately focus on a situation that calls for non-comparative hypothesis testing.

A model of non-comparative hypothesis testing

We propose that intuitive hypothesis space checking resembles the process specified in Figure 2. Specifically, we propose that people extract the salient or important features of the available observations, assess the extent to which those individual features are surprising under *H*, and then compute a global measure of surprise. This global measure of surprise provides a criterion for deciding whether to initiate the search for new hypotheses.

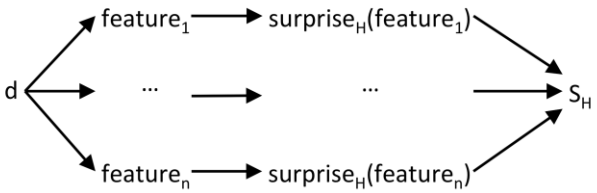

Figure 2: The reasoner extracts the salient features of the observations *d*, calculates a measure of surprise for each feature, and combines the surprise values into a global measure S_H that captures the extent to which the data are surprising given the current hypothesis space *H*.

Statisticians have proposed various measures of surprise (e.g., Bayarri & Berger, 1998; Weaver, 1948), but we focus on statistical null hypothesis testing, which is the best-known statistical procedure that can be used for hypothesis space checking. To investigate null hypothesis testing in the simplest possible setting, we focus on situations where the hypothesis space contains a single focal hypothesis h (i.e., where $H = \{h\}$), but various generalizations of our approach are applicable to composite hypothesis spaces (e.g., Bayarri & Berger, 1999; Gelman, Meng, & Stern, 1996). In null hypothesis testing, the statistician first defines a real-valued test statistic $T(d)$ that measures some property of the data; this test statistic can be viewed as one of the features in Figure 2. To evaluate the surprise of the observed value of the test statistic, the statistician then considers the expected distribution of $P(d^{\text{rep}})$ given h, where d^{rep} is a random variable representing the data that might be observed if one were to replicate the "experiment" that produced the data. By comparing the observed value of $T(d)$ to the expected distribution of $T(d^{\text{rep}})$ under h, the statistician can assess whether $T(d)$ is surprising under h. If the test statistic is defined such that greater values represent greater deviations from h, the surprise of $T(d)$ can be summarized by a *p-value*:

$$p_T(d) = P[T(d^{\text{rep}}) \geq T(d) \mid h]. \qquad (2)$$

Intuitively, the p-value represents the probability that the test statistic in the imagined replications would be at least as extreme as what was actually observed. Small p-values correspond to surprising results where the observations are unusually extreme.

In the final step of Figure 2, the reasoner combines the surprise measures for each feature into a global measure of surprise. To avoid making assumptions about how people integrate surprise ratings across different features, we focus on situations where there is a single surprising feature. In such situations, it seems reasonable to adopt the surprise value for that feature as the global measure of surprise.

Method

To evaluate our proposed model of non-comparative hypothesis testing, we conducted an experiment in which participants learned about the ancient burial sites found on a remote island chain. The burial sites were marked by "cairns" (rock piles), and each island had been occupied by one of two cultures that constructed the cairns using different procedures: the "Chaotics" placed a random number of boulders in each cairn and the "Numerologists" placed a number of boulders in accordance with a mathematical function. The instructions explained that the Numerologists used different mathematical functions on different islands but that the mathematical function was always based on the number of people buried at the site. The participants were asked to infer which cultural group occupied an island from information about the burial sites on the island.

Because the number of possible mathematical functions is infinite, we expected that participants would not be able to assess every possible explanation for the observations. We expected that when faced with this impossible task, participants would consider the hypothesis that the Chaotics occupied the island as well as various hypotheses where the

Numerologists occupied the island and used some simple mathematical function. Because the materials were designed so that no simple mathematical function would explain the observed number of boulders at the burial sites, we expected that most participants would end up with a hypothesis space that contained a single viable hypothesis: the hypothesis that the Chaotics occupied the island. We expected that participants would check this hypothesis through a procedure resembling the one depicted in Figure 2. We predicted that when the observations were sufficiently surprising, participants would be willing to attribute occupancy to the Numerologists. Critically, we expected that participants would sometimes make this attribution even when they could not identify a single mathematical function that the Numerologists might have used. As we discuss later, this finding would be difficult to explain as a consequence of comparative hypothesis testing.

The experimental materials were based on three "test statistics" that reflected the salient numerical concepts of equality and magnitude (see Table 1). The equality test statistic, for example, was defined as the count of the burial sites that had the same number of people and boulders, and we expected participants to be surprised when many of the burial sites had the same number of people and boulders.

Table 1: Test statistics

Name	Definition
Equality	number of burial sites where $b = p$
Minimum	smallest observed value of b
Repetition	frequency count for the most frequent b

Note. p = the number of people at a burial site; b = the number of boulders at a burial site.

Participants

Sixty-one undergraduates participated in the experiment for course credit.

Materials

Table 2 displays the observations presented to the participants. In the table and in the rest of the paper, we represent burial sites by two numbers separated by a dash, with the first and second numbers representing the number of people buried and boulders, respectively. Each row of the table corresponds to a different island. Twelve of the islands were designed to be surprising according to exactly one of the test statistics in Table 1 and four additional islands were designed to have no surprising features (the "None" islands). All of the islands contained either three or six burial sites, and the surprising islands were designed so that the coincidence involving the test statistic would be either moderately ($.01 < p < .15$) or highly ($p < .01$) surprising, as calculated from Equation 2.

The rightmost column shows the p-values for each island; these p-values summarize how surprising the observations would be if the Chaotics occupied the island. The p-values were calculated under the assumption that the number of boulders at a burial site could range from 1 to 100 (the instructions informed participants that this was the case). To illustrate, consider the calculation of the p-value for the first island. The observed value of the equality test statistic for this island was one: there was exactly one burial site that had the same number of people and boulders. If the Chaotics occupied the island, then the equality test statistic would follow a binomial distribution with a probability parameter of .01. Consequently, the probability that at least one of the burial sites on a three-site island has the same number of people and boulders is approximately .0297.

For the equality and minimum statistics, the four islands represented the four possible combinations of surprise condition and island size. For the repetition test statistic, we did not include a high-surprise island with three burial sites; instead, we included two moderate surprise islands with three burial sites. The reason for this was that creating a high-surprise "repetition" island with three burial sites necessitated selecting an island where each burial site had the same number of boulders. Because we were interested in situations where the participants would not be able to find a mathematical function to explain the observations, we chose not to present such an island.

Table 2: Experimental materials

Feature	Srprs.	Sz.	Burial sites	*p-value*
Equality	M	3	20-94, 39-39, 85-78	.0297
Equality	H	3	16-16, 65-65, 49-12	.0003
Equality	M	6	7-62, 33-85, 40-1, 53-26, 59-59, 94-18	.0585
Equality	H	6	12-100, 19-42, 21-21, 32-14, 75-75, 93-56	.0015
Minimum	M	3	15-86, 63-98, 84-75	.0176
Minimum	H	3	16-92, 42-97, 93-90	.0013
Minimum	M	6	5-67, 24-81, 35-72, 52-68, 57-93, 83-54	.0108
Minimum	H	6	13-75, 32-95, 35-98, 37-80, 72-85, 96-94	.0003
Repetition	M	3	3-19, 27-84, 74-19	.0299
Repetition	M	3	11-75, 39-28, 80-75	.0299
Repetition	M	6	2-5, 6-97, 31-69, 59-38, 62-52, 75-52	.1404
Repetition	H	6	12-98, 15-98, 26-4, 45-73, 60-53, 77-98	.0020
None	-	3	23-18, 40-69, 93-55	-
None	-	3	31-46, 80-24, 94-87	-
None	-	6	1-78, 43-61, 45-12, 52-35, 83-87, 91-46	-
None	-	6	1-26, 8-92, 14-36, 35-20, 40-11, 63-45	-

Note. Srprs. = surprise condition; Sz. = island size; M = moderate surprise; H = high surprise.

All of the observations were designed so that at most one of the test statistics in Table 1 would be surprising at a level greater than $p = .30$. In addition, we controlled for the distribution of even and odd numbers and for the correlation between the number of people and number of boulders.

Procedure

Participants were provided with a cover story that described their task and the Chaotics and Numerologists. Participants then completed a familiarization trial. On both the familiarization and experimental trials, the burial sites were represented by "cards" on a computerized display. Each card listed one number next to a stick figure and another number next to an illustration of a boulder pile. These numbers represented the number of people buried at the site and the number of boulders in the cairn, respectively. Participants were encouraged to re-arrange the cards by clicking and dragging them. The interface also provided buttons to automatically sort the cards according to either the number of people buried or the number of boulders. For the practice trial, the three burial sites were 31-1, 48-5, and 90-4, and participants were told that the island was occupied by Numerologists who placed a number of boulders equal to the number of prime factors of the number of people buried at the site (e.g., because 48 = 3 * 2 * 2 * 2 * 2, the burial site with 48 people had 5 boulders). This rule was intended to establish that the Numerologists were sophisticated mathematicians who had access to a wide variety of mathematical properties and rules. In doing so, our goal was to establish a universe of possible explanations that would be too large to consider in full.

Participants reported their inferences about which cultural group had occupied the island using a seven-point rating scale where the leftmost point was labeled "definitely Chaotics", the rightmost point was labeled "definitely Numerologists", and the middle point was labeled "not sure". Responses were coded from -3 ("definitely Chaotics") to 3 ("definitely Numerologists"). Participants who indicated that the Numerologists were more likely to have occupied the island than the Chaotics were also asked to indicate whether they had "discovered ANY function that the Numerologists might have used to determine the number of boulders." Participants answering affirmatively were asked to describe the function. Finally, at the end of each trial, participants were asked to list "any features, coincidences, or patterns in the burial sites that would have been surprising if the Chaotics occupied the island." Participants were provided with three text input fields and could identify up to three features, coincidences, or patterns. The responses to this prompt were intended to measure whether participants noticed the relevant features or any other features of the observations.

After completing the familiarization trial, the participants completed experimental trials for each of the 16 islands listed in Table 2. The presentation order was randomized.

Results

A preliminary analysis confirmed that participants frequently noticed the relevant features. For each feature, a majority of the participants listed the feature as surprising on at least one of the relevant trials; the proportions were .59 for the minimum feature, .72 for the equality feature and .66 for the repetition feature. A second preliminary analysis confirmed that the islands did not contain many surprising features other than the intended relevant features. Participants listed other features on only 16.8% of the experimental trials. The proportion of participants listing other features was similar across the experimental conditions: a logistic regression with categorical predictors corresponding to the surprise conditions, the relevant features, and the island sizes did not explain a significant proportion of the variance in the probability that participants noticed other features, $R^2 = .26$, $F(5, 8) = 0.55$, $p = .74$.

To evaluate our formal approach we compared the model-derived p-values and the mean culture ratings. Because people often evaluate probabilities on a logarithmic scale (e.g., Gonzalez & Wu, 1999), we adopted the logit (i.e., the log-odds) of the island p-values as the model's measure of surprise (lesser values corresponded to greater surprise). When calculating the mean culture ratings for each condition, we excluded any culture ratings for which the participant who provided the rating failed to identify the relevant feature as surprising at any point during the experiment. The rationale for this exclusion is that a participant who did not notice the relevant feature could not have been surprised by it.

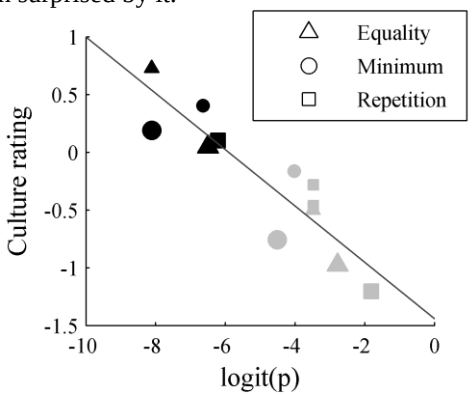

Figure 3: The culture rating as a function of the logit of the p-value. Different features are represented by different marker shapes, different island sizes are represented by different marker sizes (large markers correspond to islands with six burial sites), and different surprise conditions are represented by different shadings (black markers correspond to high-surprise islands).

Figure 3 shows the comparison of the p-values and the mean culture ratings. A linear regression confirmed that the logit of the p-values explained a significant proportion of the variance in the ratings, $R^2 = .82$, $F(1, 10) = 44.2$, $p < .001$. This relationship was essentially unchanged even when culture ratings were included for participants who failed to notice the relevant features, $R^2 = .77$, $F(1, 10) = 33.83$, $p < .001$. Inspection of Figure 3 also suggests that the islands with six burial sites may have been viewed as less surprising than the islands with three burial sites. The statistical significance of this finding was confirmed by a multistep regression that showed that island size predicts variance in the culture ratings above and beyond the variance explained by the logit of the p-values, $\Delta R^2 = .14$,

$F(1, 9) = 25.7$, $p = .001$. This finding may reflect a general tendency to underestimate the extent to which deviations from the mean become increasingly surprising for larger samples (Kahneman & Tversky, 1972).

Although our participants compared hypotheses in the sense that they reported whether the Chaotics or Numerologists occupied an island, it seems difficult to explain their inferences as the product of what we have called comparative hypothesis testing. Consider, for example, the difficulties that arise in explaining the culture ratings by appealing to Equation 1, which in the context of our experiment involves the comparison of $P(d|\text{Chaotics})$ and $P(d|\text{Numerologists})$. Note that $P(d|\text{Chaotics})$ depends only on the number of burial sites on the island: for any island with three burial sites, for example, $P(d|\text{Chaotics})$ is $(1/100)^3$. Thus, if the participants' inferences were indeed based on Equation 1, then the differences in the culture ratings must have arisen primarily because of differences in $P(d|\text{Numerologists})$.

If $P(f|\text{Numerologists})$ is a prior distribution over specific functions f, then

$$P(d|\text{Numerologists}) = \int_f (d|f)P(f|\text{Numerologists}) \quad (3)$$

The integral in Equation 3 will be large to the extent that there are many functions that are plausible *a priori* ($P(f|\text{Numerologists})$ is high) and consistent with the data ($P(d|f) > 0$). Approximating this integral using sampling or some other standard method would involve identifying one or more functions f for which $P(d|f) > 0$. Our participants, however, rarely identified even a single function f for which $P(d|f) > 0$. Recall that participants who claimed that the Numerologists occupied an island were asked whether they had found any mathematical function to explain the observations. Participants reported finding a function on only 15.5% of these occasions. Furthermore, the "functions" that these participants reported were often not fully specified functions at all. One representative participant claimed to have found a function but then wrote that "I don't have a function, but when put roughly on a graph it almost-kinda-sorta forms a wave." Summarizing his inference, the same participant later added, "I'm grasping at straws though." Figure 4, furthermore, shows that the relationship between function finding and the culture ratings is weak and, according to a linear regression, non-significant, $R^2 = .016$, $F(1, 12) = .19$, $p = .67$.

Could participants have estimated $P(d|\text{Numerologists})$ without identifying a single specific function f that might have been used by the Numerologists? Might participants, for example, have used some computational procedure that approximates the integral in Equation 3 without actually identifying any specific functions? We cannot exclude this possibility, but we do not know of any such procedure. In the absence of a known procedure that approximates the integral in Equation 3 given some plausible specification of the prior, it seems reasonable to conclude that our participants did not rely on comparative hypothesis testing.

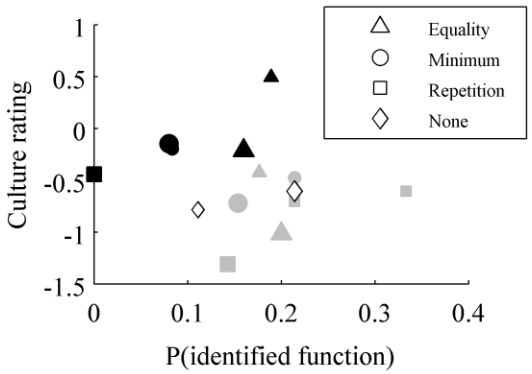

Figure 4: The mean culture ratings as a function of the proportion of participants who claimed to have identified a mathematical function that explained the observations.

As a final test of our model, we investigated whether the model-derived surprise predicted the culture ratings even after excluding trials in which participants claimed to have found a mathematical function. To do so, we recalculated the mean culture ratings while excluding any culture rating where either (1) the participant reported finding a mathematical function or (2) the participant never noticed the relevant feature (as in previous analyses). A linear regression on these recalculated culture ratings confirmed that the logit of the p-values remained strongly predictive of the culture ratings, $R^2 = .87$, $F(1, 10) = 69.43$, $p < .001$.

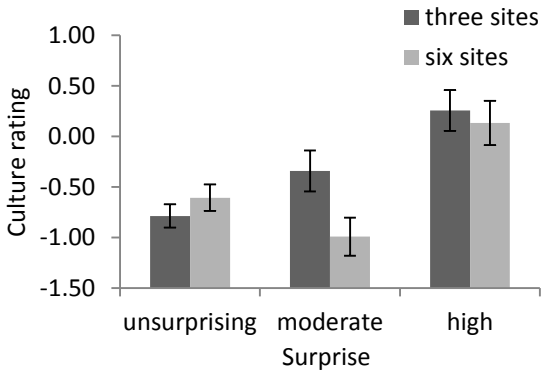

Figure 5: Culture ratings as a function of surprise condition and the number of burial sites. The error bars show standard errors.

Our analyses so far have focused on the islands that were designed to be of "moderate" or "high" surprise. We compared these islands to the unsurprising "None" islands by analyzing the mean culture ratings as a function of surprise condition. Figure 5 shows that participants were much more willing to attribute island occupancy to the Numerologists in the high-surprise condition. The similarity between the culture ratings for the unsurprising and moderately-surprising condition was not expected, but it may be that the culture ratings are only influenced by observations once the surprise exceeds a certain threshold. Figure 5 also suggests that island size might have influenced

the culture ratings, either on its own or in an interaction with the surprise condition. A within-subjects ANOVA showed that the culture ratings were influenced by both surprise condition, $F(2, 100) = 27.80$, $p < .001$, and island size, $F(1, 50) = 6.89$, $p = .01$; the interaction between surprise condition and island size was marginally significant, $F(2, 100) = 2.89$, $p = .06$. In post-hoc analyses, we confirmed that the culture ratings in the unsurprising and moderately-surprising conditions were not significantly different, $t(50) = .80$, $p = .43$, and that the culture ratings in the high-surprise condition were significantly different from those in the unsurprising, $t(50) = 5.66$, $p < .001$, and moderately-surprising, $t(50) = 6.74$, $p < .001$, conditions.

Discussion

The experimental findings suggest that people perform hypothesis space checking using an intuitive version of non-comparative hypothesis testing. The findings are not naturally explained by comparative hypothesis testing. This is not to say that comparative hypothesis testing is never useful: recall that our experiment was deliberately designed so that comparative hypothesis testing would be of limited relevance, and comparative hypothesis testing undoubtedly plays an important role in other settings. Moreover, although comparative hypothesis testing cannot explain our main experimental findings, there are reasons to believe that it influenced our participants' thinking to some extent. Participants were often reluctant to fully commit to the idea that the Numerologists occupied the island even after observing very surprising observations: even in the most surprising condition ($p \approx .0001$), the mean culture rating was only 0.73. One interpretation of this finding is that people are often unwilling to fully reject a hypothesis space until a better explanation is discovered (see also Griffiths & Tenenbaum, 2007).

Other researchers have proposed that people employ methods such as sampling to approximate Bayesian inference in situations where it is impossible for them to evaluate the entire hypothesis space (e.g., Sanborn, Griffiths, & Navarro, 2010). These methods, however, do not address the problem posed by Figure 1. Sampling from *H* may be useful if this hypothesis space is large, but this approach does not explain how a reasoner might decide that the true hypothesis lies outside *H*. Supporters of sampling might respond that the problem of hypothesis space checking never arises because the space of available hypotheses is always equivalent to *U*. This position, however, seems incompatible with the intuition that scientists and others are sometimes able to generate hypotheses and explanations that are genuinely new.

The justifications for comparative and non-comparative testing remain controversial among statisticians and philosophers (e.g., Howson & Urbach, 1989/1996; Mayo, 1996; see also Gigerenzer et al., 1990, Chapter 3), but both kinds of hypothesis testing seem necessary to account for the inferences that people make. Non-comparative hypothesis testing is especially notable for the role it plays in the discovery of new hypotheses. These discovery processes, while often mysterious and difficult to explain, are involved in many of the most interesting inferences that people make. Statistical model checking does not explain where new hypotheses come from, but it can explain why people initiate the search for new hypotheses.

Acknowledgements

This work was supported in part by NSF grant CDI-0835797 and by the Pittsburgh Life Sciences Greenhouse Opportunity Fund. We thank Angela McCarthy for assistance with data collection.

References

Bayarri, M. J., & Berger, J. O. (1999). Quantifying surprise in the data and model verification. *Bayesian statistics, 6*, 53-82.

Earman, J. (1992). *Bayes or bust?: A critical examination of Bayesian confirmation theory.* MIT Press.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 8-38.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733-759.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1990). *The empire of chance: How probability changed science and everyday life.* Cambridge University Press.

Gillies, D. (2001). Bayesianism and the fixity of the theoretical framework. In D. Corfield, & J. Williamson (Eds.), *Foundations of Bayesianism*. Dordrecht: Kluwer.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology, 38 (1)*, 129-166.

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition, 103*, 180-226.

Howson, C., & Urbach, P. (1996). *Scientific reasoning: The Bayesian approach* (3rd ed.). Open Court Publishing Co. (Original work published 1989).

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

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

Salmon, W. (1990). Rationality and objectivity in science or Tom Kuhn meets Tom Bayes. In C. W. Savage (Ed.), *Scientific Theories*. University of Minnesota Press.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117 (4)*, 1144-1167.

Weaver, W. (1948). Probability, rarity, interest, and surprise. *The Scientific Monthly, 67 (6)*, 390-392.