Heuristic Biases in Statistical Thinking

Andrew Neath*

Abstract

Our minds rely on heurisitic thinking and intuition, often with much success. However, a major bias in our heuristic thinking stems from an inability to properly account for the role that randomness plays in the world. It should be expected that formal statistical training would lead to a scientific approach free of such bias. I will argue that the traditional approach to teaching introductory statistics is, in fact, promoting methods based on the same illusions that corrupt our heuristic thinking. A consequence of this approach to teaching statistics is an issue currently faced in science where an unacceptably large proportion of "statistically" established findings fail upon attempts at replication.

Key Words: statistics education, scientific reasoning, fallacy of the transposed conditional, representativeness, base rates

1. Introduction

In this note, I will argue how the traditional approach to statistical inference is logically flawed, performs poorly in practice, and promotes bad science. Once one believes this premise, a natural question arises: Why do statistically trained educators continue to support such methods? I propose the simple reason is that our minds have fooled us into believing that the traditional justifications are perfectly correct. I will refer to the Nobel Prize winning work of Daniel Kahneman and Amos Tversky on heuristic biases in our thinking to show that even well trained statisticians can be led to support incorrect statistical interpretations.

2. Significance testing framework

Significance testing is the most commonly taught approach to a statistical decision problem. Within the significance testing framework, one is to decide between a null hypothesis (representing no effect, no difference, status quo, etc.) and an alternative hypothesis (representing a new finding). The null hypothesis is believed initially. We move off this belief if the observed data is unusual / unlikely under the assumption of the null hypothesis. In this case, we are taught to *reject the null hypothesis*. If the observed data is not unusual / unlikely, we are not taught to accept the null hypothesis, but rather to merely *fail to reject the null hypothesis*. The softer language for the decision in favor of the null H_o is intentional.

The justifications for the language used in interpreting a significance testing result mimic those of a mathematical proof by contrapositive. Let proposition A be the statement that the null H_o is true, and let proposition B be the statement that the observed data is unusual. The idea behind significance testing is based on the conditional sentence $A \Rightarrow B'$. That is, if we assume the null hypothesis is true, it follows that the data we observe should be compatible with this assumption. The contrapositive statement $B \Rightarrow A'$ is logically equivalent to the original conditional sentence. That is, the observance of unusual data under the assumption of the null hypothesis (B) seems to provide a logical justification to *reject* the null hypothesis (A'). The converse statement $B' \Rightarrow A$ is *not* logically

*Southern Illinois University Edwardsville, Department of Mathematics and Statistics, Edwardsville, IL 62026, aneath@siue.edu

equivalent to the original conditional sentence. When data is observed that is not unusual under the assumption of the null hypothesis (B'), there is not the same logical justification for concluding the null hypothesis. Thus, the softer language of *fail to reject* is taught to statistics students.

To further illustrate how significance testing is presented, consider the following excerpts. Bold type is added for emphasis. Walpole et al. (2012) write "Rejection of a hypothesis tends to **all but rule out the hypothesis**. On the other hand, it is very important to emphasize that the failure to reject does not rule out other possibilities. As a result, the **firm conclusion is established by the data analyst when a hypothesis is rejected**." Devore and Peck (2012) write how rejection of the null hypothesis would represent "**compelling evidence.**" However, "rather than make a decision to accept the null hypothesis, we avoid the conclusion that the null hypothesis is true." Sullivan (2013) writes "When the null hypothesis is rejected, there is **sufficient evidence to conclude that the alternative hypothesis is true**." However, "we never say that the null hypothesis is true." Moore, McCabe, and Craig (2012) interpret a statistically significant result as "**We conclude that the null hypothesis must be false**."

The word *reject* itself conjures the image of a strong conclusion. A definition of reject by Merriam-Webster is **to declare not to be true**. Synonyms of reject include the equally strong words **contradict, disavow, disclaim, refute**. It seems clear the interpretation presented to statistics students is that rejection of a null hypothesis is a decision made with a high degree of belief, while acceptance of a null hypothesis is a decision with a low enough degree of belief to prevent a strong conclusion. Many textbooks will refer to a courtroom analogy where a decision to find a defendant guilty, made only when the degree of belief is *beyond all reasonable doubt*, is similar to a decision to reject the null hypothesis in a significance testing problem.

The logical arguments for justifying a degree of belief interpretation to a significance testing result ignored a formal accounting of the uncertainty inherent to a statistical decision. As we shall see in the next section, when we do formally account for uncertainty, the logical arguments break down completely and significance testing is seen for what it really is: a severely flawed approach to science and decision making.

3. Fallacy of the transposed conditional

We must use probability to correctly assess the measure of belief in a statistical decision problem. As in Section 2, let A be the event that H_o is true, and let B be the event that the observed data is unusual / unlikely, under the null assumption. Significance testing is based on the conditional probability $P(B \mid A)$ being small. For a level $\alpha = .05$ test, $P(B \mid A) = .05$. When unusual data is observed (event B), we reject the null hypothesis (reject A). Recall that statistics students are taught that rejecting a null hypothesis is a firm conclusion, represents compelling and sufficient evidence, is a decision with a high degree of belief as in a courtroom setting, etc. In a formal accounting of uncertainty, a high degree of belief against the null would be measured by a small conditional probability $P(A \mid B)$. But the claim that $P(A \mid B)$ should be small follows only from the fact that $P(B \mid A)$ is small. This faulty reasoning has been called *the prosecutor's fallacy, the inverse fallacy,* or *the fallacy of the transposed conditional*. (See, for example, Ziliak and McCloskey, 2008).

It is very easy to develop an example showing how a significance testing interpretation can go horribly wrong. Consider a patient undergoing a diagnostic test for a serious disease. Let A be the null event that the patient does not have the disease. Let B be the event that the patient tests positive for the disease. Tests with good diagnostic properties are unlikely to result in a positive, conditional on the patient being free from the disease. For ease of

demonstration, take $P(B \mid A) = .05$, so that the diagnostic testing problem matches the significance testing problem perfectly. A positive test result (event B) would constitute an unusual / unlikely outcome, if the patient were free of the disease (event A). Suppose the patient does test positive for the disease. According to the interpretation we present to statistics students, the null hypothesis of no disease is rejected at the 5% significance level. and we seemingly have a high degree of belief that the patient does indeed have the disease. (Recall the interpretations presented to statistics students.) It takes only an elementary knowledge of probability theory to see that this significance testing conclusion is not valid. The belief in the event that the patient is diseased, given the positive diagnostic test, is the conditional probability $P(A \mid B)$. In the problem, we know only $P(B \mid A)$. A simple application of Bayes Theorem reveals that we need $P(B \mid A')$, a property of the test on diseased patients, and $P(A)$, the prevalence of the disease in the testing population. It is unsettling that a problem in an introductory textbook that would be vague and incomplete in the sections on probability would yield an unambiguous and strongly worded conclusion in the sections on significance testing.

A proper evaluation of belief in a significance testing result favoring the alternative is found by computing the probability that H_o is true (event A), conditional on observation of unusual data (event B). It is easy to see that

$$\begin{aligned} P(A \mid B) &= \frac{P(A)\,P(B \mid A)}{P(A)\,P(B \mid A) + P(A')\,P(B \mid A')} \\ &= \frac{\theta\,\alpha}{\theta\,\alpha + (1-\theta)\,(1-\beta)}, \end{aligned}$$

where $\theta = P(A)$, β is the probability of a type II error, and α is the probability of a type I error. Whereas α and β are familiar concepts in hypothesis testing, the importance of the role that θ plays is too often neglected. It is helpful to think of θ as representing the scientific context of the experiment. As an illustration, consider a drug in an early stage of testing. Drug development is prolonged and demanding. Only a small proportion of all drugs which begin testing are ultimately deemed effective. Let's take $\theta = .99$ to represent the context of the experiment. Since the drug is in the early stage of testing, there is a high probability that it will not be effective. For ease of demonstration, suppose a test for effectiveness is carried out at level $\alpha = .05$ with power $1 - \beta = .80$. In cases where the hypothesis test results in a decision to *reject the null hypothesis* of no effect, we have taught our students that such decisions embody a strong conclusion, analogous to a finding of guilty in a criminal trial. However, a simple calculation reveals $P(A \mid B) = .86$, a fairly high probability that the null hypothesis is true. A correct interpretation of the result is in direct conflict with the interpretation we are teaching to statistics students.

If more data in support of an effect becomes available, the belief in a positive finding will naturally increase. Scientific context can be thought of as representing the current state of information. The determination of this value is not without debate; experts may reasonably disagree on how to best assess available information. But the role scientific context plays in a proper quantification of belief is essential to good scientific practice. For example, a test for the effect of smoking during pregnancy on birth weight (Oster et al., 1988) is presented to students the same as a test for the effect of cereal eating during pregnancy on the baby's sex (Mathews et al., 2008). In both of these applications, a statistically significant result was observed. An appeal to scientific context would establish that the degree of belief in these positive research findings should be quite different. Goodman (1999) writes on how failure to address issues of scientific context has hurt the interplay between scientists and statisticians: "Such features as biological plausibility, the cogency of the theory tested, and the strength of previous results all become mere side issues of unclear rele-

vance. The methods of statistical inference in current use have contributed to a widespread misperception. This has damaged the quality of scientific reasoning and discourse."

The probability calculation demonstrates the need for a drug to pass multiple stages of testing before approval, an idea that is well understood and supported by statistics professionals. The need for replication of a positive finding before reaching a strong degree of belief holds true in any scientific inquiry. Yet, this idea is often lost in the teaching and practice of significance testing. We are pushing an interpretation that any statistically significant result is one with a high degree of belief. The consequence of this approach is alarm over the large number of statistically significant results that fail upon attempts at replication. In 2011, Bayer Healthcare reported that its scientists could not replicate 75% of positive findings in cardiovascular disease, cancer, and women's health, and in 2010, Amgen reported that its scientists could not replicate 47 out of 53 findings in landmark papers (Begley and Ellis, 2012).

Statistic educators are encouraging a definitive conclusion to a significance testing problem when we should be teaching students to take a healthy skepticism toward any single positive research finding. In the next section, I look for an explanation to why statistically trained educators continue to promote ideas which are in direct conflict with formal measures of belief.

4. System 1 / System 2 thinking

Psychologists and behavioral economists theorize two modes of human reasoning, dubbed System 1 and System 2. System 1 is known as fast thinking, and is extremely useful in solving familiar problems quickly and with little effort. Consider a very simple multiplication problem, say 2×6. The answer comes to us immediately, without effort. We did not have to think hard about what is meant by multiplication, nor did we have to work hard in performing the calculation. The problem is familiar, so we recognized the answer immediately. We may idealize System 1 as problem solving through heuristics or intuition. Heuristic thinking is incredibly valuable and useful in solving problems far beyond the simple multiplication problem used as motivation. As experts in mathematics and statistics, we are able to look at problems in our field and quickly determine a mode of attack. Intuition is not the same as instinct. Heuristic thinking is a learned skill, often requiring years of practice to tune right. When students ask "How did you know so quickly that was the best way to answer the problem?", the best response we can give is to put in the time and hard work to sharpen ones (heuristic) thinking.

System 2 is known as slow or deep thinking, and requires considerable effort. System 2 will be called upon to solve a hard math problem, or to check the validity of a complex logical argument. Consider a more difficult multiplication problem, say 234×512. We are all capable of solving this problem, perhaps in our heads without pencil and paper, but the answer arrives only after some effort. Because of the effort involved, our minds try to avoid calling on System 2. Cognitive scientists explain the process as System 2 staying in the background, serving as a check on System 1. System 2 will only step in as needed. However, as powerful as our heuristic minds can be, there are cases where System 1 can fool System 2 in staying away, leading to poor judgement and poor decision making. These heuristic biases often appear in problems which involve specifying a degree of belief. It was our inherent heuristic biases which may have prompted the quote (attributed to Persi Diaconnis) "Our minds are just not wired to do probability problems very well."

4.1 Plausibility is not probability

Let's make a connection between some common heuristic biases and the misinterpretation of a significance testing result. We will use Kahneman (2011) as our primary reference. We start with a simple experiment. Tversky and Kahneman (1983) made up the Linda problem to study heuristic errors in judgement. Linda is described as follows.

> Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

Participants are presented with eight scenarios and asked to rank from least to most probable.

(a) Linda is a teacher in elementary school.
(b) Linda works in a bookstore and takes yoga classes.
(c) Linda is active in the feminist movement.
(d) Linda is a psychiatric social worker.
(e) Linda is a member of the League of Women Voters.
(f) Linda is a bank teller.
(g) Linda is an insurance salesperson.
(h) Linda is a bank teller and is active in the feminist movement.

Although the experiment shows its age a bit, the essence of the description is clear. Linda is a good fit for a social worker, a bookstore or the feminist movement, but a poor fit for a bank teller or an insurance salesperson. Thus, it may seem that the description of Linda as a bank teller is only probable if she is also active in the feminist movement. The twist, of course, is that (f) must be more probable than (h), since event (h) is a subset of event (f). Nevertheless, 85% of respondents (doctoral students in the Stanford Graduate School of Business) ranked "feminist bank teller" as more likely than "bank teller". Kahneman (2011) explains this type of reasoning as confusing plausibility, or representativeness, with probability. System 1 hooks onto a plausible story as a quick way to aid understanding. But, as the Linda experiment illustrates, the most coherent stories are not necessarily the most probable.

I am not arguing that as large a number of professional statisticians will fall for the trick in the Linda problem. Nonetheless, the example does serve to illustrate how our minds can misjudge a measure of belief. Recall how easy it is to grasp onto the misinterpretation in significance testing. Many textbooks present the "proof by contrapositive" justification without noticing that this line of reasoning involves a probabilistic fallacy. Expanding on the Linda experiment, Kahneman (2011) writes "When people believe a conclusion is true, they are also very likely to believe arguments that appear to support it, even when these arguments are unsound." We are first presented with the interpretation to significance testing as students, with no need to question the validity of the argument. Because we believe the interpretations we were taught, we have an aspiration to believe the same justification as we pass it on to our students. When presenting the problem in a different context, say as a diagnostic testing problem, there is no trouble seeing the error in reasoning. Still, many statistics educators and textbook writers buy into a probabilistic fallacy when presenting a significance testing problem.

4.2 WYSIATI

Let's consider another form of heuristic bias. Here is an example from Gelman and Nolan (2002).

> A study of new diagnoses of kidney cancer in the 3141 counties of the United States reveals a remarkable pattern. The counties in which the incidence of kidney cancer is lowest are mostly rural, sparsely populated, and located in traditionally Republican states in the Midwest, the South, and the West. What do you make of this?

Wainer and Zwerling (2006) commented "It is both easy and tempting to infer that their low cancer rates are directly due to the clean living of the rural lifestyle - no air pollution, no water pollution, access to fresh food without additives." The trick, however, is that the counties with the highest rates of kidney cancer follow the same description. Wainer and Zwerling (2006) again point out how easy it would be to infer that "their high cancer rates might be directly due to the poverty of the rural lifestyle - no access to good medical care, a high-fat diet, and too much alcohol, too much tobacco." Of course the underlying factor is not the rural lifestyle, but the size of the county. Smaller counties exhibit greater variability, so by chance alone will have both the lowest rates and highest rates of kidney cancer. Nevertheless, it is easy for our minds to jump to a believable, yet false, conclusion.

System 1 is insensitive to the quality and quantity of information used in its heuristic reasoning. This heuristic bias is named WYSIATI - what you see is all there is. Kahneman (2011) writes "The measure of success for System 1 is the coherence of the story it manages to create. The amount and quality of information on which the story is based are largely irrelevant." As an example, sports fans are particularly susceptible to the WYSIATI bias, reacting to their teams latest success (or failure) as an overly strong indication of future performance, ignoring other relevant information.

Again, I do not necessarily believe a large number of professional statisticians will fall for the trick in this example. However, recall how the interpretation to a significance testing problem is presented as *rejecting* the null hypothesis, representing a *firm conclusion, compelling evidence, a high degree of belief*, etc. WYSIATI leads to a bias of confidence over doubt, preventing a proper consideration of the uncertainty inherent to the results from a single experiment. The overconfidence in the conclusion becomes particularly apparent after seeing how often testing problems in practice fail to yield the correct result. Ioannides (2005) writes further on this issue: "We should acknowledge that significance testing in the report of a single study gives only a partial picture. What matters is the totality of evidence. The high rate of nonreplication is a consequence of the ill-founded strategy of claiming conclusive research findings solely on the basis of a single study." Improperly teaching statistics students that a significance test can be used to establish conclusive evidence is a failing, possibly due to the WYSIATI heuristic bias.

4.3 Base rates

The commonality between heuristic biases and misinterpreting significance test results is the mistake of judging probability by representativeness. Here is a classic example (Kahneman, Tversky 1973).

> Tom W. is a graduate student at the main university in your state. Please rank the following nine fields of graduate specialization in order of the probability that Tom W. is a student in each of these fields.

business administration
computer science
engineering
humanities and education
law
medicine
library science
physical and life sciences
social science and social work

Absent any other information about Tom W., the ranking should reflect the relative sizes of the enrollments in the different fields. Kahneman (2011) refers to this as a *base rate*.

Next, we are presented with a personality sketch of Tom W.

> Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people, and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.

After reading the personality sketch, participants are asked to revise their ranking. The description of Tom W. is not a good fit with the fields of humanities and education, and social science and social work ("little feel and little sympathy for other people"). The fields computer science (a hint of nerdiness) and engineering ("neat and tidy systems") provide a better fit to the description. The trick here is that the personality sketch was created as representative of fields with a relatively small number of students, and a poor fit for fields with the most students. Nevertheless, participants ignored these base rates, even after initially being prompted to keep them in mind, in creating a final probability ranking. Despite a low base rate, computer science and engineering were rated most probable for Tom W.'s field of study.

Our minds mistake plausibility / representativeness and probability. In the Tom W. example, the heuristic bias occurs from ignoring the role that base rates play. In a hypothesis testing problem, the base rate is analogous to what we referred to earlier as the scientific context. An early stage test for a drug's effectiveness should be interpreted differently than a confirmatory test for a drug whose properties have been well studied. A positive result on a diagnostic test from a low prevalence population should be interpreted differently than a positive result from a population with higher prevalence. Kahneman and Tversky (1973) define representativeness "as the degree to which an event is similar in essential characteristics to its parent population and reflects the salient features of the process by which it is generated." Decisions from significance testing are based on the representativeness of the observed data with respect to the null hypothesis. Base rates are not considered in the interpretation of a significance test; no attempt is made to account for the role that context plays in scientific learning. The heuristic bias exhibited by participants in the Tom W. experiment parallels the fallacy of the transposed conditional. In both cases, a judgement of probability is falsely substituted with a judgement of representativeness. Kahneman (2011) writes "Anyone who ignores base rates in probability assessments will certainly make mistakes." The large number of statistically significant test results later found to be in error testifies to the validity of this statement.

5. Concluding Remarks

As it is traditionally taught, introductory statistics is a mathematics course, not a course in scientific reasoning. There is no ambiguity; all students are expected to reach the same conclusion. The core scientific principle of replication is not discussed; a significance test result is presented as one that is already decisive. Students are presented with interpretations that do not address the role of scientific context; a decision between hypotheses is treated automatically with no attention paid to the underlying science. The importance of the underlying scientific principles are almost never raised in the teaching of statistics to our future scientists.

The literature on how significance testing is misinterpreted goes back nearly as far as significance testing itself. Berkson's 1938 attack on the false reasoning behind the method appeared over 75 years ago. A complete list of authors that have attempted to awaken people's thinking on the subject is much too large to be easily summarized. I will provide a favorite which is particularly appropriate here. The following is a quote from Cohen (1994): "What's wrong with significance testing? Well, among other things, it does not tell us what we want to know, and *we so much want to know that, out of desperation, we nevertheless believe that it does*!"

In this note, I have explored the similarities between the errors committed in the justification of significance testing and heuristic biases often committed in human reasoning. The fallacy perpetrated when a significance test is misinterpreted indeed matches very closely with these heuristic biases. Perhaps the exploration of a reason for why significance tests are misunderstood will provide the push toward an approach to teaching statistics more in line with good scientific principles.

REFERENCES

Begley, C. and Ellis, L. (2012). Drug development: Raise standards for preclinical research. *Nature,* 483, 531-533.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association,* 33, 526-542.

Cohen, J. (1994). The earth is round (p less than .05). *American Psychologist,* 45, 997-1003.

Devore, J. and Peck, R. (2012). *Statistics: The Exploration and Analysis of Data.* Boston: Brooks/Cole.

Gelman, A. and Nolan, D. (2002). *Teaching Statistics: A Bag of Tricks.* New York: Oxford University Press.

Goodman, S. (1999a). Toward evidence-based medical statistics: 1. The p-value fallacy. *Annals of Internal Medicine,* 130, 995-1004.

Ioannides, J. (2005). Why most published research findings are false. *PLoS Medicine,* 2(8), e124.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.

Mathews F., Johnson P., Neil A. (2008). You are what your mother eats: Evidence for maternal preconception diet influencing foetal sex in humans. *Proceedings of the Royal Society: Biological Sciences*, 275, 1661-1668.

Moore, D., McCabe, G., and Craig, B. (2012). *Introduction to the Practice of Statistics.* New York: W.H. Freeman.

Oster, G., Delea, T., and Colditz, G. (1988). Maternal smoking during pregnancy and expenditures on neonatal health care. *American Journal of Preventive Medicine,* 4, 216-219.

Sullivan, M. (2013). *Statistics: Informed Decisions Using Data.* Boston: Pearson.

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review,* 90, 293-315.

Wainer, H., and Zwerling, H. (2006). Logical and empirical evidence that smaller schools do not improve student achievement. *The Phi Delta Kappan,* 87, 300-303.

Walpole, R., Myers, R., Myers, S., and Ye, K. (2012). *Probability and Statistics for Engineers and Scientists.* Boston: Prentice Hall.

Ziliak, S. and McCloskey D. (2008). *The Cult of Statistical Significance*. Ann Arbor, Michigan: University of Michigan Press.