ISSA Proceedings 1998 – Presumptive Reasoning And The Pragmatics Of Assent: The Case Of Argument Ad Ignorantiam

1. Three Theses

This paper focusses on three traditional distinctions commonly made by argumentation theorists. The distinctions generally correlate with one another and work together in picturing argumentation and framing puzzles about it. Not everyone holds all or any of them – maybe not even most. But the distinctions are invoked and alluded to often enough that we think it useful to challenge them directly.

First, there is a distinction to be drawn between justifying the truth or falsity of a proposition or claim and justifying acceptance or rejection of a proposition or claim. The truth or falsity of a proposition is a matter of independent reality. Acceptance or rejection of a proposition is a voluntary decision. Rational justification of acceptance or rejection is a matter of choice, a weighing of costs and benefits. Rational justification of truth or falsity is a matter of evidence, a balancing of facts. Justifying truth or falsity is a matter of proof; justifying acceptance or rejection is a matter of persuasion.

Second, a distinction should be maintained between arguments over propositions of fact and arguments about propositions of policy. It is a distinction closely related to the first in its rationale. It relies on such matters as the difference between description and evaluation, “is” and “ought”, reasons and motivations, epistemology and politics, epistemic reason and practical reason.

Third, a distinction should be maintained between demonstrative proof and plausible demonstration. The former kinds of arguments are associated with strong conclusions involving direct evidence, certainty, necessity, infallibility and the like. The latter kinds of arguments deal with a balance of considerations, presumptions, probabilities, and tentative conclusions.

One can, of course, maintain all these distinctions as conceptual distinctions, which is to say that these distinctions mean different things, they have different

implications, and they participate in different systems of concepts and puzzles. But presumably these distinctions are more than just conceptual. Presumably they point to real differences in the way in which argumentation is conducted in different domains and help to explain real differences in our sense of the quality of those arguments.

Traditionally, at least, scientific research has been held up as a paragon of demonstrative proof concerning the truth and falsity of propositions of fact. Its procedures of inference are highly formalized through statistical analysis. Its research questions are answered on the basis of quantifiable facts that are scrupulously guarded from questions of value. Its empirical claims seem to be as directly demonstrated and as certain as one can get. If these distinctions hold up anywhere, they should hold up here. In fact, there are important ways in which these distinctions blur when we examine the logic of the statistical analysis upon which modern scientific research depends.

2. Statistical Reasoning as Plausible Reasoning

The core of statistical analysis in empirical research is the logic of hypothesis testing. Factual propositions that are derived from theory and predict empirical differences (research hypotheses) are tested against observed differences. The test occurs by setting the research hypothesis against a competing, default hypothesis – typically the null hypothesis that there are no real differences. Now, it isn't news to anyone that the test of whether the observed differences best match the research or the null hypothesis is a matter of probabilistic inference. But it is worth noting that the logic of hypothesis testing is also a logic of presumptive reasoning. In fact, the statistical inference amounts to *argumentum ad ignorantiam* (cf. Walton, 1996a).

Setting very high the level of proof required to establish the research hypothesis creates a heavy presumption in favor of the null hypothesis. In the absence of compelling evidence to the contrary, normal researchers assume their data shows that no actual effects or differences are present (or, that only trivial effects or differences exist). This is what tests of statistical significance amount to (even when taken together with tests of statistical power). As Cohen (1988: 1-2) puts it:

When the behavioral scientist has occasion to don the mantle of the applied statistician, the probability is high that it will be for the purpose of testing one or more null hypotheses, i.e., "the hypothesis that the phenomenon to be demonstrated is in fact absent [Fisher, 1949, p.13]." Not that he hopes to "prove" this hypothesis. On the contrary, he typically hopes to "reject" this hypothesis and

thus “prove” that the phenomenon in question is in fact present. Let us acknowledge at the outset the necessarily probabilistic character of statistical inference, and dispense with the mocking quotation marks about words like *reject* and *prove.* This may be done by requiring that an investigator set certain appropriate probability standards for research results which provide a basis for rejection of the null hypothesis and hence for proof of the existence of the phenomenon under test. Results from a random sample drawn from a population will only approximate the characteristics of the population. Therefore, even if the null hypothesis is, in fact, true, a given sample result is not expected to mirror this fact exactly. Before sample data are gathered, therefore, the investigator selects some prudently small value a (say .01 or .05), so that he *may* eventually be able to say about his sample data,”*If the null hypothesis is true*, the probability of the obtained sample result is no more than a,” i.e. a statistically significant result. *If* he can make this statement, since a is small, he said to have rejected the null hypothesis “with an a significance criterion” or “at the a significance level.” If, on the other hand, he finds the probability to be greater than a, he cannot make the above statement and he has failed to reject the null hypothesis, or, equivalently finds it “tenable,” or “accepts” it, all at the a significance level.

The presumption is that unless the variability between observed groups is sizably greater than the variability within the groups, the observed differences should be assumed to be reflections of random error in sampling and measurement rather than reflections of real differences between populations sampled.
That the logic of statistical inference is a logic of plausible reasoning based on presumption is something that scientists and statisticians implicitly know – though commonly they explicitly disavow such knowledge. The conventional circumlocution used when a significance test fails to support the research hypothesis is that the researcher “fails to reject the null hypothesis.” This way of talking parallels the argumentation theorist’s common explanation for why ad ignorantiam appeals are fallacious: One cannot conclude that a proposition is true simply because one has failed to show that the proposition is false, or vice versa. One can only conclude that no conclusion can be drawn. One doesn’t know the status of the proposition one way or the other. For example, Jaccard (1983: 129) reminds us:
When an experimenter obtains a result that is consistent with the null hypothesis (when it falls between the range of -1.96 and +1.96 instead of outside of it) technically, he or she does not accept the null hypotheses as being true. Rather

he or she fails to reject the null hypothesis. In principle, we can never accept the null hypothesis as being true via our statistical methods; we can only reject it as being untenable.

Similarly, Williams (1992: 79), who talks about "accepting" as well as "rejecting" the null hypothesis, nevertheless warns us:

If a study results in failure to reject a null hypothesis, the researcher has not really "proved" a null hypothesis, but has failed to find support for the research hypothesis. It is not unusual to find studies with negative outcomes where the research has placed a great deal of stock in "acceptance" of null hypotheses. Such interpretations, strictly speaking, are in error because the logic of a research design incorporates the testing of some alternative (research hypothesis) against the status quo (null hypothesis). Although failure to find support for the alternative does leave one with the status quo, it does not rule out other possible alternatives. Put into practical terms, be skeptical of interpretations of unrejected null hypotheses.

Phrases like "technically" and "strictly speaking" are the sorts of euphemisms methodologists use when theory crashes into common sense but don't want to have to admit they are sunk. (Keppel, 1991, uses the euphemistic halfway phrase, "retain the null hypothesis.") And, of course, the reason such theoretical qualifications are set out in the first place is because normal researchers openly disregard them in practice.

It seems then, that the advocate of the traditional distinction between demonstrative proof and plausible argument faces a dilemma. Like so many statistical textbook authors, the advocate can conclude that normal scientific research is widely based on fallacious reasoning and needs to be corrected. Or, the advocate can conclude that well done quantitative empirical research in science really is based on a presumptive form of reasoning. Either way, demonstrative proof seems to be missing from the picture.

We think the reason it is missing is because it is not needed to redeem the rationality of scientific inference, if it ever is needed or ever exists at all. As commonsense reasoners, scientific researchers know that arguments from ignorance are legitimate forms of plausible reasoning when one has a good reason for setting a presumption in the first place. Quantitative analysis in scientific research is plausible reasoning. It is *formally rigorous* plausible reasoning, but it is a kind of plausible reasoning nevertheless: A kind in which

presumptions are established as the levels of proof (in the form of probability assessments) required to accept research hypotheses.

3. Statistical Propositions as Propositions of Policy

The level of proof required to demonstrate the research hypothesis is commonly a matter of convention. Alpha levels in significance testing are ordinarily set at .05. There can be good reason for setting this level of proof that goes beyond a purely arbitrary decision. The nature of this broader rationale once again proves instructive. For the rationale is one in which *argumentum ad consequentiam* plays the decisive role. And this suggests to us that another distinction carries little weight: the distinction between propositions of fact and propositions of policy. Argumentation theorists have long recognized that while *ad consequentiam* reasoning is an illegitimate proof of a proposition of fact, it can provide compelling support for a proposition of policy (Walton, 1996b). In general, this is because the former would involve an illicit shift from a question of what 'ought' to be, or one of value, to a question of what 'is,' or one of fact. And this is said to be an intrinsic difference between propositions of policy and propositions of fact. Yet this does not appear to be a scrupulously guarded distinction in the logic of hypothesis testing.

Go back to the question of setting the level of statistical significance in hypothesis testing. Textbook authors commonly explain that the level of proof necessary to accept and reject the null and research hypotheses is dependent on both the *risk* of inaccuracy and the *cost* of inaccuracy. In statistical jargon, this process is labeled as committing Type I and Type II errors. Type I error is committed when one rejects the null hypothesis when the null hypothesis is in fact 'true'. Type II error takes place when one accepts (fails to reject) the null hypothesis when the null hypothesis is in fact 'false'. Rosenthal and Rosnow (1991: 41) colorfully describe these two errors an inferential mistake involving "gullibility" (Type I error) while Type II error involves being "blind to a relationship."

These errors are inversely related: when the likelihood of committing Type I error is decreased the likelihood of Type II error is increased. The probability of committing either type of error is determined by setting an alpha level required to accept a hypothesis. A higher than usual alpha level (say, $p = .10$) increases the likelihood of committing Type I error while a lower than usual alpha level (say, $p = .01$) increases the possibility of committing Type II error.

When explaining the rationale for this deciding the alpha level, statistical theorists almost uniformly turn to a utility model of decision-making, calling on

researchers to balance risks and costs of the two types of errors. Summers, Peters and Armstrong explain that the goal of researchers is in deciding which error to make, and "it would make sense to choose limits that balance expected costs of Type I and Type II errors. (1981: 248)" Likewise, Mood and Graybill (1963: 279) explain, "to arrive at a reasonable value for alpha requires an experimenter to weigh the consequences of making a Type I and Type II error." Rosenthal and Rosnow (1991: 455) suggest that the balancing is in effect a practical judgment of consequences: If an investigator has decided to set alpha (a) at .05 and is conducting a test of significance with power = .40, beta (b) will be 1-.40, or .60. Then the ratio of b /a will be .60/.05 = 12 implying a conception of Type I errors (a) as 12 times more serious than Type II errors (b).

The consequentiality of factual decision-making, however, is most apparent when statistics textbooks create a practical context. Heiman (1992: 292-293) explains the reasoning with the following concrete illustration:

We typically set alpha at .05 because .05 is an acceptably low probability of making a Type I error. This may not sound like a big deal. But the next time you fly in an airplane, consider the possibility that the designer's belief that the wings will stay on may actually be a Type I error. A 5% chance is scary enough – we certainly do not want more than a 5% chance that the wings will fall off. Sometimes we want to reduce the probability of making a Type I error even further, and then we usually set alpha at .01. For example, we might have set alpha at .01 if our smart pill [a hypothetical intelligence-inducing pill] had some dangerous side-effects. We would be concerned about subjecting the public to these side-effects, especially if the pill does not work. Intuitively, it takes even more to convince us that the pill works, and thus there is a lower probability that we will make an error.

Similarly, Hays (1994: 284) explains: Within contexts such as the test of a new medication in which Type I error is abhorrent, setting a extremely small is manifestly appropriate. Here, considerations of Type II error are actually secondary. In some instances in a social science as well, Type I error clearly is to be avoided, and from the outset the experimenter wants to be sure that this kind of error is very improbable.

Jaccard (1983: 131) also illustrates the reasoning in terms of the widely used medical scenario:

The tradition of adopting a conservative alpha level in social science research evolved from experimental settings where a given kind of error was very

important and had to be avoided. An example of such an experimental setting is that of testing a new drug for medical purposes, with the aim of ensuring that the drug is safe for the normal adult population. In this case, deciding that a drug is safe when, in fact, it tends to produce adverse reactions in a large proportion of adults is an error that is certainly to be avoided. Under these circumstances a small alpha level is selected so as to *avoid making the costly error.* With a conservative alpha level, the medical research takes little risk of concluding that the drug is safe when actually it is not. Thus, the practice of setting conservative alpha levels evolved from situations where one kind of error was extremely important and had to be avoided if possible.

Keppel (1991: 56), on the other hand, talks about what is important simply in terms of the more general intellectual and academic costs and benefits of the decision:
Every researcher must strike a balance between the two types of error. If it is important to discover new facts, then we may be willing to accept more Type I errors and thus *increase* the rejection region. On the other hand, if it is important not to clog up the literature with false facts, which is one way to view Type I errors, then we may be willing to accept more Type II errors and *decrease* the rejection region.

All these authors and many others discuss the decision-making process in terms of consequences, costs, importance, seriousness, or severity of error. In other words, research conclusions are inextricably bound up in *ad consequentiam* reasoning. In fact, the seeming objectivity of the ".05″ level of significance testing is a reflection of just the opposite – an arbitrary judgment based on lack of sufficient information:
The inverse relationship of the risks of the two types of error makes it necessary to strike a reasonable balance. . . . But conventions are useful only when there is no other reasonable guide. . . . In much research, of course, there is no clear basis for deciding whether a Type I or Type II error would be more costly, and so the investigator makes use of the conventional level of determining statistical significance. (Sellitz, Jahoda, Deutsch & Cook, 1959: 418).

When making a decision regarding making type I or type II errors, the loss function associated with the two errors must be known before a rational choice concerning alpha can be made. However, experimenters in the behavioral sciences are generally unable to specify the losses associated with the two errors

of inference. *The use of the .05 or .01 level of significance in hypothesis testing is a convention.* (Kirk, 1968: 2, sec. 1.5).

Pretty clearly then, the rationale for statistical significance testing relies heavily on argumentum ad consequentiam. It seems then, that the advocate of the traditional distinction between propositions of policy and propositions of fact faces a dilemma. Unless this distinction is a chimera, either the advocate must conclude that statistical argument is grounded in a real howler (illicitly converting 'ought' to 'is'), or the advocate can conclude that scientific reasoning is not really factual reasoning at all. Neither option seems to be attractive to those who would maintain the empirical utility of distinguishing propositions of fact and policy.

4. The Pragmatics of Decision-Making

We think both dilemmas above are a reflection of still a deeper breakdown in distinctions: that between justifying the truth and falsity of propositions and justifying the rationality of their acceptance or rejection. We will not bother to rehearse the argument that statistical decision-making is concerned primarily with the latter and only indirectly with the former. The briefest review of the language quoted above should be convincing enough. Quantitative empirical research in science does not justify the truth or falsity of empirical propositions per se; rather it justifies the rationality of accepting or rejecting such propositions. Scientific theory and empirical knowledge is a matter of *deciding* what to *treat* as true or false. All of the language of statistical inference works at that level. It is a meta-level. It should not be surprising then, that ad consequentiam reasoning – matters of utility and usefulness rather than truth – should rest at the heart of empirical knowledge and reasoning. And it should not be surprising either that statistical inference and scientific reasoning is plausible reasoning based on practical presumptions. But if that is what we find in this domain of knowledge, where exactly would we find anything else?

REFERENCES

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jaccard, J. (1983). *Statistics for the Behavioral Sciences*. Belmont, CA: Wadsworth.

Keppel, G. (1991). *Design and Analysis* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Kirk, R. E. (1968). *Experimental Design: Procedures for the Behavioral Sciences.* Belmont, CA: Wadsworth.
Mood, A. M. & F. A. Graybill (1963). *Introduction to the Theory of Statistics* (2nd ed.). New York: McGraw-Hill.
Rosenthal, R. & R. L. Rosnow (1991). *Essentials of Behavioral Research* (2nd ed.). New York: McGraw-Hill.
Sellitz, C., M. Jahoda, M. Deutsch & S. W. Cook (1959). *Research Methods in Social Relations* (rev. ed.). New York: Holt, Rinehart and Winston.
Summers, G. W., W. S. Peters & C. P. Armstrong (1981). *Basic Statistics in Business and Economics* (3rd ed.). Belmont, CA: Wadsworth.
Walton, D. N. (1996a). *Arguments from Ignorance*. University Park: Pennsylvania State University.
Walton, D. N. (1996b). *Argumentation Schemes for Presumptive Reasoning.* Mahwah, NJ: Lawrence Erlbaum Associates.
Williams, F. (1992). *Reasoning with Statistics* (4th ed.). Ft. Worth, TX: Harcourt Brace Jovanovich.