A Formal Test of Substantive Significance

Justin Esarey
Assistant Professor of Political Science
Emory University
jesarey@emory.edu

February 13, 2010

Abstract

While formal tests exist for statistical significance, researchers have traditionally relied on informal arguments to demonstrate the substantive significance of their results. I introduce the first formal test for the existence of a substantively meaningful relationship in quantitative data. The test takes a rational choice perspective toward evidence, using Bayesian statistical decision theory to ask whether it makes sense to believe in the existence of a statistical relationship given a researcher's view of the consequences of correct and incorrect decisions. The test generates a critical test statistic c^* with a clear interpretation: if a relationship of size c^* is not important enough to influence future research and policy advice, then the evidence does not support the existence of a substantively significant effect. A replication of findings from the *American Journal of Political Science* and *Journal of Politics* illustrates that statistical significance at conventional levels is neither necessary nor sufficient to accept a hypothesis of substantive significance using c^*. I make software packages available for Stata and R that allow political scientists to easily use c^* for inference in their own research.

Introduction: statistical inference and rational choice under uncertainty

Political scientists are expected to demonstrate that an empirical finding is large and important enough to matter, not just that it exists. A variety of work across the social sciences has convincingly shown that a statistically significant finding is not necessarily a substantively significant finding, which I define to mean a finding worth integrating into the body of knowledge used to shape future research and policy advice (for a review, see Gill, 1999; see also

McCloskey, 1998, Chapter 8; Ziliak and McCloskey, 2008). But how *do* we decide whether a result constitutes convincing evidence of a substantively significant relationship? Many researchers go beyond t-testing to calculate and showcase substantively relevant marginal effects, usually an estimate of the change in a dependent variable Y associated with a change in an independent variable X and the uncertainty around this estimate (King, Tomz and Wittenberg, 2000; Gelman, Pasarica and Dodhia, 2002; Kastellec and Leoni, 2007). A researcher then uses this information to argue that an effect is qualitatively large and certain enough to be substantively meaningful (Miller, 2008; Ziliak and McCloskey, 2004).

While the informal reasoning that a researcher uses to argue for the substantive significance of a result is usually sound, s/he may have a difficult time communicating to others precisely how the strength and uncertainty of an effect, along with a scientific aversion to mistakenly accepting the existence of a relationship between X and Y where none exists, combines to form his/her judgment (Lunt, 2004). It may be difficult for that researcher to use consistent standards across multiple studies, particuarly when those studies are in different substantive areas. It will be even more difficult for that researcher to relate his/her judgments of substantive significance to judgments that others make. For all these reasons, researchers may find it useful to have a formalized test statistic for substantive significance, akin to the t-test for statistical significance, that quantitatively crystallizes this judgment in a transparent, standardized, and communicable way.

In this paper, I introduce the first method to formalize the notion of substantive significance. To do so, I use Bayesian statistical decision theory (Wald, 1950; DeGroot, 2004 {1970}; Pratt, Raiffa and Schlaifer, 1996; Manski, 2007, ch. 12) to apply principles of rational choice under uncertainty to statistical inference, creating a c^* statistic that researchers can use to clarify and regularize the principles that they already use to make informal judgments about substantive significance. This statistic combines quantitative measures of (1) the level at which an effect becomes substantively meaningful, (2) the uncertainty surrounding the result, and (3) the researcher's desire to avoid false positives (type I errors) even if

more false negatives (type II errors) are created.

The c^* statistic allows a researcher to share exactly how and why s/he made a judgment about substantive significance in objective and transparent terms. A researcher can also precisely determine how sensitive that judgment is to changes in the strength and uncertainty of a result or in his/her aversion to false positives, a marked advantage over informal reasoning. Finally, researchers can more easily compare their judgments to those of other researchers, making it clear why and how researchers may agree or disagree about what it takes for a result to be substantively meaningful.

A replication analysis of papers recently published in *American Journal of Political Science* and *Journal of Politics* demonstrates that judgments of statistical significance drawn from a t-test or confidence interval can differ from judgments of substantive significance drawn using c^*. I find that effects that are strongly statistically significant sometimes have small c^* values that are probably not substantively meaningful, which is consistent with the arguments of prior work. But I also find that some effects that are statistically insignificant have c^* values large enough to be considered substantively meaningful. Thus, statistical significance is not just insufficient for substantive significance, but also not necessary—an extension of previous findings attributable to the formalization of substantive significance.

The logic of substantive significance

How do social scientists decide what constitutes convincing evidence of a substantively significant relationship? I believe that the informal reasoning process already used by most political scientists can be fruitfully compared to the hypothesis testing framework with which we are already familiar, a comparison that will prove useful in constructing a formal test statistic for substantive significance. To demonstrate the usefulness of this comparison, suppose that a researcher finds that the presence of a strong, politically independent central bank is related to higher incomes for a country's poorest quintile. In fact, Romer and Romer (1999) found

that every one percent increase in inflation rates is associated with a 5.71% decline in the income of the poorest fifth of the population, with a standard error of 2.93%.[1] Using conventional thresholds for the t-test, this effect is marginally statistically significant—that is, we can tentatively rule out the possibility of zero relationship.[2] But is this finding substantively significant, and why (or why not)?

A researcher's default belief—akin to a *null hypothesis*—is that future beliefs and actions should be predicated on the assumption that central bank independence has no effect on poverty. S/he should not include the independence of the central bank as a variable when examining the effect of political institutions on poverty. S/he should not teach students that the political insulation of monetary policy actually helps society's poorest. Any advice that the researcher provides to policymakers should be consistent with no relationship between central bank independence and poverty.

In examining the empirical evidence, the researcher is deciding whether to revise this default belief—that is, to start conditioning his/her actions on the idea that strong and independent central banks *are* associated with lesser poverty. This revision is akin to an *alternative hypothesis*. If the researcher begins acting on this alternative belief, and that belief is correct, s/he will receive benefits over continuing to believe in the status quo of no substantively significant relationship. His/her policy advice will be more efficaceous at eradicating poverty. His/her teaching will be more informative to students and enable them to have a better understanding of the links between political institutions and economic outcomes. Future research will probably be more accurate—at the least, we will have gained a new control variable. This future research might also be more interesting and informative, because the finding raises new questions about the interplay between political institutions and poverty.

[1] See (1999) p. 37, Table 5, column 3. The dependent variable in their model is $\ln(income)$ and the key independent variable is a country's average inflation rate. For such log-level models, the beta coefficients can be interpreted as percentage changes in the dependent variable with respect to unit changes in the independent variable. The data is a cross-section from 1988 (or nearby years, when necessary).

[2] The two-tailed p-value is 0.0512.

In every case, these benefits are proportional to the magnitude of the relationship. If the creation of an independent central bank caused the incomes of the poor to rise by 200%, the academic and policy value of that knowledge would be immense. This effect would be a key finding for political economists, an important example of the power of institutional design, and a crucial tool for alleviating suffering at home and abroad. By contrast, a 0.01% rise in income is so small that it can safely be ignored; indeed, spending time studying and making political use of such a tiny effect probably presents a greater opportunity cost than it is worth.

But the researcher must weigh the potential benefits of accepting what I have called the alternative hypothesis against its costs—namely, the possibility of falsely accepting this alternative. Research effort may be squandered trying to theoretically explain and empirically elaborate on a non-existent relationship between central bank independence and poverty. Misinformed policy advice would result in wasted effort at best and increases in poverty at worst. A researcher's professional reputation might be damaged by endorsing a conclusion that turns out to be misleading. And just as benefits are proportional to the magnitude of an effect, the more incorrect that the alternative hypothesis is, the more damaging it is to accept it. It is harmful to believe that central bank independence alleviates poverty if it has no effect, but disastrous if central bank independence actually *increases* poverty.

Most political scientists probably believe that the consequences of falsely accepting an alternative hypothesis are worse than the consequences of ignoring a correct alternative hypothesis. Indeed, in the context of statistical significance testing, political scientists are used to putting more priority on avoiding a false positive (mistakenly concluding that some statistical relationship exists) than avoiding a false negative (mistakenly concluding that there is no statistical relationship). This practice is consistent with science's healthy skepticism toward novel claims. The aversion to risk embodied by this approach helps reseachers to avoid mistakes that push future theorizing and data collection down blind alleys and misinform our teaching and policy advice.

To summarize, I believe that the following are the key tenets[3] of the approach that political scientists do take—and ought to take—when making judgments of substantive significance:

1. The null hypothesis (no substantively significant relationship between X and Y) is the status quo.

2. Belief in a *correct* alternative hypothesis (a substantively significant relationship between X and Y in a positive or negative direction) improves on the status quo; belief in an *incorrect* alternative hypothesis is worse than the status quo.

3. The benefit/harm to accepting an alternative hypothesis (relative to maintaining belief in the null) is proportional to how correct/incorrect the belief is.

4. Related to the point above, a relationship must be above some critical size before it is important enough to act on; that is, belief in tiny and substantively meaningless effects causes harm.

5. Political scientists should be averse to the risks presented by an alternative hypothesis; that is, they should believe that belief in *incorrect* alternative hypotheses is proportionally more harmful than a failure to believe in *correct* alternative hypotheses.

These will serve as the assumptions around which I will construct a formal test statistic for substantive significance, and I will often refer back to these assumptions as I lay out the construction of that statistic. The goal is for that statistic to quantitatively embody these assumptions, which I believe that most political scientists already use when thinking about substantive significance. As a result, the statistic should prove a useful tool to standardize, clarify, and make consistent our judgments of statistical significance.

[3]For a similar analysis of the essential components of substantive significance, see Miller (2008).

Rational choice under uncertainty: a formal framework for Bayesian statistical decision theory

The test statistic that I propose is built on the foundations of Bayesian statistical decision theory, which presents and formalizes a useful way of looking at evidence. As far as I am aware, none of this work has been directly applied to the problem of substantive significance. It may be that prior researchers did not believe that the concept of substantive significance was precise enough, or consistent enough across multiple fields of substantive inquiry, to inform the design of a test statistic. But in addition, some statistical decision theorists see their work as fundamentally separate from hypothesis testing. This perspective is alluded to by Manski:

> Why did statistical decision theory lose momentum long ago? ...Another reason may have been diminishing interest in decision making as the motivation for statistical analysis. Modern statisticians and econometricians tend to view their objectives as estimation and hypothesis testing rather than decision making (Manski, 2007, p. 244).

But some work in statistical decision theory, including very foundational work, *does* address the problem of binary hypothesis testing—though to a different end than the one to which I put it. For example, Wald (1950, p. 18) says that the typical binary hypothesis is a "special case of the general decision problem" and notes that size and power (the probability of incorrect and correct rejection of a null hypothesis, respectively) are "special cases of the notion of risk in the general decision theory" (Wald, 1950, p. 20). This is quite consistent with the underpinnings of my own approach.[4] As I construct my own test statistic for

[4]The most pertinent prior work focuses on decisions about a coefficient's sign (i.e., whether $\beta > 0$) based on a dataset. See, e.g., Schervish (1995, pp. 239-246), Robert (2001, Chapter 5, especially pp. 254-259), and DeGroot (2004 {1970}, pp. 244-247).

substantive significance in the following sections, I will note some linkages to and variations from the literature in decision theory.

Inference as a rational bet

When researchers decide whether evidence is sufficient to accept an alternative hypothesis, like the hypothesis of a substantively significant positive or negative relationship, Bayesian statistical decision theory argues that their decision should be fundamentally similar to the one made by a rational bettor (Pratt, Raiffa and Schlaifer, 1996, pp. 11-46). Political scientists are trying to decide whether empirical evidence makes some scientific proposition a good bet worth believing in (at least provisionally) or a bad bet that needs to be rejected. If we are falsificationists, we may say that we are trying to decide whether the null hypothesis or the alternative hypothesis is a *better bet* given the evidence. Good falsificationists should be reluctant to reject the null hypothesis without especially strong evidence: we should be *loss averse bettors* with respect to the new and uncertain alternative hypothesis, as indicated by assumption (5) above.

In other words, a person should approach the inference problem laid out in the previous section in the same way that s/he would approach a bet on a game of chance. Consider a simple illustration: does it make sense to take a bet with a 50% chance to win $6 and a 50% chance to lose $4? While the expected payoff is $1, this does not necessarily mean that turning down the bet is irrational: a person may fear the $4 loss more than s/he values the $6 gain. The person may dislike uncertainty about the future, and thus pay a psychic cost for taking chances. Losing $4 might prove proportionally more catastrophic than the benefit provided by winning $6; consider whether most people would risk the last $4 they had, money that they needed to eat, on a game of chance.

Likewise, and for the same reasons, analysts want to be wary of accepting a hypothesis of a positive relationship between two variables—even when the balance of evidence indicates a positive relationship—if there is a substantial possibility that the relationship is actually

negative (or vice versa). In particular, scientists are cognizant of the outsized consequences of a false positive—namely, mistaken policy advice, harmful teaching, and wasted research resources. Of course, when it comes to social research, there are not cash payoffs determined by a randomizing device like a throw of the dice. But as discussed earlier, there are certainly potential costs and benefits associated with switching from a belief in the null to a belief in the alternative. What political scientists want to know is whether changing their belief, and changing their research agenda, teaching, and policy advice accordingly, is a good bet—is likely to change the status quo in a desirable direction—compared to standing pat. That is, they want to know whether accepting a hypothetical statement—something like "increases in X tend to be associated with increases in Y" or "$\beta \geq c$"—is a good bet compared to standing pat on the belief that no such effect exists. Drawing on assumptions (1), (2), and (4) from the prior section, the relevant question should be: "Given that falsely accepting an incorrect prediction is more important to me than falsely ignoring a correct one, and given the minimum size threshold for a relationship to be substantively important, do my findings constitute evidence of a substantively meaningful positive (or negative) relationship between X and Y?"

The analogy to betting enables us to leverage the wealth of formal knowledge about rational choice under uncertainty built over the last century and apply that knowledge to statistical inference procedures.[5]

Choice under uncertainty[6]

We start by assuming that a decision maker i has (or behaves as if s/he has) a utility function $u_i(w)$ that links objective outcomes w to the subjective value that the decision maker derives

[5] A considerable amount of research demonstrates that people frequently deviate from the predictions of expected utility theory (and rational choice theory more generally). However, the present task is *not* to accurately describe human behavior, but to provide a framework that will help researchers to draw statistical inferences that rationally balance the concerns of aversion to mistakes, the presence of uncertainty, and the need for substantive importance. The tools of expected utility theory and loss/risk aversion are well-suited to the fundamentally *prescriptive* task of guiding people to rational decisions under these conditions.

[6] For helpful references to this material as treated by other statistical decision theorists, see DeGroot (2004 {1970}, Chapter 7) and Pratt, Raiffa and Schlaifer (1996, Chapters 3 and 4).

from receiving this outcome. In the next subsection I will discuss the form that this utility function should take when drawing statistical inferences, but for now I leave the function abstract for clarity's sake. When facing a choice under uncertainty—should I take a bet, or not?—a rational decision maker should determine whether the expected utility of the bet exceeds zero. That is, if a bet that costs c has payoffs of $\{w_1, w_2, ..., w_k\}$ with corresponding probabilities $\{p_1, p_2, ..., p_k\}$, $\sum_j p_j = 1$, then the rational chooser should determine whether:

$$\sum_j u_i(w_j)p_j - c > 0 \tag{1}$$

If the payoff and probability is continuous rather than discrete, then we exchange the point probability p_i with a probability distribution $f(w)$ and write:

$$\int u_i(w)f(w)dw - c > 0 \tag{2}$$

Bayesian statistical decision theory replaces the language of bets and payoffs with the language of statistical estimates. Thus, when examining empirical evidence, the decision maker faces an array of potential values of a parameter β whose probability of being correct is associated with a posterior probability density $f(\beta|\text{data})$. The analyst makes a decision d which then yields a utility contingent on the true value of the parameter. How decisions connect to utility is a function of the particular kind of decision that the analyst is trying to make.

The decision faced by researchers and policy makers in the setting I propose is *binary*: accept the existence of a substantively meaningful relationship in some direction, or reject the existence of that relationship. To decide whether to accept or reject the relationship, we need to determine whether the researcher's expected utility of acceptance is greater than the expected utility of rejection:

$$E[u(\text{accept})] - E[u(\text{reject})] = \int \left[u(\text{accept}|\beta) - u(\text{reject}|\beta)\right] f(\beta|\text{data})d\beta \tag{3}$$

In the terminology of Bayesian statistical decision theory, $u(\text{accept}|\beta) - u(\text{reject}|\beta)$ is the *loss function* associated with deciding to accept a hypothesis of substantive significance (Robert, 2001, pp. 60-65; Pratt, Raiffa and Schlaifer, 1996, pp. 489-493; Schervish, 1995, pp. 144-146). The level of loss is contingent on the unknown value of β. The expected value of loss, given in equation 3, represents the *risk* associated with the decision to accept the hypothesis of substantive significance.

Naturally, the form that $u(\text{accept}|\beta) - u(\text{reject}|\beta)$ should take is a possible point of contention: this form will directly impact our judgment of whether statistical evidence is acceptable. I will argue in the next section that the assumptions that I made about the goals of a social scientist imply a specific mathematical form for $u(\text{accept}|\beta) - u(\text{reject}|\beta)$.

Knowledge as payoff

The hallmark of the political scientist's problem is the question of whether a change in X will cause a change in Y. While the framework I am about to describe can apply to any estimated parameter, for simplicity let us consider a linear model estimated via Ordinary Least Squares:

$$Y = \beta_0 + \beta X + \varepsilon$$

If a researcher accepts a positive (or negative) relationship between an independent variable X and a dependent variable Y rather than rejecting it, assumptions (2) and (3) say that the importance of that decision should be related to the strength of the relationship:

$$\frac{\partial Y}{\partial X} = \beta$$

That is, the "payoff" is the influence of X on Y, which in this framework is given by β.[7] If

[7]This statement is true in the context of OLS; for GLM models in other families, such as logit and probit, the quantities of interest may need to be computed separately.

the value of β is in the predicted direction, the payoff from accepting the prediction should be positive. If the value of β is *not* in the predicted direction, the payoff from accepting the prediction should be negative (i.e., we are harmed by accepting that prediction).

For political scientists, the payoff lies in finding and describing a new relationship, and stronger relationships mean a greater power to predict and explain. To reiterate an earlier point, finding that independent central banks are associated with slightly more economic growth (on the margin of substantive significance) is not as notable as finding that central banks are associated with large differences in growth that could gradually move a country into the ranks of the developed world. For governments and corporations, changes in some factor Y are usually directly related to monetary payoffs (profits, government revenues, GDP growth, and so on). As a result, the size of the β term in a linear regression serves as a handy measure of the payoff of accepting a conclusion for both policy makers and academics. Larger predictions in the correct direction mean greater benefit for both groups, and both groups's goals are more frustrated by larger predictions in the wrong direction.[8]

Assumption (4) emphasizes that an effect must breach some critical size threshold before it should be considered substantively significant. As a result, the net research, policy, and teaching benefit that comes out of a result should be a function of how large it is compared to that threshold. Calling this threshold c, we should believe that the net utility of accepting the alternative hypothesis (and rejecting the null) is a function of $\beta - c$, the extent to which β exceeds the critical threshold for substantive significance.

The parameter c may be viewed as the size of β at which a relationship between X and Y becomes *substantively significant* (worth using to shape future research and policy advice) and below which there is no substantively important relationship between X and Y. For example, factors that increase the income of the poor by less than 0.1 percentage points may be too unimportant to be worth recommending to policy makers; they simply do not have

[8]Of course, the scale of X and Y will determine the scale of β and doubling these scales will not double β's importance: importance is a function of the size of β relative to the scale of X and Y and our substantive judgment of the importance of changes in Y.

an appreciable effect on the state's welfare. Nor are effects of this size important enough to warrant inclusion in a curriculum or future research projects: there are other influences more central to growth and more worthy of academics' limited time and attention. In short, findings smaller than c are actually detrimental to accept as substantively significant.

What this framework indicates is that a belief in an effect's substantive significance can be usefully conceptualized as a belief that $\beta > c$ (if our prediction is that β is positive) or that $\beta < c$ (if our prediction is that β is negative). This is the mathematical translation of the verbal statement that the relationship between X and Y is substantively significant, worth using as a basis for scientific, political, and business judgments.

To summarize, in order to be consistent with the process that political scientists use to make judgments of substantive significance, findings consistent with a theory should yield positive utility, and findings inconsistent with a prediction yield negative utility. Therefore, $u(\text{accept}) - u(\text{reject})$ should be positive when $\beta > c$, if our prediction is that β is positive, and negative otherwise. If our prediction is that β is negative, then $u(\text{accept}) - u(\text{reject})$ should be positive when $\beta < c$ and negative otherwise. But this guidance is imprecise; how should we determine the exact form of $u(\text{accept}) - u(\text{reject})$ and the appropriate value for c? And what of the desire of researchers to avoid false positives at the expense of a greater propensity for false negatives, assumption (5) above? We answer the first question by way of answering the second.

Loss aversion and risk aversion

So far, the outlined framework explicitly handles the presence of uncertainty in statistical results and incorporates substantive importance into the inference process. But what about a political scientist's belief that mistakenly accepting a false relationship is more important than mistakenly rejecting a true relationship, as embodied in assumption (5)?

Loss aversion describes the tendency to undervalue gains in comparison to losses. Risk aversion describes the tendency to undervalue risky choices in comparison with certain

choices. To return briefly to an earlier example, we might expect a loss averse person to turn down a bet that gave a 50% chance of a \$6 win and 50% chance of a \$4 loss, even though the bet has a positive expected value, because a \$4 loss would have a more significant effect on the person's welfare than a \$6 gain. A risk averse person might turn down this same bet, but may also prefer to accept a certain payment of \$4.50 over a risky investment with a 50% chance of a \$6 gain and a 50% chance of a \$4 gain. A strictly loss averse person would not make this choice, as the risky choice always results in a gain and has a higher expected value.

Theorists usually represent these preference structures with utility functions that map payoffs onto a value function that over- or undervalues certain payoffs. For simple loss aversion, one way to accomplish this task to simply suppose that a gain is worth only a fraction as much as a loss. That is, losses are valued proportionately more than gains, but doubling a loss (or gain) doubles the negative (or positive) utility experienced by the person (Tversky and Kahneman, 1991).

Utility functions that try to capture risk averse preferences generally curve downward, so that increasing payoffs yield ever-smaller increments in utility while increasing losses are increasingly damaging to well-being. The degree of risk aversion that a person exhibits can be measured by the extent to which the utility function curves downward. That is, we say that an individual i values a level of wealth w according to a function $u_i(w)$, then suppose that $-\frac{u_i''(w)}{u_i'(w)} < 0$. This ratio is known as the *Arrow-Pratt coefficient of absolute risk aversion.*

An example of a utility function consistent with a loss averse and a risk averse decision maker is plotted in Figure 1. A loss and risk neutral decision maker assigns a utility value in proportion to the size of a payoff. The loss averse decision maker tends to overvalue losses relative to gains. The risk averse decision maker tends to assign decreasingly larger values to positive payoffs and increasingly more negative values to negative payoffs.

Both risk and loss aversion are easy to mathematically represent in the context of statis-

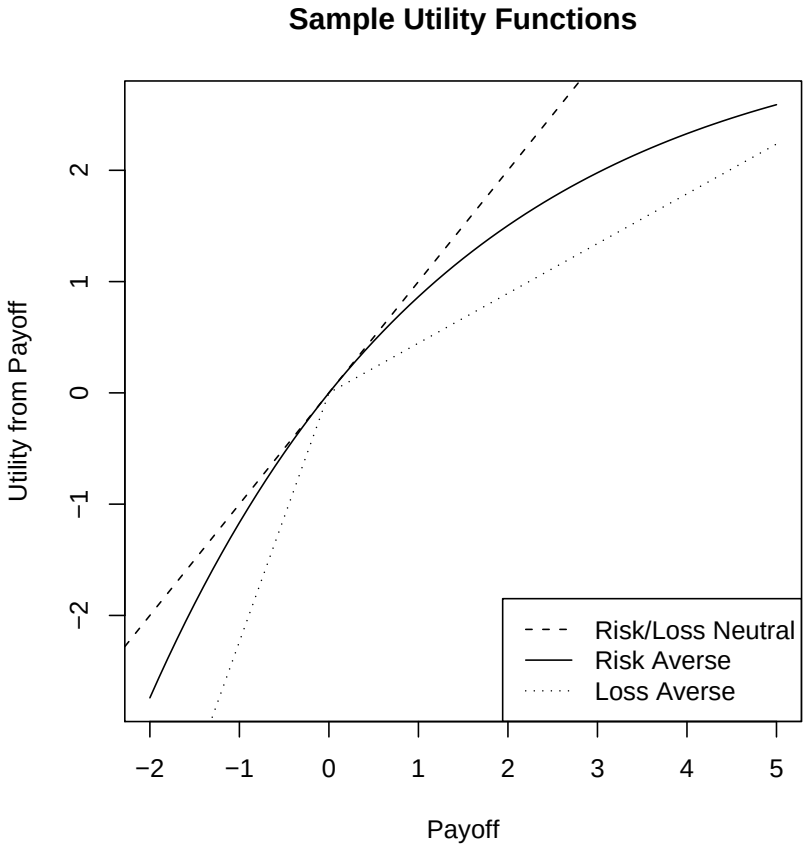

Figure 1: Sample Utility Functions

tical inference. For the alternative hypothesis of a positive β,[9] simple loss aversion can be modeled with a kinked linear utility function:

$$u(\text{accept}|\beta, \gamma, c) - u(\text{reject}|\beta, \gamma, c) = \gamma^{-\text{sign}(\beta - c)} [\beta - c] \tag{4}$$

That is, utility is gained in proportion to distance from the threshold for substantive significance c, with larger correct predictions yielding a greater benefit (and larger incorrect predictions yielding greater harm). This structure is consistent with assumption (3) from my earlier characterization of the reasoning process that political scientists employ to make judgments about substantive significance. γ is a parameter that describes the degree of loss aversion that the researcher exhibits, the degree of kink in the loss averse utility function depicted in Figure 1. This parameter can be determined by conscious choice, a procedure analogous to selecting an α level for conducting a t-test. Indeed, like α, γ calibrates the extent to which a researcher is willing to trade off type I error (false rejection of the null hypothesis) against type II error (false acceptance of the null hypothesis); loss aversion presumes that type I errors are more important to avoid.

Risk aversion is also concisely represented with well-known utility functions, such as the *constant absolute risk aversion* (or CARA) function:

$$u(\text{accept}|\beta, \delta, c) - u(\text{reject}|\beta, \delta, c) = \frac{1}{\delta} (1 - \exp(-\delta [\beta - c])) \tag{5}$$

This risk aversion function has the desirable property that the Arrow-Pratt coefficient is constant, δ, for all values of β. Larger values of δ indicate a higher degree of risk aversion, or greater concavity of the risk averse utility function depicted in Figure 1.[10]

[9]The utility from accepting an alternative hypothesis of a negative relationship uses the same functions below, but replaces β with $-\beta$ and c with $-c$ so that negative β values yield positive utility and so that the substantive threshold for significance is less than zero.

[10]Negative values of δ indicate risk-seeking preferences, or convexity of the utility function, a condition under which people will accept bets with an expected payoff of *less* than zero simply because they enjoy risk. I neglect this condition because, normatively, scientists and policy-makers should not want to accept decisions that are wrong on average simply because of the thrill that they might be correct. Similarly, I

A focus on the role of loss or risk aversion in statistical inference, which springs from a desire to be cautious when coming to new scientific conclusions on the basis of uncertain evidence, distinguishes my work from the typical treatment of one-sided hypotheses (such as the hypothesis that $\beta > c$) in statistical decision theory. More typically, a 0-1 loss function is specified, where the loss = 1 if a correct decision is made and = 0 otherwise. For the hypothesis that $\beta > c$, this amounts to a step function: a horizontal line at $u = 0$ for $\beta \leq c$ and at $u = 1$ for $\beta > c$. Under this loss function, the expected risk is equivalent to a frequentist one-tailed p-value when $c = 0$ (Robert, 2001, pp. 80-81).

A 0-1 loss function is ideal when the objective is to just to maximize the probability of correctly rejecting a null hypothesis (i.e., the *power* of a test) given a specified probability of incorrectly rejecting the null (the *size* of a test, or α-level for a t-test), as in tests of statistical significance (Schervish, 1995, pp. 242-243). But I have argued that a test for substantive significance must factor in the consequences of correct and incorrect decisions, including the *magnitude* of those successes and mistakes, in order to mirror our understanding of the concept. Using a 0-1 loss function would be equivalent to trying to judge substantive significance with a p-value—an idea that social scientists would find deeply unsatisfying (Ziliak and McCloskey, 2008).

Summarily, it is reasonable to assume that social scientists want to be loss or risk averse when deciding whether to accept the substantive significance of new findings. Combined with our previous assumption—that findings consistent with a theory should yield positive utility, and findings inconsistent with a prediction yield negative utility—$u(\text{accept}) - u(\text{reject})$ should either be kinked (in the case of loss aversion) or curved (in the case of risk aversion). The degree of kink/curve should be proportional to the degree of loss/risk aversion desired, which is a matter of choice akin to the choice of an α parameter in a traditional t-test.

ignore values of $\gamma < 1$ representing loss-seeking preferences (and $\gamma = 0$, for which u is undefined).

How can a researcher choose γ and δ?

Before proceeding to statistical inference, the researcher must decide on a level of loss aversion γ or risk aversion δ. In the case of loss aversion, the kinked but linear form of the utility function makes it reasonably simple for analysts to pick a γ. If, for example, an analyst decides that being wrong about the substantive significance and direction of a coefficient is four times worse than being correct, then $\gamma = \sqrt{4} = 2$.

Determining a risk aversion coefficient δ is not as intuitive, but experimental economists and political scientists have developed a procedure for determining a person's level of risk aversion (Holt and Laury, 2002). Recall that individuals' willingness to accept bets is dependent on their loss/risk aversion level. The idea behind these procedures is straightforward: a choice of two lotteries (one riskier, one safer) is offered to a subject. This choice is repeatedly offered, gradually decreasing the riskiness of the risky lottery. A person should eventually switch over from the safe option to the risky option; the point at which the person switches over determines his/her δ.

There is an even simpler way[11] to determine a critical δ that also focuses researchers more narrowly on the issue of statistical inference. A researcher should ask him/herself the following question:

> Suppose that $c = 0$ (any relationship is substantively meaningful) and evidence that you gathered presented you with two[12] possibilities: there is a 95% chance that the true standardized coefficient β describing the relationship between X and Y is $\beta_H = 1$, and a 5% chance that the true relationship is some value of $\beta_L < 0$. How low would β_L have to be before you refused to conclude that $\beta > 0$?

The idea here is to determine how willing the researcher is to make a bet on the basis of

[11] This procedure is similar to a procedure suggested by Pratt, Raiffa and Schlaifer (1996, pp. 77-78), though they repeat it for many choices to construct a less parametrically bound utility function. A similar question can be used to solve for a loss aversion coefficient γ, but γ is intuitively straightforward enough that such a procedure should be unnecessary.

[12] The posterior is presented as a distribution with two point masses to simplify calculations and make the situation easy to understand.

evidence. If β_L is extremely negative, then accepting this evidence might not be worth the risk of being wrong, even though the chance of being wrong is rather unlikely. On the other hand, if β_L is closer to zero, the probable scientific and policy benefit from accepting this evidence outweighs the improbable consequences of being wrong.

Once the researcher reports β_L, the degree of risk aversion can be determined by numerically solving the following equation for δ:

$$0.95\frac{1}{\delta}\left(1-\exp(-\delta)\right)+0.05\frac{1}{\delta}\left(1-\exp(-\beta_L\delta)\right)=0 \tag{6}$$

Finding the solution to this equation is equivalent to finding the level of risk aversion that would make a person indifferent to accepting or rejecting this evidence. For example, if the researcher replied that a $\beta_L = -2$ would be the point at which he/she no longer accepted this evidence, then $\delta = 1.36$. If the researcher replied that $\beta_L = -4$ would dissuade him/her, then $\delta = 0.551$. The lower the β_L that the researcher will accept, the less-curved the utility function is and the closer to zero that δ gets.

Inference and parameter choice

Ultimately, the conclusion that a researcher draws about substantive significance will depend on the γ or δ coefficient chosen. This fact may trouble some readers because the choice of this parameter is ultimately arbitrary, dependent on the researcher's own preferences or ones that the scientific community has arrived at through convention.

It is important to note that this choice is no more arbitrary than the choice of an α level for statistical significance in a conventional t-test, with which the research community is highly comfortable. Furthermore, the informal substantive interpretation of results is no less arbitrary and far less transparent than an explicit choice of γ or δ. A researcher who examines an estimated effect and its standard error would have a hard time communicating to a colleague precisely *why* the effect was (or wasn't) large or certain enough to be substantively

significant. With γ or δ, a researcher can clearly and numerically indicate his/her level of loss or risk aversion and exactly how that degree of aversion caused him/her to interpret the strength of the result.

Finally, and perhaps most importantly, a researcher can also perform and report a sensitivity analysis that allows others to see how differences in loss/risk aversion change the inference that a researcher draws. Thus, while any particular choice of γ or δ may be arbitrary, it is straightforward to report the range of γ or δ over which an inference is supported. This makes the selection of any particular γ or δ less important and less arbitrary to inference, much as the reporting of a p-value makes the selection of any particular α less important to a finding of statistical significance. I will show how such a sensitivity analysis can be conducted in the next section, where I will lay out the nuts and bolts of testing for substantive significance using c^*.

Conducting a critical test for inference

For those using t-tests or 95% credible regions, a p-value of less than .05 is often used as the threshold for statistical significance. This threshold represents a *critical test* in the sense that it is relatively easy to implement and presents a clear and unambiguous yardstick against which results can be measured to determine their acceptability. What constitutes a critical test for substantive significance in the framework that I now propose?

Recall from equation 3 that the statistician's objective is to determine whether the expected utility of accepting the conclusion that β is substantively significant is greater than zero. Thus, we can combine equation 3 with equation 4 to determine the critical condition needed to conclude that β is substantively significant under loss aversion:

$$\int \gamma^{-\text{sign}(\beta - c)} \left[\beta - c\right] f(\beta|\text{data}) d\beta > 0 \tag{7}$$

The condition is similar for the CARA risk averse utility function:

$$\int \frac{1}{\delta} \left(1 - \exp(-\delta \left[\beta - c\right])\right) f(\beta|\text{data}) d\beta > 0 \tag{8}$$

The "Bayesian" in Bayesian statistical decision theory comes from the fact that both conditions rely on the posterior probability of β given the data, $f(\beta|\text{data})$, which is determined by Bayes' rule:

$$f(\beta|\text{data}) = \frac{f(\text{data}|\beta) f(\beta)}{\int f(\text{data}|\beta) f(\beta) d\beta}$$

This will, of course, require the analyst to specify a prior, $f(\beta)$, before determining whether these conditions are met. I will address the issue of prior specification after first discussing how, given a prior, an analyst can check whether these conditions are met. In fact, there are two ways to check these conditions when performing a statistical analysis.

First procedure: compute utilities using a chosen critical c

To conduct a critical test, a researcher settles on a value of c corresponding to the minimum substantively relevant value of $\frac{\partial Y}{\partial X}$. The researcher also decides how loss or risk averse s/he wants to be, perhaps by employing a community standard similar to the $\alpha = .05$ convention. Finally, the researcher determines whether condition 7 (or 8) is met using the selected c and γ (or δ). An expected utility of greater than zero indicates that the null hypothesis of no substantively significant effect should be rejected, while an expected utility of less than or equal to zero indicates that the null should *not* be rejected.

Second procedure: determine a maximum c necessary to accept evidence

An alternative procedure for conducting a critical test presents the same information in a different form. As before, we are primarily interested in whether condition 7 or 8 is met. Rather than taking c as an external given, though, we can instead determine the c^* that

solves:

$$\int \left[u(\text{accept}|\beta, \gamma, c^*) - u(\text{reject}|\beta, \gamma, c^*)\right] f(\beta|\text{data}) d\beta = 0 \tag{9}$$

We substitute δ for γ if using the risk aversion framework.

A researcher must believe that a β coefficient of size c^* is substantively meaningful in order to conclude that the analysis presents evidence of a substantively significant relationship. If the researcher's own minimum threshold for substantive significance is higher than c^*, then the null hypothesis of no substantively significant relationship cannot be rejected. The value of c^* will also be the per-unit cost of a policy intervention below which this evidence warrants using changes in X to cause a change in Y. That is, c^* is the maximum per-unit cost at which a policy intervention is justifiable.

In the terminology of statistical decision theory, solving equation 9 is the equivalent of finding an *admissible* decision rule for substantive significance. An admissible rule is one that always yields the decision with the lowest possible expected loss (Robert, 2001, p. 74; Schervish, 1995, pp. 153-154). In this case, the admissible decision rule is to accept the alternative hypothesis of a substantively significant relationship if the threshold for substantive significance is smaller than c^*, and to reject that hypothesis otherwise. By construction, following this rule always maximizes expected utility (i.e., minimizes expected loss).

While using the same basic principles and information, this procedure has an advantage over the first: it creates a critical statistic c^* that can be reported in the analysis and interpreted by interested readers. Like a p-value, a c^* allows readers to decide individually whether they are scientifically comfortable with rejecting the null hypothesis on the basis of this evidence. The technique also lends itself to easy sensitivity analysis over the choice of γ and δ parameters, as is shown below.

Computing c^* and choice of a Bayesian prior

Computing a c^* is a process that the vast majority of mainstream statistical software packages are already capable of through add-on packages. I have developed software for the R statistical computing environment that, when given either a γ or a δ, will compute a c^* immediately after a linear regression or other generalized linear model. The software will also determine c^* when given manually entered information from an published table (without the accompanying dataset). The software has been ported into Stata for the loss aversion framework.[13]

My software package assumes a truncated uniform prior distribution $f(\beta)$ that gives equal positive probability to values of β that are $\pm 8\sigma$ of the estimated $\hat{\beta}$ and 0 probability to all other values. This structure assumes minimalistic knowledge of the underlying parameters before examining a dataset, consistent with a very cautious approach toward inference. It also allows for a Bayesian interpretation of frequentist results: for the classical linear regression model, $f(\beta|\text{data})$ takes a multivariate t distribution with $n-k$ degrees of freedom (Gelman et al., 2003, pp. 355-357). The software therefore uses a t distribution for the posterior when computing c^* in these cases. For generalized linear models with this prior estimated via maximum likelihood, $f(\beta|\text{data})$ is asymptotically normal as $n \to \infty$ (Gelfand and Ghosh, 2000, pp. 4-8), and therefore the software uses the normal distribution in these instances.

Computing c^* is a matter of finding the roots (in c^*) of equation 9. This class of problem is already solved via iterative maximization algorithms, such as Newton-Raphson,[14] but we are required to compute an integral at each iteration. For results with a t-distributed or normally distributed posterior, we can use standard quadrature approaches to quickly compute the integral for each step of the maximization process.

Some analysts may wish to impose stronger prior beliefs, which will change the distri-

[13]The software for both R and Stata is available at my website, `http://userwww.service.emory.edu/~jesarey/`. Nathan Danneman ported the R code into Stata.

[14]Many maximization algorithms involve finding a root of $\frac{df(x)}{dx}$ to find a maximum of $f(x)$. The root-finding capability of these algorithms is easily adapted to non-maximization root solutions.

bution of $f(\beta|\text{data})$. If the prior is conjugate with the posterior, the analyst must replace the multivariate t or normal distributions with the appropriate posterior distribution before calculating c^*. For more exotic posteriors, where $f(\beta|\text{data})$ is difficult to analytically express, an approximation can be computed via Markov Chain Monte Carlo methods, stored, and then used in the integration of equation 9.

Sensitivity Analysis

As alluded to in the prior section, the conclusion that a researcher draws about substantive significance will depend on the γ or δ coefficient that s/he chooses at the start. A reader might therefore ask how sensitive the inference drawn is to the choice of this parameter, and whether that sensitivity can be succinctly reported post-estimation. Fortunately, it is straightforward for an analyst to determine the sensitivity of his/her inference to the choice of γ/δ.

To demonstrate, let's return to the example provided by Romer and Romer (1999) from the introduction. Recall that Romer and Romer found that every one percent increase in inflation rates is associated with a 5.71% decline in the income of the poorest fifth of the population, with a standard error of 2.93%. Is this effect substantively significant? For the example, I will adopt the loss aversion framework and set $\gamma = 2$, meaning that incorrectly accepting the substantive significance of a directional relationship is four times more important than correctly drawing that conclusion. In this setting, $c^* = -4.07$, meaning that if we think that a 4.07% decline in the income of the poorest fifth of the population for every one percent increase in inflation is substantively significant, then we should accept this alternative hypothesis of a negative and substantively significant relationship between inflation and the income of the poor.

But how robust is this inference to the choice of γ? To answer the question, we can choose multiple values of γ ranging between 1 (the minimum) and some large value (say, 4), compute c^* for each of these values of γ, and plot the resulting relationship to determine how

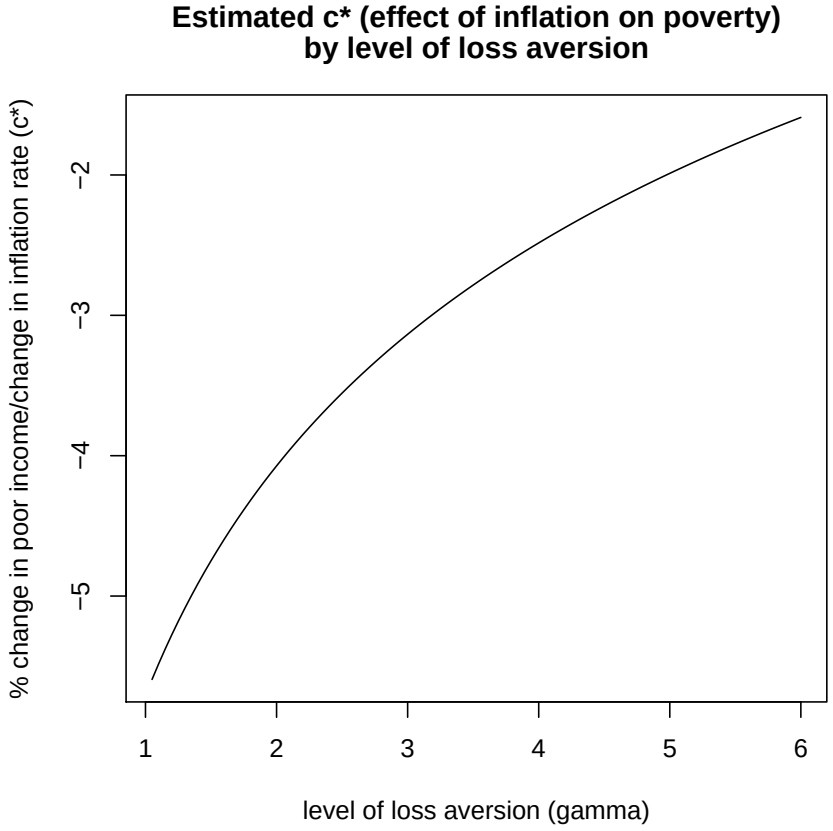

Figure 2: Sensitivity Analysis for Romer and Romer (1999)

inference would change as degree of risk aversion changed. The result is depicted in Figure 2.

For small changes in $\gamma \approx 2$, the value of c^* changes little: it hovers around -4, indicating that those who believe that mistakenly accepting a substantively insignificant effect is about four times as important as correctly accepting such an effect should accept the conclusion that tighter inflation policies help the poor if a $\approx 4\%$ change in income is substantively meaningful. Indeed, even those who are half as risk averse (that is, those for whom mistaken acceptance is only twice as important as correct acceptance), $c^* \approx 4.9\%$—reasonably close to the original calculation.

c^* is not a random variable

It is important to note that, unlike other statistics that political scientists are used to computing (like $\hat{\beta}$ or a t-ratio), c^* is not a random variable and does not take a distribution. In the Bayesian framework for inference (and in contrast to frequentist ways of thinking), any uncertainty in the estimates is encapsulated in $f(\beta|\text{data})$, the posterior distribution of β given observed data. The data are considered fixed, and as a result, we can consider the form of $f(\beta|\text{data})$ fixed contingent on some prior $f(\beta)$: indeed, the posterior is a deterministic function given by Bayes' rule. Calculation of the c^* statistic using equation 9 integrates out uncertainty over the unknown quantity, β, using $f(\beta|\text{data})$. Thus, c^* is a fixed characteristic of the posterior distribution the same way that the mode of that posterior distribution is also a fixed statistic (given a data set and a prior).

An intuitive argument for why c^* is not random can be gleaned from the earlier analogy between betting and statistical inference. The price at which accepting a bet is indifferent to refusing it is not random, despite the fact that the outcome of a bet is random. A bet either makes sense given the expected value of its (randomly determined) payoffs, or it does not: it is a money-making proposition in expectation, or it isn't. Likewise, c^* is meant to represent the smallest definition of substantive significance that can be sustained from a particular dataset and a set of prior beliefs. It represents the cutpoint between when the optimal decision in expectation is to embrace the alternative hypothesis (c^* is large enough to be considered substantively significant) or to stick with the null (it isn't). This c^* cutpoint is fixed for the same reason that the maximum profitable price of a bet is fixed. By contrast, a coefficient β is uncertain—it is the outcome of the bet, in this analogy—and this uncertainty is encapsulated in $f(\beta|\text{data})$.

Loss aversion or risk aversion: which approach to use?

To this point, I have given equal emphasis to the risk and loss aversion frameworks in developing the inference procedure. Both can be implemented into the critical test described

above, and risk averse utility functions have been widely employed in theoretical models of behavior. However, I believe that there are at least three reasons to favor the simple loss aversion framework when making statistical inferences.

First, I believe that loss aversion is a better analog for the decision process that researchers should use in making statistical inferences. When making decisions with data, a scientist should be interested in (1) determining whether the evidence supports the conclusion that a substantively significant relationship exists and (2) minimizing error of any kind, but (3) trading off less Type I error in favor of more Type II error. The third goal requires researchers to value the possibility of mistaken predictions more than that of correct predictions, as in loss aversion, but does not imply that they should undervalue large correct predictions compared to smaller correct predictions, as a risk averse person would.

Second, the CARA utility function presents scaling issues that are not easily avoided and whose correction would create further problems. The utility value of equation 8 is not independent of the scale of β: doubling the value of β (by for example doubling the scale of the independent variable Y) would change the utility value and the inference drawn from a set of data. The loss aversion framework faces no such difficulty.

To illustrate, suppose that a researcher with $\delta = 0.5$ found a point-mass posterior distribution with a 10% chance of $\beta = -0.1$ and a 90% chance of $\beta = 0.2$. For this situation, the threshold for substantive significance is fairly large, over 3/4 the size of the estimated positive β ($c^* = 0.168$). But if the researcher simply increases the scale of the variables such that there is now a 10% chance of $\beta = -4$ and a 90% chance of $\beta = 8$, the necessary threshold for substantive significance drops to only about 7/100 the size of the positive β ($c^* = 0.561$), a much more demanding standard!

In order to remedy this shortcoming, it is necessary to standardize β coefficients by adjusting the Xs and Ys to have a mean of 0 and a standard deviation of 1 when using a risk aversion approach.[15] This standardization puts all analyses on the same playing field

[15]One could also use the constant relative risk aversion (CRRA) utility function to solve this problem, but this choice creates a new problem: the CRRA function cannot handle coefficients less than or equal to zero.

with respect to inference, regardless of the scale of their variables: specifically, inferences will be drawn according to the proportion of the total possible variation in Y that a change in X can cause. But standardization of variables can cause a new set of problems (King, 1986), problems that can be avoided by simply using the loss aversion framework to draw inferences.

Finally, the loss aversion framework is much simpler to understand, use, and present than the risk aversion framework without compromising the core benefits of the approach. As the previous section makes clear, the γ coefficient that a researcher chooses to set his/her level of loss aversion has a straightforward interpretation: it is the extent to which the researcher values losses more than gains. By contrast, a special procedure is necessary for a researcher to choose the appropriate CARA risk coefficient δ.

For these reasons, I believe that researchers should employ the loss aversion framework and the c^* critical test statistic to draw inferences from empirical results. In the following section, I will demonstrate how this can be done using the software packages that I provide.

Inference with c^*: a brief demonstration

The c^* statistic can be immediately applied to interpret existing research, as I will show using two recently published articles from prominent general interest journals of the discipline. The purpose of this exercise is not to demonstrate an error in these researchers' analysis or to single them out for criticism, as there is nothing incorrect or unique about computing p-values. I do want to demonstrate that c^* provides additional, useful information that can enhance our understanding of statistical results. The replication should underscore how c^* focuses researchers on the substantive meaning of their coefficients, while still accounting for the uncertainty intrinsic to any statistical estimate. In addition, the analysis conclusively demonstrates that statistical significance is neither necessary nor sufficient to show

the substantive significance of a finding.[16]

Clinton study of representation

First, I will re-examine some of the critical results from Joshua Clinton's 2006 article on representation in Congress in the *Journal of Politics* (Clinton, 2006). In this article, Clinton examines the relationship between a survey-based ideology measure of residents of American congressional districts in the year 2000 and legislator ideal points estimated on the basis of voting records.[17] In his OLS regression analysis, Clinton finds that there is a positive relationship between the conservatism of a Republican legislator's voting record and the degree of conservatism expressed by his/her Republican constituents. The same relationship exists between a Republican legislator's record and their Democratic constituents' ideology. The conservatism of Democratic legislators' records, by contrast, is associated with the conservatism of their Republican, but *not* Democratic, constituents.

While Clinton uses both OLS regression and an errors-in-variables regression designed to correct for measurement errors, I focus on the OLS results in this replication.[18] I perform the same OLS regression that Clinton ran, calculating a c^* for a loss-averse researcher with $\gamma = 2$. This γ means a researcher believes that falsely concluding that a directional relationship is substantively significant is four times worse than correctly drawing that conclusion. The results are shown in Table 1.

First, I examine the c^* based on the loss averse utility function for same party constituent ideology for Democrats. In order to accept that the conclusion that increases in the conservatism of Democratic constituents tend to substantively increase the conservatism of a Democratic legislator's voting record, we would need to believe that a coefficient of .00932 or larger was a substantively significant effect. I doubt that many researchers would believe

[16]Note that both of these statements can be proven by example: to show that p is not necessary for q, one example of q and $\neg p$ suffices, while to show that p is not sufficient for q, one example of p and $\neg q$ suffices. Thus, the demonstrations to follow do constitute a formal proof of the propositions.

[17]The legislators ideal points are estimated in Clinton, Jackman, and Rivers (2004).

[18]The c^* approach is easily applied to EIV regression, but describing this technique would distract from the central purpose of the present article.

Table 1: Constituent influence on "key votes" in Congress, Table 3 from Clinton (2006)

	OLS Rep.			OLS Dem.		
	β	s.e.	loss c^*	β	s.e.	loss c^*
Wgt. Same Party Avg. Ideology	1.614*	.4214	1.382	.1907	.3290	.00932
Wgt. Different Party Avg. Ideology	.6997*	.3632	.4995	2.167*	.3241	1.988
Constant	.4602*	.1182		-.1.145*	.0919	

Dependent variable = legislator ideology score from Clinton, Jackman and Rivers (2004). $N = 222$ (Republicans) and 210 (Democrats). $R^2 = 0.09$ (Republicans) and 0.26 (Democrats). A * indicates statistical significance, $\alpha = 0.05$ (one-tailed).

this effect to be substantively important, leading me to reject the hypothesis of substantive significance. In this case, the coefficient's substantive insignificance matches its statistical insignificance.

The c^* statistic for Republican constituents' effect on Democratic legislators is more subtle: in this case, we need to believe that a one point increase in constituent conservatism being associated with a 1.989 point increase in a legislator's ideology score is substantively significant. Because the independent and dependent variables are both attitudinal measures, relative changes may be more informative than raw scores: if constituent conservatism changed by one standard deviation (about .11 of a point on a 5 point survey ideology scale), a Democrat at the median of his/her party would move to the 67th percentile under this change. If that effect passes the bar of substantive significance, then we should accept this evidence as supporting a substantively significant positive relationship between the two variables.

For Republican legislators, both c^* statistics are smaller than the c^* of 1.989 above. Consider the calculated substantive significance threshold for the effect of Democratic constituents on Republican legislators: we must decide whether it is substantively significant that a one point increase in Democratic constituent conservatism is associated with a .5003 point change in Republican legislator ideology. If Democratic constituents increased in conservatism by one standard deviation (about .09 of a point on a 5 point survey ideology scale), a Republican at the median of his/her party would move to the 52nd percentile under this

change. Although this coefficient is statistically significant (distinguishable from zero), I doubt that this effect could be classified as substantively significant. Indeed, when estimating the errors-in-variables regression, Clinton also discards this coefficient (on the basis of statistical significance).

In all these cases, the inferences I draw are a combination of (1) the evidence provided by Clinton, (2) the fact that I value losses four times as much as gains, and (3) my judgment of how large an effect must be before it is substantively significant. Other readers may have different thresholds for substantive significance, thresholds that we could debate on substantive grounds using illustrative examples. Including the c^* values in the table allows these readers to come to their own conclusions using this evidence (much as p-values allow readers to draw different conclusions if they wish to use an $\alpha \neq 0.05$). Most importantly, the criteria that informed my decision are clear, and the consequences for using different criteria are easily and quantitatively established.

Baek study of political communication and voter turnout

Next, I will re-examine the findings of Baek's cross-national study of links between political communication and voter turnout (Baek, 2009), recently published in the *American Journal of Political Science*. Baek hypothesizes that "information rich environments promote political engagement and participation by lowering information costs for the electorate" (p. 377). He seeks to test this hypothesis by determining the relationship between average voter turnout rates and measures of information availability in a cross-section of countries. Among other predictions, Baek believes that the presence of a publicly-owned broadcasting system will boost turnout because it provides a greater quantity of substantive news and political coverage than a privately owned, advertising-driven system. The results from one of his statistical models is depicted in Table 2, including the loss aversion based c^* coefficients for this regression (using $\gamma = 2$).

The variable *Public Audience Share*, which measures the proportion (0-100) of the televi-

Table 2: Political communication and voter turnout, Table 1 (Model 3) from Baek (2009)

	β	s.e.	loss c^*
Degree of Democracy	1.92*	0.77	1.49
Compulsory Voting	20.19*	3.46	18.26
District Magnitude	0.06*	0.03	0.04
Unicameralism	5.37*	2.63	3.90
Human Development Index	0.25	0.18	0.15
Socially Owned Enterprises	0.16	0.12	0.09
Free TV Time	14.77*	4.36	12.34
Campaign Funding Limits	-3.47	2.49	-2.08
Public Direct Funding	2.44	2.91	0.82
# of Newspaper Subscribers	-0.00	0.01	0.00
Partisan Press	0.51	2.87	0.00
Access to Paid TV Advertising	1.22	2.86	0.00
Public Audience Share	0.15*	0.05	0.12
Constant	-1.60	16.98	

Dependent variable = average voter turnout (0-100) between 1995-2004. N=66, $R^2 = 0.57$. A * indicates statistical significance, $\alpha = 0.05$ (two-tailed). Robust standard errors reported.

sion audience captured by public broadcasting stations, is positively associated with turnout. Note the c^* value of .12: if we believe that a .12 percentage point increase in turnout for every 1 percentage point increase in public audience share is a substantively meaningful change, then we can accept that there is a substantively significant positive relationship between public audience share and voter turnout.

Yet despite its statistical insignificance,[19] the *Campaign Funding Limits* dummy variable (= 1 if a country puts legal limits on campaign contributions or spending) presents a $c^* = -2.08$. That is, we should accept the existence of a substantively meaningful negative relationship between the presence of legal restrictions on campaign spending and voter turnout if we believe that a 2.08 percentage point decline in voter turnout in the presence of these restrictions is a substantively meaningful difference. This difference in voter turnout is roughly on the order of the effect of *Public Audience Share*: comparing their c^* values, the presence of campaign finance laws is equivalent to a 17.33 percentage point increase in

[19]Two-tailed t-tests are reported in the table, as in the original article, but this coefficient is also insignificant in a one-tailed test ($p = 0.1045$).

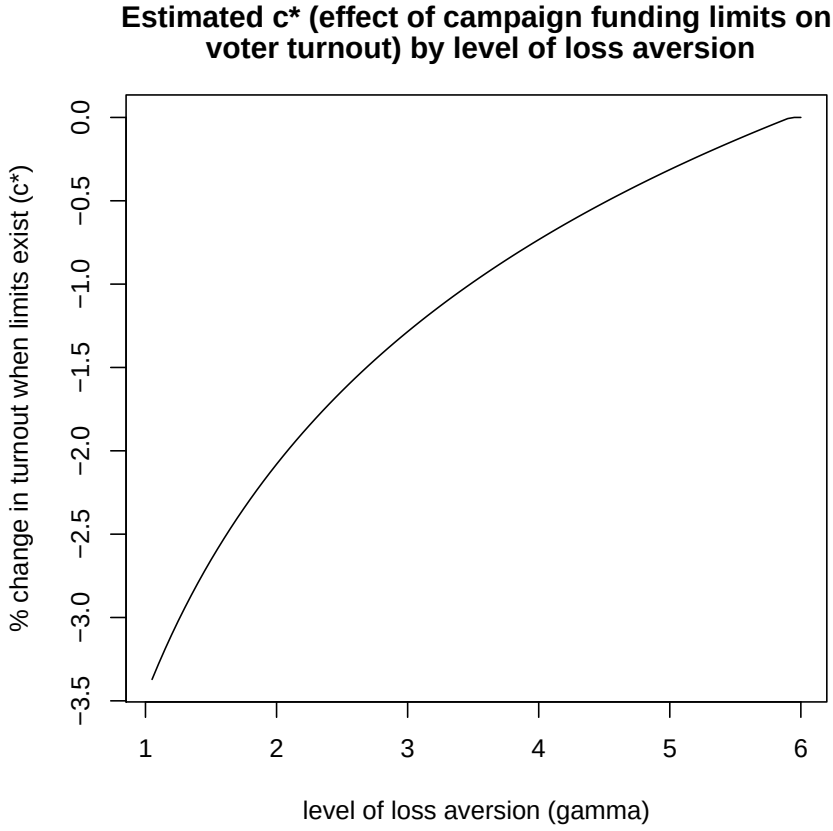

Figure 3: Sensitivity Analysis for Baek (2009)

the public audience. Although belief in the positive association between *Campaign Funding Limits* and voter turnout is a higher-risk, higher-reward bet, the c^* threshold indicates that we should probably accept this bet if we are willing to accept the bet presented by the effect of *Public Audience Share*.

Is our conclusion about the substantive significance of campaign funding limits sensitive to our level of loss aversion, γ? The sensitivity analysis depicted in Figure 3 indicates that the effect is almost certainly substantively significant for reasonable levels of loss aversion: only when mistakenly accepted conclusions are over 12 times as important as correctly accepted ones does the estimated c^* drop below a 1% effect. Thus, we can be reasonably confident that most researchers would accept the existence of a substantively significant relationship between campaign funding limits and voter turnout.

Lessons from replication

The replication analysis illustrates some important characteristics of the c^* statistic. First, it shows that statistical significance is neither necessary nor sufficient to indicate that a coefficient is substantively significant. Political scientists are quite used to the idea that statistical significance is not enough to demonstrate substantive meaning, and they routinely supplement t-tests and confidence intervals with illustrations designed to show the reader the important consequences of their finding. It may be more surprising for political scientists to see a substantively significant coefficient that cannot pass a conventional t-test for statistical significance. The relationship between legal limits on campaign finance and voter turnout is statistically insignificant in Baek's study. It is true that the belief in this relationship is riskier than a belief in the relationship between public television's audience share and voter turnout. Yet even accounting for this additional risk—and a healthy aversion to mistakenly accepted relationships—it is probably still rational to accept this relationship and make future research and policy decisions accordingly.

Most importantly, the replication analysis shows the potential for c^* to facilitate clear and substantively-rooted discussions about empirical findings. For example, there may be debate over whether campaign finance laws actually have an effect on voter turnout. The process of generating a c^* statistic can help make the nature of the disagreement clearer. Not all people have the same threshold for substantive significance: perhaps a 2.08 percentage point decline in voter turnout is not large enough to be relevant to some. In this case, each side can make arguments about the practical political consequences of a 2 point drop in turnout. On the other hand, some people may be more loss-averse than others even if they agree on the threshold for substantive significance; for a person who values losses eight times more than gains ($\gamma = 2\sqrt{2}$), a 1.40 percentage point decline in voter turnout must be substantively significant in order to believe in the connection to campaign finance restrictions. In this case, the discussion should be about the right tradeoff between type I and type II errors in this particular context. Such discussion can be facilitated by sensitivity analyses like that of

Figure 3.

Conclusion

In this paper, I have illustrated that Bayesian statistical decision theory can help political scientists structure and standardize their reasoning about substantive significance, and that the c^* statistic facilitates this process. I recommend that empirical researchers supplement their t-tests and confidence intervals with calculated c^* statistics based on a conservative loss aversion coefficient (such as the $\gamma = 2$ that I used in this paper). The additional work required to implement this recommendation is trivial: with the R and Stata packages that I have made available, calculating the coefficient is a one-line command that takes only a moment to execute. In return, the researcher gets a quantitative indicator of whether his/her data provides persuasive evidence of a substantively meaningful parameter estimate.

I suspect that the use of this statistic could, at least at first, create debate about the appropriate level of loss aversion γ that researchers should employ when testing hypotheses about substantive significance. But in many cases, such as in the examples above, the sensitivity analysis that I prescribe suggests that results are robust to a reasonable level of disagreement about γ. In any event, the judicious use and reporting of that sensitivity analysis allows researchers to draw their own conclusions without imposing a level of risk aversion on them.

I also expect authors to make extended arguments about the substantive importance of their results, and especially about whether a particular value of c^* is large enough to be substantively meaningful. But researchers have always had to make judgments about the substantive significance of their results; with c^*, these judgments are just made more transparent and communicable. It is easier to determine the precise contributions of loss aversion, coefficient uncertainty, and coefficient magnitude to the overall judgment of substantive significance, and to ask how the judgment would change if any of these factors were

different. In short, the c^* statistic helps us bring arguments about statistical evidence back onto substantive grounds. It clarifies precisely how (and how much) disputants diagree and allows readers to form their own opinion based on their own aversion to false positives.

Some important work remains for future research. Many quantities of interest cannot be read directly off of a coefficient table, such as the marginal effect of an independent variable on the change in dependent variable probability in a logit or probit model. These quantities must be computed using a software package such as Clarify (King, Tomz and Wittenberg, 2000). The marginal effect of variables tied up in interaction terms (and especially the variation around these marginal effects) must also be separately computed (Brambor, Clark and Golder, 2006). In both of these cases, it would be useful to determine whether the calculated marginal effects were substantively significant. While there is no reason that a c^* cannot be computed for these effects, a specialized software package must be written to compute these effects.

References

Baek, Mijeong. 2009. "A Comparative Analysis of Political Communication Systems and Voter Turnout." *American Journal of Political Science* 53:376–393.

Brambor, Thomas, William Clark and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14:63–82.

Clinton, Joshua. 2006. "Representation in Congress: Constituents and Roll Calls in the 106th House." *Journal of Politics* 68(2):397–409.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98:355–370.

DeGroot, Morris. 2004 {1970}. *Optimal Statistical Decisions*. Wiley Interscience.

Gelfand, Alan E. and Malay Ghosh. 2000. Generalized Linear Models: A Bayesian View. In *Generalized Linear Models: A Bayesian Perspective*, ed. Sujit K. Ghosh and Bani K. Mallick. Marcel Dekker chapter 1, pp. 3–22.

Gelman, Andrew, Cristian Pasarica and Rahul Dodhia. 2002. "Let's Practice What We Preach: Turning Tables into Graphs." *The American Statistician* 56(2):121–130.

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 2003. *Bayesian Data Analysis, Second Edition*. 2 ed. Chapman & Hall/CRC.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3):647–674.

Holt, Charles A. and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92:1644–1655.

Kastellec, Jonathan and Eduardo Leoni. 2007. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics* 5(4):755–771.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30:666–687.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44:347–361.

Lunt, Peter. 2004. "The Significance of the Significance Test Controversy: Comments on 'Size Matters'." *The Journal of Socio-Economics* 33:559–564.

Manski, Charles F. 2007. *Identification for Prediction and Decision*. Harvard University Press.

McCloskey, Deirdre. 1998. *The Rhetoric of Economics*. University of Wisconsin Press.

Miller, Jane E. 2008. Interpreting the substantive significance of multivariable regression coefficients. In *2008 Proceedings of the American Statistical Association, Statistical Education Section*. URL: `http://policy.rutgers.edu/faculty/miller/2008regression_coefficients.pdf`.

Pratt, John W., Howard Raiffa and Robert Schlaifer. 1996. *Introduction to Statistical Decision Theory*. MIT Press.

Robert, Christian P. 2001. *The Bayesian Choice*. 2nd ed. Springer Verlag.

Romer, Christina D. and David H. Romer. 1999. "Monetary Policy and the Well-Being of the Poor." *Economic Review of the Federal Reserve Bank of Kansas City* 1:21–49.

Schervish, Mark. 1995. *Theory of Statistics*. Springer.

Tversky, Amos and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *Quarterly Journal of Economics* 106(4):1039–1061.

Wald, Abraham. 1950. *Statistical Decision Functions*. Wiley.

Ziliak, Steven and Deirdre McCloskey. 2004. "Size Matters: The Standard Error of Regressions in the American Economic Review." *The Journal of Socio-Economics* 33:527–546.

Ziliak, Steven and Deirdre McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press.