

Why Theory Choosing is Better than Hypothesis Testing

Donald Wittman

Department of Economics, University of California, Santa Cruz
(eMail: wittman@ucsc.edu)

Abstract Hypothesis testing is the standard approach used in scientific work, but it is the wrong methodology to use in choosing which theory most accurately describes the data. This is because hypothesis testing is asymmetric and does not specify the alternative hypothesis in sufficient detail. A more appropriate methodology is to engage in model choice where the contending theories are treated symmetrically. This chapter first explains the problems with hypothesis testing, then provides the methodology of theory choice, and finally provides examples where hypothesis testing leads to inappropriate conclusions.

Keywords model selection, theory choosing, hypothesis testing

1. Introduction

Classical hypothesis testing is the name of the game in economics and in science, more generally. But in most cases, scientists are choosing theories: is the universe ever-expanding or will it collapse; which will get us out of the economic recession – lower taxes or greater government expenditures; and are financial markets characterized by rational expectations or irrational exuberance. Classical hypothesis testing is ill-suited to such a task. In this chapter, I first point out problems with hypothesis testing and then suggest theory choosing as a more satisfactory approach. The argument proceeds via a series of examples.

2. Coin Tossing

I start with a coin-tossing example. Suppose that Person A claims that a coin is fair and person B claims that the coin is weighted so that it will fall on heads 90 percent of the time. We can let A and B stand for two theories. Suppose that in four tosses of the coin it landed on heads three times. The likelihood of this happening if the coin is fair is $[4!/(3!1!)](.5)^3(1 - .5)^1 = 4(.0625)$; the likelihood of this happening if the coin has a 90 percent chance of landing on heads is $4(.9)^3(.1) = 4(.0729)$. In this simple example (and assuming that the loss functions are the same), theory B would be chosen over theory A as theory B is more likely. The likelihood ratio of A to B would be $.0625/.0729$. Clearly, if person A had instead thought the coin had a $3/4$ chance of landing on heads, then the data would have been more consistent with theory A and theory A would have been chosen. This is what I mean by theory choosing. One chooses the theory that is a better predictor. The likelihood ratio test also provides a measure of how confident one should feel in one's choice of theory.¹

Now let us consider hypothesis testing. Suppose that you believe the coin is unfair (the alternative hypothesis). You then set up the null hypothesis: probability of heads equals $.5$. Rejecting the null hypothesis that the coin is fair leads you to the conclusion that the coin is unfair. But unfair means that the probability of heads is not equal to one half. There are two fundamental problems with this approach. (1) 'Not evenly balanced' is not a predictive theory because (almost) every possible set of outcomes is consistent with the coin not being evenly balanced. We are comparing a point hypothesis to two intervals, one on each side of one half. The alternative hypothesis is useless as a theory as it does not limit the set of possible predictions. (2) Furthermore, the test is asymmetric. We really do not directly consider the probability of the alternative hypothesis being right. Only if we reject the null hypothesis do we accept the alternative hypothesis. But if we must go with one theory, then we have to choose which is best by treating both hypotheses in the same way.

I will concentrate on point 2 as point 1 will be partially resolved in dealing with the first point. In hypothesis testing we see how likely one would observe heads on 3 out of 4 tosses if $P = .5$. If it is unlikely, then we reject the hypothesis that $P = .5$ (at some level of significance) and accept the hypothesis that $P \neq .5$ (at that level of significance). Note that hypothesis testing does not ask what the probability of observing $3/4$ is if the coin is

¹ Note that this approach is different from the standard likelihood-ratio test, which compares the likelihood of the theory (say A) being true to the maximum likelihood given the data. For an introduction to the methodology outlined here, see Winkler (2002).

not evenly balanced. In order to make the approach more symmetric we go back to the likelihood ratio test approach employed earlier. The likelihood of $P = .5$ when heads comes up 3 out of 4 times has already been calculated as $4(.0625)$. But what is the likelihood of heads coming up when $P \neq .5$? To answer this question more structure is needed than is provided in hypothesis testing. The believer in the theory that the coin is unfair needs to also have a theory regarding the distribution of unfair probabilities. That is, ‘the coin is not evenly balanced’ is too vague of a theory to test properly without some notion of the probability distribution. Here, I will assume that the distribution of possible values of P is uniformly distributed on $[0, 1]$, which is consistent with no additional knowledge regarding P.²

Using the same numerical example as before, the mean likelihood if the coin is unfair would be:

$$4 \int_0^1 P^3 [1 - P]^1 dP = 4 \int_0^1 \left(\frac{P^4}{4} - \frac{P^5}{5} \right) = 4 \left(\frac{5}{20} - \frac{4}{20} \right) = 4(.05).$$

The mean likelihood ratio of fair to unfair is $625/500$. So, given this data, we choose the ‘coin is fair’ theory as it is more likely.

Note that choosing theories is about choosing the theory that fits the data the best, while hypothesis testing is about finding data that disconfirms the null hypothesis. Of course, likelihood ratio tests are used for hypothesis testing, but in such cases, the likelihood ratio compares the likelihood of the data given the hypothesis to the maximum likelihood given the data. Here the likelihood ratio is determined by the likelihood of one theory to the likelihood of the other (or equivalently the likelihood ratio of theory A to the maximum likelihood is compared to the likelihood ratio of theory B to the maximum likelihood).

3. Stock Markets

Rational expectations theory argues that stock markets are martingales. More formally, $x_t = \alpha + \beta x_{t-1} + \gamma x_{t-2} + \varepsilon_t$, where $\alpha = \gamma = 0$ and $\beta = 1$. That is, $x_t = x_{t-1} + \varepsilon_t$. Suppose, that someone held, for want of a better phrase, an accelerator theory, which is formalized as follows: $x_t = 2x_{t-1} - x_{t-2} + \varepsilon_t$. In deciding which theory is the best, one would choose the theory that is most likely; that is, the one that fits the data the best. This is theory choosing, and, in this simple example, the tools for choosing one theory over the other are readily available – one would use a likelihood ratio

²Note that each theory can have its own prior distribution, as is the case here.

test of the two theories. Now if ε_t were binomially distributed, we could proceed as before. However, ε_t is likely to have a continuous distribution, such as the normal, in which case the likelihood of a point is 0. In such cases, we need to generate an interval around the point. This can be done in a variety of ways. The two most likely candidates are: (1) a priori, for example $[\hat{\beta} - 1, \hat{\beta} + 1]$, where $\hat{\beta}$ is the estimate of β ; and (2) ex post by generating a confidence interval from the data equal to 1, 5 or 10% such that there is a .5, 2.5 or 5% chance of the point estimate being on either side of the interval (note how this is the converse of the standard approach to hypothesis testing and confidence intervals).

Typically, one does not have a well-specified alternative to the rational expectations model. Instead, one has a more diffuse alternative, and this is where the trouble begins. First, the null and alternative hypotheses are reversed. Under the standard way of testing hypotheses, the alternative hypothesis is the theoretical construct and the null hypothesis is the contrary hypothesis set up for testing purposes. For rational expectations, the alternative hypothesis, $\beta = 1$, becomes the null hypothesis, and a believer in rational expectations hopes that the null hypothesis is not rejected, which is a very weak result. Ignoring the reversal of the null and alternative hypotheses, let us proceed by pretending that we are sociologists who believe that markets are irrational ($\beta \neq 1$).

I have not done any stock market regressions since the recent stock market crash, but before then I did, and virtually always I would end up rejecting the hypothesis that $\beta = 1$. Ignoring the possibility of performing a joint test over many sets of data (because in many cases, joint tests are not possible as the separate sets of data are not available), this means that I would reject the rational expectations hypothesis and accept the irrational financial market hypothesis even though β might equal .997.

Now the problem is that we are comparing a point to a continuum minus the point. So we are comparing a predictive theory to a theory without predictive power. Indeed, this is the essence of the problem with hypothesis testing. So how should we proceed? The methodology is much the same as outlined in the coin tossing experiment, but here there is an additional wrinkle. It is much harder to specify the range and distribution of possible alternatives in an a priori way as the methodology would demand. To illustrate, if $\beta \neq 1$, what do you the reader think is the set of possible values of β ?³

³ A theory of irrationality would also have to account for the value of γ and the coefficients of the other lagged variables, but we will ignore such problems here.

One approach is to not be *a priori* but instead use the data from a different perspective. Suppose that you have observed the following price series: 3, 1, 5, 4, 2. One could calculate the unbiased estimates of the mean and variance of the observations (not the mean and variance of the changes) and then use the mean and standard deviation as a way of determining the distribution of possible changes from one period to the next.

However one goes about determining the distribution of possible changes from one period to the next, note that such a determination is a requirement for choosing between $\beta \neq 1$ and $\beta = 1$. While I have not proven it, it appears that, for most distributions that are likely to be chosen, theory choice will end up choosing $\beta = 1$ more often than hypothesis testing will not reject $\beta = 1$ (and, as already noted, the latter is an extremely weak result). That is, ordinary hypothesis testing is biased against the efficient market hypothesis.

4. Are Voters Rational?

The misuse of hypothesis testing is not merely hypothetical. In his book, *The Myth of the Rational Voter*, Bryan Caplan argues that voters are irrational because their views on the economy are significantly different in a statistical sense from economists' views, whose views Caplan assumes are rational. Like the hypothesis testing of stock market rationality, Caplan is comparing the rationality point estimate to an ill-defined continuum of possibilities of the irrationality model. Rejecting the rationality hypothesis, he accepts the irrationality hypothesis. It is a very lopsided test as every observation possible is consistent with irrationality theory. At the end of the day, one has to choose whether 'voters are irrational' is a better theory than 'voters are rational' in predicting behavior. And so, the methodology of theory choosing outlined in this paper is preferred. Once again, the method requires the theory of irrationality to explicitly state the distribution of possible irrationalities before a comparison can be made. The results depend on one's priors regarding the distribution of irrationality. Because, diffuse priors are likely, it does appear that ordinary hypothesis testing is biased against rational voter theory.

5. Demand Curves and Other Asymmetric Tests

Ignoring the problem of simultaneity (and statistical methods such as two-stage least squares to correct for it), let us consider the estimation of demand: $\text{Log}(Q) = \text{Log}(\alpha) + \beta \text{Log}(P) + \varepsilon$, where β is elasticity and P is

price.⁴ The standard way of testing downward sloping demand is to have the null hypothesis be $\beta \geq 0$ and the alternative hypothesis be $\beta < 0$. Unfortunately, $\beta \geq 0$ is not really a theoretical hypothesis but a straw-man hypothesis required by the protocol of hypothesis testing. It seems that the most likely alternative theory to the standard economic theory is that people do not pay attention to prices ($\beta = 0$), which we will label as the inattention theory.

In order to choose between the economic theory and the inattention theory, one has to specify the prior distribution for each theory. To move the argument along, assume that the prior distribution for the economic theory is a uniform distribution of β on $[-2, 0]$ and that the prior distribution for the inattention theory is a uniform distribution of β on $[-1, 1]$. If ε has a continuous distribution, we will also need to specify a confidence interval around $\hat{\beta}$. We will assume that the confidence interval is symmetric around the estimated value of β . In such cases, the choice of theory does not depend on the size of the confidence interval (but of course the likelihood ratio does).

In this case, any $-2 \leq \hat{\beta} < -2/3$ will lead one to choose the economic theory over the inattention theory; while any $1 \geq \hat{\beta} > -2/3$ will lead one to choose the inattention theory over the economic theory. So the inattention theory will be chosen over the economic theory even when $\hat{\beta}$ is negative, as long as it is not 'too negative'. When seen from the viewpoint of theory choice, it appears that the standard hypothesis testing approach is biased in favor of accepting the law of downward sloping demand as any $\hat{\beta}$ such that $0 > \hat{\beta} > -2/3$ and $\hat{\beta}$ is statistically different from 0 would reject the hypothesis that $\hat{\beta} \geq 0$ and accept the hypothesis that $\hat{\beta} < 0$.

Once again, the likelihood ratio test provides a standard for how confident one feels with one's choice.

6. Concluding Remarks

Are voters rational or irrational; does watching television make children more or less violent; and do managers maximize the returns to stockholders or do they maximize the size of the firm. In all these examples, the researcher has to choose among competing theories as to which theory explains the facts the best. Faced with such questions, one should use the methodology outlined in this paper rather than engage in hypothesis testing.

⁴ We will not be undertaking the usual convention of multiplying through by -1 to obtain the measure of elasticity.

Nevertheless, hypothesis testing has been the name of the game. This paper is an attempt to change the balance toward theory choice.

References

- Caplan, Bryan (2007) *The Myth of the Rational Voter: Why Democracies Choose Bad Policies*, Princeton University Press.
- Winkler, Robert (2003) *Introduction to Bayesian Inference and Decision, 2nd Edition*, Gainesville, FL: Probabilistic Press.