Statistical Inference in Merger Analysis

Working Paper

Ref: 2004-01

By

Paula Ramada

November 2004

Statistical Inference in Merger Analysis

By

Working Paper

Paula Ramada

November 2004

Contents

Page

Tables & Figures

Page

Abstract

Statistical inference has gained increasing policy relevance in the last couple of decades as antitrust policy and merger analysis raise the weight put on results of a statistical nature, based on econometric fitting of data to models. In the context of merger analysis, statistical inference is often used in the guise of a hypothesis test, where the test is on whether the merger will be harmful to consumers (and/or possibly to other competitors). The aim of this note is to review the methodological choices regarding the use of statistical inference. We focus on the implications of the choices of two main components of a hypothesis test: the choice of the null and alternative hypotheses; and the choice of the level of significance of the test. We discuss the interpretation of resulting outcomes and their implications for policy decision-making.

1 Introduction

The goal of statistical inference in econometrics is to use the principles of statistics to make inferences about observed data. This analysis takes place in one of two frameworks, classical or Bayesian. The overwhelming majority of empirical study in econometrics has been done in the classical framework and this is therefore where our present focus lies. The three main stages of inferential statistics are sampling, estimation, and hypothesis testing. In this note we discuss a number of issues concerning the latter of the three.

Hypothesis testing in merger analysis is often concerned with the test of whether or not the merger will be directly harmful to consumers and perhaps also indirectly through harm to other competitors. The statistical inference approach to this problem is to construct a hypothesis relative to merger impact and then use statistical formulas based on collected data to decide relatively to that hypothesis. Thus, the first important consideration of this process is the formulation of the hypothesis to be tested. The statistical methodology will give an answer relative to a given hypothesis, and the conclusions that can be drawn on that basis are often different from those that would result from a test based on a different hypothesis. In particular, finding strong statistical evidence against a specific hypothesis may not inform the researcher as to what the correct alternative to that hypothesis may be.

The second consideration is with respect to how demanding the researcher will be in order to conclude against the initially formulated hypothesis. Regardless of what the initial hypothesis about merger harm is, we have to decide how much evidence we will require before rejecting that initial hypothesis. The researcher must be aware that a statistical procedure is not without error, and that therefore it is possible to reject the initial hypothesis in error, as indeed it is possible to not reject the initial hypothesis in error. The seriousness of these two errors must be considered and, to the extent possible, they must be minimised.

In this note we will first briefly retrace the epistemology of hypothesis testing and how it fits within the scientific method. The most important difficulty with statistical inference is that the process of going from the knowledge of the particular cases of which we have information to the knowledge of the cases of which we do not is an inductive process. Unlike deductive methods, induction is not conclusive and is subject to a number of problems.

We will proceed with a brief overview of what a hypothesis test is, which are its main component elements and, most importantly, the two types of errors involved.

The two main sections of this note deal with the alternative ways in which we can formulate a testable hypothesis, on the one hand, and with the strength of statistical evidence that we wish to require in order to reject the chosen hypothesis, on the other. Our view is that these two elements are often chosen in a mechanical way and this may result in a sub-optimal use of the tools of statistical inference in merger analysis and anti-trust policy in general.

The ultimate goal of this note is thus to step back from a mechanical use of hypothesis testing and, by looking afresh at its theoretical foundations, understand the implications of alternative modelling choices, interpret the obtained results correctly, and understand the limitations of the method.

2 The foundation of hypothesis testing - falsification versus affirmation

When using empirical observation to make inductive inferences, we have a much greater ability to falsify a principle than to affirm it. One of the most significant contributions to epistemology by the Austro-British philosopher Karl R. Popper has focused precisely on this. Popper sustains that the scientific method does not use inductive reasoning, but rather hypothetical-deductive reasoning. Although the movement from the data evaluating a hypothesis to a conclusion on the latter goes from the specific to the general, i.e. in an inductive direction, induction does not exist as a reasoning process or inference. In other words, there is no method that enables us to infer or to verify hypotheses or theories based on experimentation, or even to conclude that they are very probable in any meaningful way. Popper illustrated the point with this now-classic example: if we observe swan after swan, and each is white, we may infer that all swans are white. We may observe 10,000 swans, all white, and feel confident in our inference. All of this evidence may make us think that we can prove that all swans are white. However, all it takes is a single observation of a non-white swan to disprove the assertion that all swans are white.

This view of the scientific method implies that the task of a researcher is to propose a hypothesis as a tentative solution to a problem, confront the prediction deduced from the hypothesis with actual experience, and evaluate whether the hypothesis is rejected or not by the facts. As theories cannot be verified, we can only maintain them if they withstand our attempts to reject them. Consequently, the test of a theory consists of criticism or a serious attempt at falsification, in order to reject it if it is false.

Based on this theoretical underpinning, a common approach to statistical inference takes the following form: we develop a scientific theory; we construct a null hypothesis, which relates to that theory in some specific way; and then we do our best attempt at disproving it. This process may seem contrived and confusing -- as if we're building a straw man and then knocking him down -- but is in fact logically mandated by the Popperian view described above. All of our current statistical theory, used in hypothesis testing, is based on this approach. The 'P-value' often reported in statistical and econometric work is exactly the mathematical probability that we would find the observed values if the null hypothesis were true. The lower that probability the greater the confidence we have in that we have falsified the null hypothesis.

This epistemological approach is relatively silent as to what an appropriate null hypothesis should be. In light of the above, however, the statistical inference methodology should be understood as something that gives us the possibility to refute statements but not the possibility to confirm statements. This distinction may be difficult in practice, particularly when we are not testing a specific theory but rather searching for some empirical regularity, such as whether a given occurrence in a sample of individuals can be generalised to a larger population. Consider, as an example, a medical experiment where in a sample of 1000 individuals, 500 are treated with a given drug and 500 are given a placebo. Suppose then that we find that in the treated population the incidence of the disease under study was 20% lower. The aim of statistical inference in such a situation could simply be to determine whether such a difference is or not "significant". In other words, does it imply efficacy of the drug under test for a more general population? In statistical terms the question may be phrased as "how likely is it that we would get a 20% better result in the treated group if the drug were in fact *ineffective*?". This phrasing implies that our null hypothesis is that the drug is ineffective and we are trying to assess how strong our evidence against it is.

From a test thus formulated we would conclude either that the impact of the drug is not significantly different from zero or that it is significantly different from zero. We would not, in any event, conclude that the impact of the drug is a 20% decrease in disease incidence.

We could instead formulate our question in the following terms: "how likely is it that we would get a 20% better result in the treated group if the drug were in fact effective?" If 'effective' is defined as a decrease in disease incidence of 20% or more, we would probably get a very high P-value for this test. But, would this be a strong conclusion? The result is no doubt consistent with an effective drug but we would be left wondering how likely would it be to obtain that same result by chance, i.e., even though the drug is in effect useless. Under this second test formulation it would not be possible to answer this question.

In merger analysis, a testable null hypothesis is e.g. "the merger will not have a significant impact on prices". It may seem that it is not possible to reject this statement while at the same time not "accepting" its negation. If we conclude that we reject the hypothesis that a given merger will not have a significant impact on prices, are we not then stating that the merger will have a significant impact on prices? In fact, the two alternative ways of testing will not yield the same conclusions. It we test the hypothesis "the merger has no effect" and reject it we may be reasonably convinced that merger has some effect, or at least we can tell how confident we are that that is the case. However, if we test the hypothesis "the merger has significant price impact" and do not reject it, we cannot be sure with what probability we mistakenly fail to reject a true hypothesis. This point is explained further below.

3 Elements of hypotheses testing in merger analysis

A hypothesis test is a procedure that details how a sample is to be inspected to determine if it agrees reasonably well with a given hypothesis. It is essentially a decision rule that tells us when to reject or not reject the hypothesis. Decision rules are seldom infallible; false hypotheses may be accepted and true hypotheses may be rejected. The classical testing procedure involves the statement of a "null" or maintained hypothesis and an "alternative". These are conventionally denoted H_0 and Ha, respectively. The formulation of the statistical hypotheses is the first step in testing a hypothesis.

The second step is to choose the level of significance. This represents the probability of rejecting a true null hypothesis. The level of significance used may vary, but an often-used level of significance is 5%. This significance level means that there is only a 5% chance of rejecting the null hypothesis when it is true. We use this level of significance to protect us against rejecting a true null hypothesis. The level of significance determines the size of the critical (or rejection) region. The critical region is the set of values for the test statistic for which we will decide to reject the null hypothesis. This region will be larger when the significance level is increased. We do not reject our hypothesis if the test statistic falls within the acceptance region.

There are two types of errors connected with hypotheses testing. A type I error occurs when we reject a H_0 which is true and the significance level of the test, α , is the probability of making a type I error. A type II error occurs when we accept a H_0 which is false. The probability of a type II error is β . The smaller the value of β , the better is the test. Alternatively (1- β), i.e. the probability of rejecting H_0 when it is false (denoted the power of the test), should be as large as possible. The power of the test is the measure of how well the test of hypothesis is working. A low value of (1- β) means that the test is working poorly.

In hypothesis testing both α (the probability of type-I error) and β (the probability of type II error) should be small. Typically, the acceptable level for a type I error, or α , is usually set in advance by a researcher, but the type II error or β for a given test is not always possible to compute as it requires estimation of the distribution of Ha, which is unknown in most of the cases. The power of the test, given by (1- β) is different for each value for which Ha is true. Typically the power will be higher relative to alternatives that are 'far' from the null hypothesis but the power (of the same test) may be low for other values of the alternative hypothesis. This relationship can be summarised by a curve, known as a power curve.

Ideally we would like to make the probability of both types of errors as small as possible. Unfortunately, this is not possible. To illustrate this consider the results of a jury trial viewed as a hypothesis test. The accused is presumed innocent until proven guilty, so the corresponding hypotheses would be:

H_0: The defendant is innocent.

Ha: The defendant is guilty.

In this case a type I error would be sending an innocent defendant to jail. A type II error would be setting a guilty defendant free. If we were to set the burden of proof so high that we were sure that we never sent an innocent person to jail, we would probably end up setting all defendants free. Whereas if one were to set criteria so that a guilty party was never set free we would send quite a few innocent people to jail.

Hypothesis testing thus involves the following steps:

- Specify the Null Hypothesis (H_0) and the Alternative Hypothesis (Ha).
- Compute the appropriate test statistic based on the sample data. The sampling distribution, if the Null Hypothesis were true, is assumed to be known.
- Think of the problem as a decision problem, and consider the relative importance and probability of Type I and Type II errors.
- To control for Type I error choose a significance level α and consider only tests with probability of Type I error no greater than α .
- Among the tests, select one that makes the probability of a Type II error as small as possible (that is, power as large as possible). If this probability is too large, you will have to take a larger sample to reduce the chance of error.
- Compute the Acceptance and the Critical Regions based upon the sampling distribution.
- Reject the Null Hypothesis H_0 if the test statistics falls within the critical region and do not reject otherwise.

4 The null and the alternative hypotheses

The choice of the null and alternative hypotheses is not cast in stone. Rather, it is a significant first step in the process of hypothesis testing and thus deserves more attention than is frequently the case. This section is about the significance of the choice of the null hypothesis. We will discuss this choice in light of two different issues. First, we will look at the correct interpretation of what the test results signify relative to the original question of the researcher. Second we look at the implications of systematically choosing a null hypothesis of "zero" or "no effect", in terms of what will become the more likely conclusions from the procedure.

4.1 The null hypothesis and the interpretation of the test

One of the questions raised by the methodology of hypotheses testing is whether results and conclusions are sensitive to the choice of the null hypothesis. A hypothesis test is a test relative to a given null hypothesis. The test conclusion is whether or not that null is rejected. The test does not provide a conclusion relatively to the alternative hypothesis. This is an important element for the interpretation of a test and it will guide us as to the more appropriate formulation of the null hypothesis. We have to consider, essentially, what hypothesis do we really want to put to a test.

We obtain a powerful result from the exercise of hypothesis testing only if we reject the null hypothesis. If we reject the null, we can do that with a pre-determined degree of confidence. This degree of confidence can be set arbitrarily high. Tests with 95% or 99% degrees of confidence (i.e. levels of significance of 5% and 1%, respectively) are not uncommon. In a 95% test, if we reject the null, we would be 95% "confident" that the null is indeed false. This means that, upon rejection of H_0, H_0 is very unlikely to be true, and we can pre-set exactly how unlikely we want that to be.

What this implies in a merger impact analysis is that if we define H_0 as "the merger has no impact on prices" and we reject that at a high level of confidence, then we will be quite confident that the assertion "the merger will not have an impact on prices" is untrue. Thus, we can reject the hypothesis that the merger will not have an impact on prices with a high degree of confidence. However, we cannot conclude, with a certain defined level of confidence, that the merger will have an impact on prices.

Alternatively, we could be interested in investigating whether we would be able to confidently reject the assertion that the merger will have, say, 5% or more impact on prices. This type of test would lead to a different statement of the results: rejecting H_0, in this case, implies that we are highly confident that it is not the case that the merger will have a price impact of 5% or more.

However, again, we cannot conclude, with a given level of confidence, that the merger will have an impact on prices of less than 5%.

When H_0 is "the merger has no price impact", we are looking for strong statistical backing against the "no impact" scenario. This means that we will require strong statistical evidence as a condition to stopping the merger.

When H_0 is "the merger has an appreciable price impact", we are looking for strong statistical evidence to back a decision of letting the merger go ahead.

Which of the two is the best way to proceed? Clearly it seems that there cannot be a single answer to this question. The answer may vary from case to case and it may also depend on what other information has been gathered. If a market investigation by the competition authorities has lead to the preliminary view that the merger would, in all likelihood, be harmful to consumers, the competition authorities would probably want to obtain very convincing statistical evidence to the contrary before deciding to let the merger go ahead.

Consider, therefore, the situation where the prior of the authorities is that the merger will be harmful, i.e. the non-statistical evidence points to a decision of not letting the merger go through. If H_0 is "the merger is harmful", a rejection of the null would point to allowing the merger (contrary to the previously held opinion). If we reject this H_0 in a, say, 95% test, it implies that the statistical evidence is strongly against the assertion that the merger is harmful.

On the other hand, if the prior is that the merger is harmful but the null hypothesis is "merger not harmful", then the test would never result in taking a decision contrary to the initial prior. If we maintain H_0 we do not know with what degree of confidence we are doing that so that we would not be able to use that result as a basis of a policy decision. If we reject H_0 we would find additional evidence in favour of the initial prior. The process thus described does not therefore put our prior to the test. We are never allowing for the possibility of strongly rejecting our prior. It's an exercise in "affirmation", i.e. it may contribute one more piece of evidence in favour of a given view, but it is overall inconclusive in that it will never make us change our initial view.

In general, when the conclusion from hypothesis testing is that we do not reject H_0, the test has not been helpful, in the sense that it has not helped us forming an opinion either way. There is often the somewhat misguided view that a test that does not reject H_0 does therefore lend support to H_0. This, however, is not correct. A statement such as "the test gives evidence in favour of the null hypothesis" is often an incorrect way of stating the result because this statement can generally not be made precise in the sense that we cannot actually compute the probability with which we would mistakenly fail to reject a false null.

4.2 Null hypotheses and decision making: an example

In this sub-section we illustrate with an example how the choice of the null hypothesis may change the way in which results are reported.

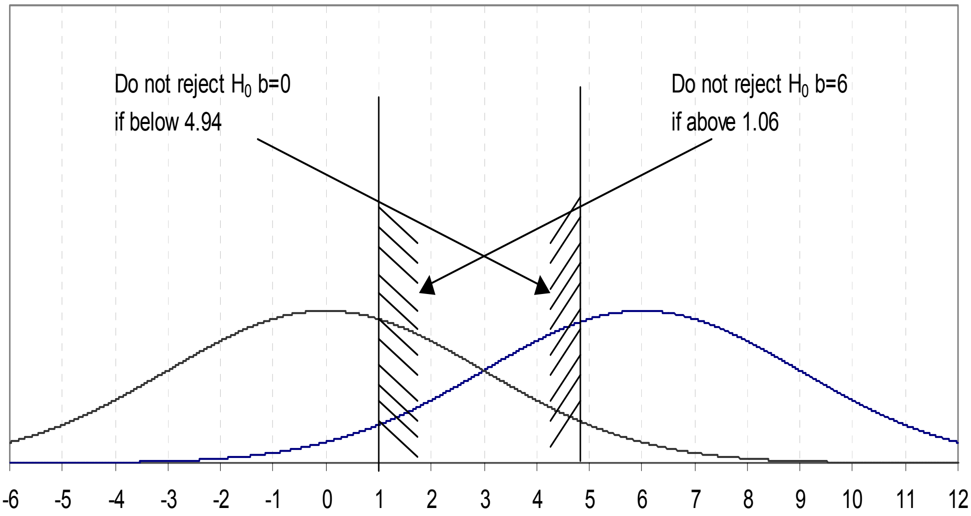
In the figure below we depict the rejection bounds for two different hypotheses testing procedures. In one case (the left hand side distribution) we test the null hypothesis b=0, in the other case (right hand side distribution) we test the null b=6. The estimator, in both cases, is assumed for simplicity to have normal distribution with standard deviation equal to 3.

When we test b=0 we do not reject for parameter estimates below 4.94. When we test b=6 we do not reject for values above 1.06. Thus there is wide range of possible parameter estimates that would lead us to reject neither of the two nulls. There is an interval of parameter estimates, between 1.1 and 4.9, where we would neither reject that b=0 nor that b=6.

If this were a merger case, our parameter estimate could refer to the percentage estimated price impact of the merger. If we started with a null hypothesis that the merger has no price impact, and a parameter estimate in the above range, we would not reject that null. We would be tempted then to consider that we had found econometric and statistical backing to let the merger go through. If however we departed from a null hypothesis that the merger is harmful and that prices will increase about 6% as a result of the merger, and with any estimate in that same range, we would also not reject that hypothesis. This could then lead us to believe that we had obtained some evidence against the merger.

This type of analysis, clearly, is internally inconsistent. It is impossible that an estimate of e.g. 3 is simultaneously evidence that the merger is likely to be harmful and likely not to be harmful. The right interpretation can only be that an estimate of 3 gives evidence in favour of neither of the two hypotheses.

Figure 1: The choice of the null hypothesis: example



The example thus illustrates how a non-rejection of a given null hypothesis may have little practical meaning. We should therefore consider inconclusive the conclusion of "not rejecting H_0".

What we try to point out with this example is that it would seem inappropriate for a decision in a merger case, or indeed in any other context, to be based on a statistical conclusion where we do not reject H_0. This is precisely because not rejecting, in this context, is a statement without statistical force. In particular, there may be a wide range of other hypotheses that would also not be rejected[1].

[1] There is a readily available statistical tool that gives us all the "null hypotheses" that would not be rejected – this is the confidence interval around the parameter point estimate.

5 Significance levels and power of the test

In this section we discuss the trade-off between type I and type II errors and how the level of significance of a given exercise of hypotheses testing should be set to take into account the specific circumstances of each situation.

Researchers do not often appear to devote much attention to the choice of a significance level. The most common approach is that a convention is followed. There is rarely an attempt to balance or even highlight the trade-off implied by this choice.

In practical applications, the most commonly chosen level of significance is perhaps 5%. It is also very commonly the case that the power of the test is not reported. In most cases, in fact, the power cannot be computed because it depends on the true state of the world when the null hypothesis is false. The researcher does not know what the true state of the world is when the null is false. For example if the null is that a given parameter is equal to 2, when this null is false the true value of the parameter may be 10 or it may be 2.01. Holding everything else constant, a statistical test will have more power in the former case than in the latter. Thus, power can be computed relative to relevant values for the alternative hypothesis. In many applications, a whole range of possible values for the parameter, in the event of the null being false, should be considered.

Researchers often set a, say, 5% probability for type I error but will not report the power of the test or the probability of type II error. There are two essential problems with this way of constructing the test. First, the resulting probability of a type II error may be extremely high. Second, that probability is dependent on the type I error probability, and the approach often chosen fails to highlight this trade-off. In fact, the more demanding the test is in terms of a type I error, the worse it will do in terms of type II error. This should concern researchers since it is unlikely that in all possible circumstances type I errors are always “more serious” than type II errors.

A 5% significance test may be a test with very low power. We may have a 5% probability of committing a type I error while having a 20% probability of committing a type II error. In this case we would be 4 times more likely to commit a type II error than a type I error. One would think that, in such a case, it could be acceptable to have a slightly higher chance of a type I error in exchange for a decrease in the probability of type II error. A decision has to be made as to what type of trade-off is acceptable. Such decisions must depend on the consequences, and associated costs, that would result from each of the two types of errors.

In the world of medicine, a null hypothesis might be "this drug will be no more effective than a placebo.” A type I error would be concluding that the drug does work when it actually does not. A type II error would be to conclude that the drug does not work when it actually does. One could argue that a type II error should be minimized here if one agrees the possibility of wasted resources is a low price to pay in the search for a potentially life-saving therapy.

In the legal world, the null hypothesis is generally "this person is innocent." A type I error would be judging the person guilty when he/she is innocent. A type II error would involve declaring the person innocent when he/she is guilty. If one accepts the thought that it is better to release a guilty person than to convict an innocent one, then it would be important to minimize the chances of a type I error. Should this be generalised to all types of legal proceedings though? Rubinfeld (1995) provides the following view:

> "Courts often accept conventional practices of the statistics profession without considering whether such practices are valid in the context of litigation. The most apparent of these practices has been the determination of a statistical level of confidence associated with the burden of persuasion set by a court – preponderance of the evidence, clear and convincing evidence, or proof beyond a reasonable doubt. I have some doubt as to whether a specific level of statistical significance should be attached to a particular burden of persuasion. But I am convinced that if significance levels are to be used, it is inappropriate to set a fixed statistical standard irrespective of the substantive nature of the litigation."[2]

In the context of legal proceedings, type I errors involve the cost of concluding that an activity was illegal when in fact it was not, while type II errors involve the cost of wrongly concluding that an activity was not illegal, when in fact it was. Courts have implicitly understood the relative costs that are imposed by different choices of "burden of persuasion". For example, Justice Harlan, quoted in Rubinfeld (1995), stated:

> "The standard of proof influences the relative frequency of these two types of erroneous outcomes. If, for example, the standard of proof for criminal trial were a preponderance of the evidence rather than proof beyond reasonable doubt, there would be a smaller risk of factual errors that result in freeing guilty persons, but a far greater risk of factual errors that result in convicting the innocent. Because the standard of proof affects the comparative frequency of these two types of erroneous outcomes, the choice of the standard to be applied in a particular kind of litigation should, in a rational world, reflect an assessment of the comparative social disutility of each."

Courts in civil cases ought to acknowledge explicitly that setting the standard for statistical proof involves just such an assessment of comparative social

[2] Rubinfeld, Daniel L.,1995, "Econometrics in the Courtroom", *Columbia Law Review*, vol. 85, HeinOnline --- 85 Colum. L. Rev. 1048 (1985)

costs. The specific details of the proposed analysis will depend in part upon one's appraisal of the relationship between prior information about the liability of the defendant and the information that is presented at trial.

In a merger impact assessment, if we take as a null hypothesis that the merger has no significant price impact, we should then interpret the meaning of type I and type II errors relative to that null and evaluate the seriousness, or make judgement relative to the "cost", of each type of error.

In this context a type I error corresponds to concluding against a merger when indeed that merger would not have had a significant price impact. A type II error corresponds to a judgement that the merger will have no price impact when it will in fact have a significant price impact.

Decreasing the chance of one type of error frequently increases the chance for the other error type. In real-life situations, one can decrease the probability of both error types by collecting more data or having more information available. However, one must frequently decide which type of error should be minimized.

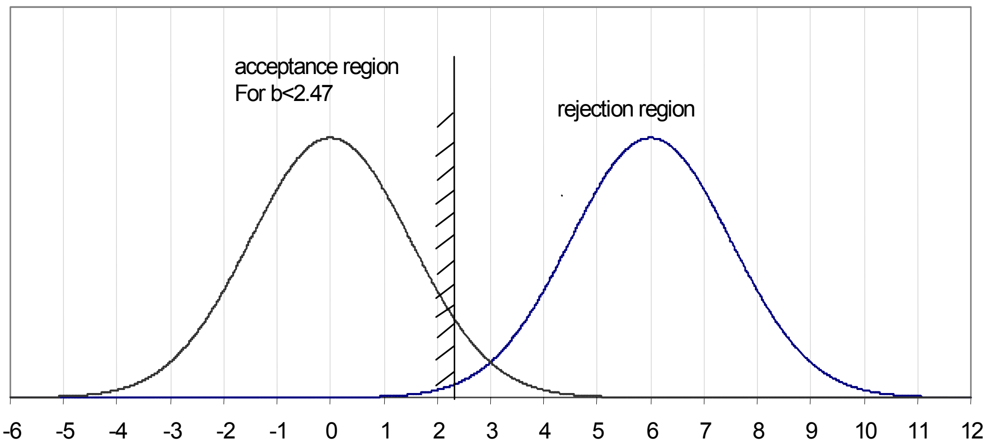
5.1 Type I and type II error trade-off: an example

To illustrate the trade-off that we have been discussing in this section, we provide below an example of possible hypotheses testing situations and the impact of changing the significance level of the test.

To make the example simple we consider a situation where the alternative hypothesis is a number rather than a range (the more common situation). This abstracts from an additional problem in analysing the strength of a hypothesis test, which is the fact that we cannot easily compute the power of the test. The power of the test will vary for different values of the alternative hypothesis.

The figure below depicts the distribution of the test statistic under the null hypothesis and the distribution of the test statistic under the alternative hypothesis. We start with an example where the variance of the estimates is low (standard deviation 1.5 in the first example). The null hypothesis is that the coefficient is equal to zero and the alternative hypothesis is that it is equal to 6. In a 5% one-sided test we reject the null hypothesis for values of the statistic above 2.47.

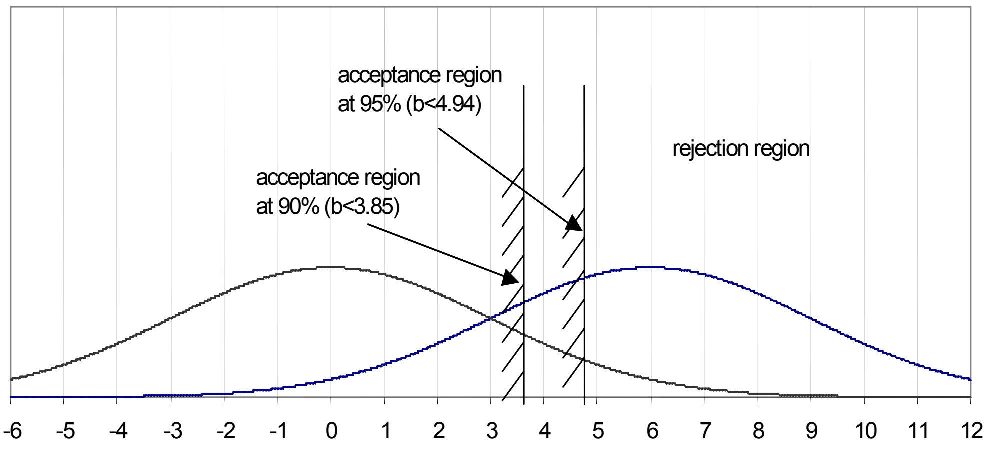
Figure 2: Hypothesis test at 5% with low estimate variance (standard deviation 1.5)



The figure can help us visualise the probability of failing to reject the null when the alternative hypothesis is the true value of the parameter. This corresponds to the area under the alternative hypothesis distribution below the critical value market in the figure. When the alternative hypothesis is true, a test statistic below 2.47 (according to which we would not reject the null) has a probability mass of .0093. This implies that the probability that we fail to reject the null when the null is indeed false is less than 1%. The probability of a type II error is thus 0.9% and the power of the test is 99.1%.

In the figure below we consider the situation where the parameter estimates have higher variance (standard deviation equal to 3 instead of 1.5). We compare a 95% one-sided test, i.e. a test with a probability of a type I error fixed at 5% with the previous figure and with a 10% test, on estimates of the same variance. When the null is given by H_0: b=0, we will reject for values of the test statistic above 4.94. But if the alternative is true, a test statistic below 4.94 has a probability mass of 0.36. This implies that, when the alternative is true, we fail to reject the null with 36% probability. The probability of a type II error is thus 36% and the power of the test is 64%.

Figure 3: Hypothesis test at 5% and 10% with high estimate variance (standard deviation 3)



We have lost power compared to the previous testing situation because, in the present case, we have low precision of our estimates. This can happen if we have insufficient data points or a lot of variation in the sample data. The comparison between the two situations above illustrates the point that, for a given level of significance, i.e. a fixed level of type I error probability, the researcher will, in general, have no control over the power of the test, or the probability of a type II error. What the researcher can do is to evaluate and consider the trade-off between the two types of error and choose a combination that is deemed appropriate for the situation at hand.

It seems, in this test, that a disproportionate weight is being given to type I errors, relative to type II errors. It is 7.2 times more likely to commit a type II error than a type I error. This may be reasonable in some situations but it certainly may not be reasonable in others. We may wonder, then, how much extra power would the test have if we accepted an increase in the probability of type I error. We illustrate this trade-off in the figure, moving the significance level from 5% to 10%.

We consider therefore a 90% one-sided test, i.e. the probability of a type I error is fixed at 10%. When the null is given by H_0:b=0, we will reject for values of the test statistic above 3.85. But if the alternative is true, a test statistic below 3.85 has a probability mass of 0.24, this implies that we will fail to reject the null with about 24% probability, if the alternative is true. The probability of a type II error is thus reduced to 24% and so the power of the test is now 76%.

We have thus illustrated the trade-off between type I and type II errors. By accepting a type I error increase from 5% to 10% we were able to reduce type II error from 36% to 24%.

Whether this is a trade-off that we want to contemplate depends on the relative importance that the researcher puts on type I and type II errors. The situation where we have a probability of type I error of 5% and of type II error of 36% would be acceptable in cases where the perceived costs associated with a type I error are much higher than those for a type II error. In the example, we lost 5% in higher to error I probability and gained 12% in lower error II probability.

This example is very simplified, in order to illustrate the trade-off between the two types of errors. In most applied situations we will not know the distribution of the test statistic under the alternative hypothesis. In a given merger we may take as the null hypothesis that the merger has no impact while the alternative is likely to be that the merger has "some significant impact" but this does not correspond to a number. For example, if we think that under the alternative hypothesis it is equally likely that the price impact of the merger is any number between 3% and 8%, the assessment of the power of the test in this case would be more complicated than what we have seen in the examples presented above.

5.2 Significance levels and the relative costs of errors

The choice of the significance level of the test will result in different combinations of probabilities of error I and error II. Both errors' probabilities will also depend on the variance of the estimators. In some circumstances the availability and quality of the data will be better than others. When there are fewer data and data of a less reliable nature our inference process is more subject to error. If, in all circumstances, we hold fixed the level of error I probability we will have in some cases very large probabilities of error II and very small in other cases.

But our relative concern about error I and error II should in fact not depend on how good our data are. This is a consideration that should be made based on the relative seriousness of the two types of error in this context: letting an anti-competitive merger go through versus preventing a merger that would have caused no consumer detriment.

We may, for example, decide ex-ante that the two types of errors are equally serious. Or that a type I error is "twice" as serious as a type II error. Or indeed we may form any other consideration. After doing so, we should then pose the question about how to reflect these concerns in the choice of the significance level for hypothesis testing.

We provide an example below to illustrate these choices. In the example, we are testing whether a parameter is equal to zero (e.g. merger price impact). We consider three cases with progressively increasing standard errors. The alternative hypothesis implicit in the computation of probabilities of error II is the hypothesis that the parameter is equal to 6.

In order to make the analysis as we exemplify below, the researcher has to make a decision on a level for the alternative hypothesis. This can be done case-by-case, based on some distinguishing feature of the alternative state of the world, i.e. the state of the world where our null hypothesis is not true.

A common formulation for a hypothesis test is e.g. H_0: b=0 H_a: b>0. The alternative hypothesis thus encompasses a very wide range of values. In practical terms, however, we often do not care about the power of a test to distinguish between 0 and 0.1. In the example of a merger case we may be concerned about distinguishing cases where the impact of the merger on prices is zero from cases where this impact is, say 6%. If the true price impact of the merger is 6% and our statistical test is unable to distinguish this from a situation where the impact is zero, then we would be unsatisfied with the test procedure.

In the table below we provide the levels of error I and error II probabilities for different choices of significance levels, for each of the three cases of standard deviation of the estimates.

In the first row we provide the outcome from the standard approach of a fixed 95% confidence level. As we can see, this choice has widely different outcomes in the three cases. In the case where the standard deviation is lowest it results in a possibly excessively low probability of error II, especially when we consider that we could have reduced error I further by increasing confidence to 98% (second line). In contrast, when the standard deviation is high the 95% confidence level results in a probability of error II of 36%, highlighted in the table. With the same confidence level of 95% in the first case we are five times more likely to commit error I than error II while in the other case we 7 times more likely to commit error II than error I. Since our relative preoccupation over the two types of errors should not be dependent on the standard deviation of the estimates (something that depends entirely on data features), it seems unreasonable that a constant confidence level would serve us well in all circumstances.

The example illustrates the implications of the choices of significance level when we put different relative weight on error I and error II. The second row, for example, describes the choice of confidence level that, for each case, would minimise the sum of the probabilities of both errors. This corresponds to a situation where the researcher cares equally about both types of error. We see that as the standard deviation increases we need to become less and less demanding in terms of error I otherwise probability of error II will become very large.

Table 1: Example

	case 1: standard error 1.5				case 2: standard error 2				case 3: standard error 3			
	confidence level	prob. of error I	prob. of error II	Rejection	confidence level	prob. of error I	prob. of error II	Rejection	confidence level	prob. of error I	prob. of error II	Rejection
Standard approach	95%	5%	1%	>2.5	95%	5%	9%	>3.3	95%	5%	**36%**	>4.9
minimise sum of errors	98%	**2%**	**2%**	>3.0	93%	**7%**	**7%**	>3.0	84%	**16%**	**16%**	>3.0
minimise sum with double weight on error I	99%	1%	4%	>3.3	96%	4%	11%	>3.5	91%	9%	25%	>4.0
minimise sum with double weight on error II	96%	4%	1%	>2.7	89%	11%	4%	>2.5	75%	25%	9%	>2.0

As the standard deviation of our estimates increases, we have to accept higher probabilities of both types of error. We have also constructed the confidence levels that should be chosen for two different objectives – in the third row we present the case where we want to minimise a weighted average of the probabilities of the two errors but we care twice as much about error I than we do about error II. In the fourth row we do the converse.

It is interesting to note that in case 1 the 95% level corresponds to a very high relative weight put on error II (more than twice the weight put on error I). While in case 3, a 95% confidence level corresponds to a very high relative weight put on error I (more than twice the weight put on error II). This again illustrates the point that the choice of the significance level should be adapted to the particular features of each situation.

6 Conclusions

In this note we have discussed hypothesis testing in the context of merger impact analysis, focusing on the implications of the choices of two main components of a hypothesis test: the choice of the null and alternative hypotheses; and the choice of the level of significance of the test.

The way in which a hypothesis test about potential merger impact is usually formulated is based on a null hypothesis that the merger will cause no harm combined with a typical significance level of 5%. We have argued that, under likely scenarios, this formulation will too often result in a misleading conclusion that the authorities should let the merger proceed.

We have discussed both components of the hypothesis test: the choice of the null hypothesis and the choice of the significance level.

We resorted to the epistemology of hypotheses testing to look for guidance in the choice and interpretation of the null hypothesis. In a sense, this study concluded that we can only really learn something through rejection. The important conclusion from that analysis was that a result of non rejection of the null hypothesis is a result of little prescriptive consequence. This was illustrated with an example where both the null and the alternative hypotheses were non rejected.

First, we make the argument that a non rejection of the null hypothesis is, by itself, an outcome lacking statistical force. This is because there could be a very wide range of null hypotheses that would also not be rejected by a similar test procedure. In particular, we could “not reject” that the merger causes no harm but also, with the same approach, “not reject” that the merger is harmful.

The conclusion from here is that “not rejecting” is a soft conclusion or even a non-conclusion. A problem then poses itself when a competition authority fails to reject the hypothesis that a given merger is harmful. Is it then appropriate to let the merger go ahead based on that non rejection, which, as we have argued, may have little statistical meaning?

Our second argument is that non rejection of the null may be a very likely outcome in tests where the level of significance is set in a somewhat mechanical way. Researchers often require very low error I probability while allowing a considerably higher error II probability. This implies that, while they are unlikely to reject a null that is true they may be relatively likely to fail to reject a false null.

The additional point we make after that, though, is that tests are constructed in such a way that there is often excessive bias towards outcomes where we fail to reject. When we demand a 95% confidence level, and the variance of our estimates is relatively high, we will “not reject” a very wide range of values for our null hypothesis. But, at the same time, these tests, while with a fixed error I probability at 5%, have very high, often many times higher, error II probability.

We then moved to a discussion of the choice of significance levels and the trade-off between error I and error II probabilities. In practical applications we will find tests where a mechanical choice of 95% confidence levels leads to very low power (i.e. high probability of error II – failing to reject when the null is false).

In conclusion we would consider less biased a methodological approach where a careful consideration of these choices is made. In order to decide to stop a merger, an authority should attempt to reject the hypothesis that the merger has insignificant price impact for consumers. In order to clear a merger, an authority should attempt to reject the hypothesis that the merger will cause a significant price impact. In either case, the choice of the significance level should take into account the resulting probability of failing to reject when a given alternative is true. In each case, some judgement should be attempted as to the relative social cost of the two possible errors in this context: failing to stop an anti-competitive merger and stopping a competitive one.

One possible avenue is to let prior information guide the choice of the null hypothesis and let the relative costs of the two types of errors guide the choice of the level of significance. There are therefore two distinct types of judgement that the researcher must do. The researcher may start out with a very strong prior that the merger will not be harmful. That should be translated into the statement of H_0. But, at the same time, the researcher may believe that it is potentially more costly, from an economic welfare perspective, to let through a harmful merger than to stop a harmless one.

7 References

Bakan, David, (1966) "The test of significance in psychological research," Psychological Bulletin, Vol. 66, pp. 423-437.

Bolles, Robert C., (1962) "The difference between statistical hypotheses and scientific hypotheses," Psychological Reports, Vol. 11, pp. 639-645.

Bunge, Mario (ed.) (1964). The critical approach to science and philosophy: essays in honor of Karl R. Popper, New York: Free Press of Glencoe, Inc.,.

Harlow, L. L., Mulaik, S. A. and Steiger, J. H. (1997). What if there were no significance tests? Lawerance Erlbaum Associtės, Mahwah.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.) (1997). What if there were no significance tests? Mahwah, NJ: Erlbaum.

Levin, J. R. (1993). Statistical significance testing from three perspectives. Journal of Experimental Education, 61, 378-382.

Mahwah, NJ: Erlbaum. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd. ed.). Hillsdale, NJ:Erlbaum. Darlington, R. B. (1990). Regression and linear models. New York: McGraw-Hill.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A Review of null hypothesis significance testing. RESEARCH IN THESCHOOLS, 5(2), 3-14.

Popper, Karl R. (1962), Conjectures and refutations, New York: Basic Books.

Popper, Karl R. (1959), The logic of scientific discovery. New York: Basic Books.

Rozeboom, William W. (1960), "The fallacy of the null-hypothesis significance test," Psychological Bulletin, Vol. 67, pp. 416-428.

Rubinfeld, Daniel L. (1995), "Econometrics in the Courtroom", Columbia Law Review, vol. 85, HeinOnline --- 85 Colum. L. Rev. 1048

Shaver, J. P. (1993) What statistical significance testing is, and what it is not. Journal of Experimental Education, 61, 293-316.

Zumbo, B. D., & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. The Statistician, 47, 385-388.