OVERCHARGE ESTIMATION: MAKING STATISTICAL EVIDENCE MORE MEANINGFUL[1] [2]

Peter Bönisch, Roman Inderst

Abstract

Economic evidence that seeks to evaluate the effect of an infringement is typically presented as a point-estimate of the effect along with its statistical significance. Such evidence alone, however, is often of little value to the judge, in particular if it is impossible for her to infer whether a significant finding is lacking due to poor data or to the actual absence of an effect. To compensate for this limitation, we suggest a simple extension to the *standard procedure* that yields a more nuanced yet still intuitive picture of the underlying data within the current statistical testing framework.

Bullet points (instead of abstract)

Overcharges in cartel cases are typically presented as a point-estimate of the effect along with its statistical significance.

We suggest a simple addition that, compared to statistical significance alone, yields a more nuanced yet still intuitive picture of the underlying data within the current statistical testing framework.

In particular, this may allow a judge to infer whether a significant finding is lacking due to poor data or to the actual absence of an effect.

[1] Our paper closely follows the concept of *severity* introduced by DG Mayo and A Spanos 'Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction' (2006) 57 *British Journal for the Philosophy of Science* 323-357 and recently put into a broader context in DG Mayo *Statistical Inference as Severe Testing. How to Get Beyond the Statistics Wars* (2018). We intend to publish an R package providing functions to produce the statistical output suggested below and some additional explanatory material. Please check the GitHub page (https://github.com/peteboe) for updates.

[2] We thank Calogero Brancatelli for his assistance.

1 Motivation and Focus

Courts frequently have to decide whether an infringement has resulted in potential damages for the plaintiff. Economic expert evidence provided in such cases is typically presented as a so-called *point estimate* of the suspected effect (e.g., the estimation of a price increase of 10 Euro or 12 %), together with an indication of the statistical significance, often expressed by a number of "significance stars". Even the finding of a positive effect of 10 Euro may be presented as statistically insignificant, and then interpreted as supporting the argument that the infringement had no effect. On the other hand, even a small but statistically significant effect may be considered reliable evidence that the infringement had a specific effect.[3] In both cases the realized empirical evidence is evaluated in a dichotomous yes-no framework. In what follows, we refer to the presentation of this type of evidence as the current *standard approach*.

In addition to misinterpretation another drawback of the *standard approach* is that the plausibility, reliability or trustworthiness of the empirical evidence submitted is often unclear. This limits judicial willingness to rely on economic expertise and restrains the overall impact of economic reasoning in an important field of application. More clarity and transparency on the crucial question of whether available economic data provide evidence for an effect of an infringement under study would clearly improve the impact of economic expert evidence. In this article, we propose the presentation of additional information in a way that would enable judges to better contextualize the findings and to support a judgement of whether the presented evidence itself is sufficient to prove an infringement.

Our argument ties into a larger, ongoing debate in the statistical profession, as well as in many applied fields of the social sciences and has recently arrived in the field of economics. This debate more generally centres on issues of interpreting, communicating and presenting empirical evidence and, in particular, on the potential misuse and misinterpretation of standard statistical tests.[4] Unfortunately, the current state of the economic and statistical debate does not provide a coherent and shared "solution" to the issues. Regulators and other legal practitioners, particularly judges, however, have to make decisions within the existing methodological framework. To support the interpretation of empirical economic evidence in these cases, we suggest to present *additional statistical information* in an intuitive manner, supported by graphical illustrations.

Throughout this article, we restrict ourselves to resolving the question of whether an infringement resulted in a significant effect. Moreover, we simplify the discussion by taking the perspective of a court-appointed expert, from whom the court expects evidence that supports its judgement as to whether the infringement had an effect. As all our examples relate to simple comparisons of the mean

[3] Both interpretations are not justified and further discussed below as the *fallacy of non-rejection* and the *fallacy of rejection*, respectively. For the purpose of exposition, we deliberately exaggerate the fallacious interpretation of statistical results as the *standard approach*. However, in practice, even well trained experts are prone to over-interpret the outcomes of statistical tests or to use imprecise language communicating statistical results.

[4] In the course of this debate some researchers even suggest abandoning statistical significance testing (more or less). For a recent critical assessment of the *standard approach* described above see, for instance, RL Wasserstein, AL Schirm and NA Lazar, 'Moving to a World Beyond "p < 0.05"' (2019) 73:sup1 *The American Statistician*, 1-19 or BB McShane, D Gal, A Gelman, C Robert and JL Tackett, 'Abandon Statistical Significance' (2019) 73:sup1 *The American Statistician* 235-245. For a critical view on the former see, for example, DG Mayo 'Don't throw out the error control baby with the bad statistics bathwater: A commentary' (2016) 70:2 *The American Statistician, Online Discussion,*

<https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108?scroll=top&needAccess=true>.

from the infringement period and market with the mean from a suitable comparator period and market, we also steer clear of all issues that relate to the adequate specification of the relevant statistical model.

More concretely, to further refine the focus of this article, we mostly consider the case where the expert's evidence, at least in the *standard approach*, seems unsupportive of an effect, as no statistically significant effect was found.[5] To interpret this alone as evidence that an effect is indeed absent is, however, wrong. Not being able to refute the *hypothesis of no effect* is not equivalent to having found evidence in support of it. To support decision-making in such circumstances we suggest to test whether the observed statistical evidence is more or less conclusive (or probable) if the (hypothesized) actual effect were of a different size, say a price increase of 10 Euro, 15 Euro or 20 Euro, rather than not being positive at all. We present a graphical analysis that makes the answer to this question transparent and useful. As we discuss in this article, this way of presenting evidence should allow even statistically untrained parties, such as courts, to learn about the credibility of empirical evidence and to interpret statistical tests in a more nuanced way than is currently possible within the coarse dichotomous yes-no scheme of the *standard approach*. Our suggestion certainly does not rule out other ways to enhance the statistical evidence that is presented in (follow-on) cases that would enable courts to make appropriate decisions. Instead, this article simply seeks to stimulate the ongoing debate and to contribute to the literature on how to enhance the presentation and interpretation of statistical evidence.

2 Example: Starting from the "Standard Approach"

A Simple Comparison of Means

In this article, we explicitly do not wish to suggest a particular application, such as to some type of *hardcore cartel*. For sake of simplicity, we suppose that the available evidence consists of a series of observed price differences between the affected market and a comparison market. It is helpful to suppose that the expert uses a simple comparison of means (a compactor approach) to provide his statistical evidence. The expert's empirical evidence thus consists solely of the mean of these price differences. Consequently, the key question is whether or not the observed mean difference constitutes sufficient evidence to refute the hypothesis that the infringement did not have a positive price effect.

Representing the *Standard Approach* as Hypothesis Testing

The expert would typically communicate the mean difference together with a measure of its statistical significance. For instance, the mean difference between the prices of the affected market and those of the comparison market could be 12 Euro.[6] While the assumed comparability between the two markets ensures that the mean difference should be close to zero in the case of no effect, in practice

[5] In in this article we focus on the case where the *standard approach* leads to an estimated effect that is not deemed to be statistically significant. There is however also a mirror fallacy, the *fallacy of rejection* (of the hypothesis of "no effect") The availability of ever-larger data sets in certain cases, combined with potentially economically small effects that are, however, leveraged by millions of transactions, will likely increase the relevance of this fallacy in future cases. We intend to address this issue in follow-on work.

[6] While we suppose throughout the analysis that the comparator analysis already takes into account all determinants that would account for systematic differences between the two markets, it is typically not feasible to rule out all confounding factors that affect each individual price realization.

and with only few observations and probably measurement error, the difference can be either positive or negative, depending on the actual price realizations. The concept of statistical significance takes this uncertainty into account. It can be best understood by considering the *standard approach* as the statistical testing of a hypothesis.

The convention is to use the hypothesis (also called the *null hypothesis*) that the effect of the infringement was zero – or, depending on circumstances, zero or negative (e.g., if the infringement could have resulted in efficiencies). The expert would then ask whether the available evidence allows him to refute this hypothesis (of no effect) with sufficient confidence. Such confidence is expressed as the admissible so-called *type I error*, i.e., the probability that the hypothesis of no effect was erroneously refuted. Concretely, in the considered case of the calculation of a mean difference, this translates into the choice of a positive hurdle: From an *ex-ante* perspective, when the observed mean difference lies above this threshold,[7] the expert concludes that by refuting the hypothesis of *no effect* the probability of an error is not higher than a given level. This level is often set at 5 %,[8] which means that, from an *ex ante* perspective, the expert would only consider a mean difference as *statistically significant* when, given this observation, the presumption of an effect would lead to an error in not more than 5 out of 100 cases. Intuitively, the threshold for the mean difference is higher when there are only few observations, so that an observed mean difference, say of 12 Euro, is more likely to be the result of (randomly fluctuating) unspecific circumstances of the individual price realizations, rather than of a systematic difference between the two markets.

Bringing in the *type II error*

When the effect is deemed not to be statistically significant, that is the estimate of the effect is below the predefined level of the admissible *type I error*, from this alone it would be wrong to conclude that this already provides strong evidence to support the hypothesis that there is no effect. In fact, going beyond the logic of hypothesis testing, our proposal aims at providing information that potentially allows the judge to conclude whether the available evidence speaks more or less in favour (of a particular size) of a possible effect.

As a bridge towards the introduction of such additional evidence, we now bring in the so-called *type II error*. Recall first that the *type I error* of 5 % is the (ex-ante) probability of erroneously refuting the null hypothesis of no effect that the expert wants to maximally tolerate. As we explained, a 5 % probability corresponds to a threshold on the observed mean difference, let's say of 15 Euro. That is, the estimated mean difference would have to be at least 15 Euro for the expert to consider it statistically significant, as he would then make a *type I error* of at most 5 % when refuting the hypothesis of no effect. But when the expert estimates in the example a mean difference of 12 Euro, he can clearly *no*t exclude that there was even a sizable positive effect. The ex ante probability with which one would erroneously stick with the null hypothesis of no effect, although a specific effect is

[7] Generally speaking, this approach amounts to defining a so-called test statistic, for which then the derived hurdle is compared with its actual observation. In the chosen example, the test statistic would typically be the so-called *p-value*, which we introduce below. This takes also into account that the expert typically needs to estimate the variance, i.e., the (unexplained) variation in the observations. For ease of exposition, we basically take the variance as given or known.

[8] See, for instance, also European Commission, Practical Guide: Quantifying Harm in Actions for Damages Based on Breaches of Article 101 or 102 of the Treaty of the Functioning of the European Union, § 88. Unfortunately, the practical guide also does not distinguish between what is a convention in economic hypothesis testing and what would be desirable at court.

present, is called the *type II error*. Given some chosen (maximum) *type I error*, in our case of 5 %, and given a specific alternative hypothesis of the true effect, we can make a straightforward calculation of the resulting *type II error*. The lower the expert's tolerance with respect to the *type I error*, the larger will be the *type II error* for a given effect size. Applied to our example, when the expert, for instance, wants to tolerate only a *type I error* of 1 %, rather than of 5 %, he risks more often to not refute the hypothesis of no effect if the actual damage was indeed positive and even substantial.[9]

When an expert thus concludes that he found no statistically significant evidence to reject the absence of an effect at a chosen level for the *type I error*, it would thus clearly be instructive to also report information on the associated *type II error*.[10] After all, when a judge translates the expert's finding into a judgement against an effect, he risks making precisely such a *type II error*. Unfortunately, this information is typically not provided in the *standard approach* as described above. One reason for this is that in order to calculate a *type II error*, one must formulate an alternative hypothesis, e.g., that the true effect was, for example, 10 Euro. Similarly to the calculation of the *type I error* for a specific null hypothesis, a *type II error* is then calculated for a specific alternative hypothesis (or, likewise, for specific stipulated differences from the *null hypothesis* of no effect). An expert may shy away from assuming such possible scenarios. However, as we discuss next, providing the associated statistical information is necessary so that a judge can obtain a more complete picture of the available evidence, which will let him decide whether this evidence speaks in favour of a specific effect or whether the evidence remains inconclusive. Importantly, the judge would then be less likely to commit the error of concluding from a statistically insignificant effect that an effect is absent or very small, because this conclusion is not warranted, unless additional evidence is provided.

3 Judging whether the Expert's Finding of "No Significance" Supports the Conclusion that an Effect is Absent

Recall that in our example a chosen maximum *type I error* of 5 % translates into a certain hurdle, so that the observed mean difference is only deemed to be statistically significant when it surpasses this hurdle. Precisely, we assumed that, given the observed variance in the data, this hurdle was 15 Euro and the actually observed mean difference was only 12 Euro. In other words, in our example the expert did not consider the observed difference of 12 Euro as sufficiently convincing evidence against the absence of an effect and would have treated any observed difference below this 15 Euro threshold as *equally insufficient* evidence against the *null hypothesis*. Instead of treating all possible effect sizes (of 1, 5, 8 or 12 Euros) below the 15 Euro threshold equally as indicating no significant deviation from the *null hypothesis*, the expert should, however, ask what result the test would have produced for a specific alternative hypothesis if the true effect was indeed positive and of a given

[9] The complementary probability of a *type II error* is called the *power of a test* and gives the probability of correctly detecting an effect of a certain size.

[10] This has been proposed in R Inderst, N Frank, K Oldehaver, ´Zur Diskrepanz zwischen gerichtlichen Beweisfragen in Kartellschadenersatzverfahren und den Ergebnissen des ökonomischen „Standardansatzes" bei statistischen Analysen´ (2019) 39 *ZWeR* (in print). This is clearly helpful in particular so as to contrast this with the (maximum) *type I error* that the expert tolerates. The key departure of the following suggestion in this article is that while the *type II error* is calculated from an ex-ante perspective, we take the de facto realized estimate into consideration.

size. This constitutes an important step beyond the *standard approach* since the expert now also scrutinizes the statistical test procedure itself and not just the dichotomous result of the test.[11]

In our case, this boils down to the following simple question: What was the probability of obtaining a larger mean difference than the actually observed value of 12 Euro if the true effect was in fact 5 Euro or even 10 Euro? This question thus directly relates to the *capability of the test at hand to detect a discrepancy* of this size. We can also frame this in terms of the complementary probability, so that then the expert would ask: If the true effect was, say, 5 Euro, what was the probability to observe an effect at least as small as the actually observed mean difference of 12 Euro (and, consequently, as insignificant as this)? We next present a graphical illustration of these questions.

A Graphical Tool

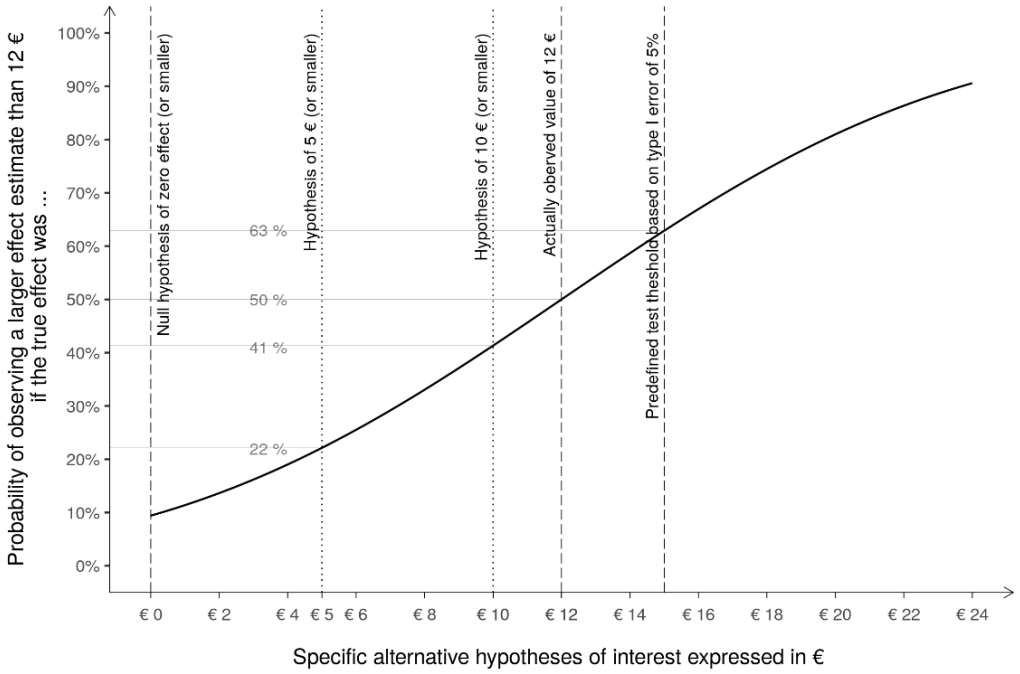
Figure 1 presents the probabilities discussed above on the vertical axis and specific alternative effect sizes to be evaluated on the horizontal axis. To read Figure 1, consider first the value zero at the horizontal axis. In this case, we evaluate exactly the *null hypothesis* of no effect. The black solid line in Figure 1 then depicts the probability with which, under this hypothesis of no effect, an observed mean price difference would be larger than the actually observed value of 12 Euro.[12] Intuitively, the black line must increase, as the higher the hypothesized true effect, the more likely it is that we observe an effect estimate larger than the actual value of 12 Euro.

Along the horizontal axis of Figure 1 we can now evaluate specific effect sizes based on this criterion. If the true effect was, for instance, 5 Euro, we would still only with a probability of about 22 % observe a mean difference larger than 12 Euro. For an effect of 10 Euro this probability increases to 41 %. Obviously, at the actually observed value of 12 Euro it would be equally likely to observe a mean difference below and above this value, so that the black line further increases to the value of 50 %.

[11] This is why the *severity* assessment of a statistical test is described as a *meta-statistical check* (see D Mayo and A Spanos, cf. FN 1).

[12] In our example, this probability is about 10 %. In this special case of exactly evaluating the *null hypothesis* this coincides with the so-called *p-value* expressing the probability of observing a deviation from the *null hypothesis* larger than the observed one if the *null hypothesis* is true.

Figure 1: A Graphical Tool For Interpreting Statistical Test Results



The axis labels in Figure 1 provide an immediate interpretation of the graph in terms of the probability that a "more convincing evidence" than the actually observed mean of 12 Euro is realized if there is a given positive true effect, of, say, 5 Euro or 10 Euro. For example, if the true effect was indeed 5 Euro, we would have observed an estimated difference larger than 12 Euro only with a probability of about 22 % - while we would have the complementary probability of 78 % of seeing a similarly statistically insignificant result. Instead of committing the *fallacy of non-rejection*, we would conclude in the present case that the capability of the test to detect an effect of a size larger than, for instance, 5 Euro was probably too low to rule out its existence.[13]

This line of reasoning not only allows for a more nuanced interpretation of non-rejected hypotheses but also reflects the well-known principle that the weaker the hypothesis we claim, the less likely it is refuted by a statistical test and *vice versa*.[14] The hypothesis of an effect of 5 Euro (or smaller), for instance, would less likely have produced a more contradictory result than the observed 12 Euro than the hypothesis of 10 Euro (or smaller). Hypotheses farther right in Figure 1 therefore survive a stricter statistical test (because the test is more capable to detect a contradicting effect size) and an expert would place more confidence in non-rejecting them. At the same time, however, these hypothesis allow for larger effect sizes to be true.

[13] While in this article we are not interested in an assessment of the size of an effect, note that Figure 1 is also informative in this respect. From the available evidence and Figure 1 we may possibly rule out quite confidently that the effect was 20 Euro or more, at least based on the given test results, since in this case the mean difference would have been larger than the observed value with a probability of almost 80 %.

[14] Generally speaking, the empirical content of a hypothesis reflects the *states of the world* that would falsify it. In the present case this means that the more a hypothesis rules out, the more we learn about the true effect size by rejecting it.

4 Summing Up

We believe that providing a graphical illustration as given in Figure 1 together with the estimated effect would give the judge additional insight. When the expert would now have typically concluded only that the evidence does not allow him to confidently reject the hypothesis of no effect, the judge can put this immediately into a broader context.

For instance, the judge may be more inclined to rule *against* an economically meaningful effect of (larger than) 5 Euro under the given circumstances if the capability of the test to detect contradictory evidence would have been more convincing, for example at 45 %. In the present example, however, we noted that when the actual effect was 5 Euro (or smaller), Figure 1 shows that there was a probability of merely 22 % that contradictory evidence would have been obtained. In the present case, the judge may thus consider the presented evidence as not sufficient to rule against an effect of 5 Euro although the statistical test resulted in an insignificant finding. This means that even when the case is far from clear cut, the information contained in Figure 1 should provide an important background for the communication between the judge and the expert. Most of all, the additional illustration of the test results should reduce the risk of committing the *fallacy of non-rejection*, i.e., that the expert's failure to reject the (null) hypothesis of "no effect" leads to the wrong conclusion that no effect is present (although the test was not capable of detecting any meaningful effect size).

For our example, we can now sum up the available statistical evidence.

First, there is the "standard" evidence of the expert report, i.e., the mean difference between the affected market and the comparator market of 12 Euro, together with the expert's finding that this difference was not statistically significant. The judge should now know, in addition, that the latter finding is equivalent to the testing of the hypothesis of no effect under a given maximum level for the *type I error*, which is the maximum (ex-ante) error probability that the expert will tolerate when he rejects the hypothesis of no effect. In the present case this was 5 %.

Second, the expert may be asked to provide the information contained in Figure 1 nicely summarizing the credibility of non-rejecting different hypotheses given the available statistical test. This overcomes the coarse yes-no paradigm of the *standard approach*. In particular, in our case the judge could read from this that if the true effect was of an economically meaningful size of 10 Euro we would have expected to find a more contradictory effect estimate then 12 Euro with a probability of 41 %. That is, the available statistical test would be capable of detecting such a discrepancy (or a larger one) with about 41 %. With the same line of reasoning the judge my nevertheless confidently reject an effect of 20 Euro since the test would have detected contradictory evidence with about 80 %. The former yes-no scheme of the *standard approach* thereby gets a more nuanced interpretation since the sensitivity of a statistical test differs for different hypothesized effects. The test performance and hence the quality of the empirical evidence at hand is explicitly taken into account. In practice, many debates evolving around allegedly conflicting empirical evidence put forward by the parties involved might be resolved by evaluating the quality of the available statistical evidence.

We admit that there is no simple metric that could be applied to this now more comprehensive picture to reach one of the three possible conclusions: that the evidence was (more) conclusive of an effect, (more) conclusive against an effect or ultimately (still) inconclusive. An expert's finding whether an estimate is by some conventional standard statistically significant can clearly not replace neither an overall assessment of an economic expert nor the overall judgement on the case. However, providing

and interpreting additional statistical evidence such as the capability of the test to detect different effect sizes should provide a stronger basis both for the communication between the expert and the judge and the formation of a final judgement.