ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Bönisch, Peter; Inderst, Roman

Working Paper

Using the Statistical Concept of "Severity" to Assess Seemingly Contradictory Statistical Evidence (with a Particular Application to Damage Estimation)

LawFin Working Paper, No. 3

Provided in Cooperation with:
Center for Advanced Studies on the Foundations of Law and Finance (LawFin), Goethe University

Suggested Citation: Bönisch, Peter; Inderst, Roman (2020) : Using the Statistical Concept of "Severity" to Assess Seemingly Contradictory Statistical Evidence (with a Particular Application to Damage Estimation), LawFin Working Paper, No. 3, Goethe University, Center for Advanced Studies on the Foundations of Law and Finance (LawFin), Frankfurt a. M., https://doi.org/10.2139/ssrn.3702906

This Version is available at:
https://hdl.handle.net/10419/244683

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

FOUNDATIONS OF LAW AND FINANCE

LawFin Working Paper No. 3

Using the Statistical Concept of "Severity" to Assess Seemingly Contradictory Statistical Evidence (With a Particular Application to Damage Estimation)

Peter Bönisch | Roman Inderst

Center for Advanced Studies on the Foundations of Law and Finance | House of Finance | Goethe University
Theodor-W.-Adorno-Platz 3 | 60629 Frankfurt am Main | Germany
www.lawfin.uni-frankfurt.de

Electronic copy available at: https://ssrn.com/abstract=3702906

Using the statistical concept of "severity" to assess seemingly contradictory statistical evidence (with a particular application to damage estimation)

Peter Bönisch, Compass Lexecon*

Roman Inderst, Goethe University Frankfurt**

When parties present divergent econometric evidence, the court may view such evidence as contradictory and thus ignore it completely, without conducting closer analysis. We develop a simple method for distinguishing between actual and merely apparent contradiction based on the statistical concept of the "severity" of the furnished evidence. Again using "severity", we also propose a method for reconciling divergent findings in instances of mere seeming contradiction. Our chosen application is that of damage estimation in follow-on cases.

* pboenisch@compasslexecon.com.

** inderst@finance.uni-frankfurt.de. Inderst thanks the DFG Center for Advanced Studies on the Foundations of Law and Finance for support (project FOR 2774).

Electronic copy available at: https://ssrn.com/abstract=3702906

1. Introduction

In light of the increasing number of antitrust damage proceedings, there is a growing need to legally assess the statistical evidence furnished by third-party experts. This sometimes leads to a paradoxical situation in which the importance of statistical evidence for judicial decision-making may decrease, despite the growing frequency with which such evidence is being furnished to the courts. This is the case as when faced with apparently contradictory statistical findings, courts may be inclined to rather ignore them so that such evidence may tend to simply cancel each other out. This creates a danger of the courts resorting to ad hoc heuristics when rendering a decision, while disregarding an important source of available information. In such a case, there is an inverse relationship between the wealth of statistical information provided to the courts and the extent to which the judgment actually relies on statistical evidence.

In this article we argue that one reason for this development is the prevailing tendency to interpret statistical test results in an excessively mechanistic manner ignoring the design and data limitations of the underlying study. Based on this insight, we suggest an alternative or at least complementary interpretation and, drawing on the statistical concept of the "severity" of the presented empirical evidence, propose a method for considering conflicting evidence in judicial practice. The concept of the severity of statistical evidence was introduced by Deborah Mayo and Aris Spano, notably in their 2006 article.[1] We have already built on this concept in an earlier article, albeit there we only discussed its application as a correction against the misinterpretation of standard statistical tests.[2] In this article, we show how this concept can provide a powerful and intuitive tool, also from the perspective of a statistical layman, to assess seemingly contradictory statistical evidence. The interpretation that we provide in this text may also prove particularly helpful for judges who often request to a probabilistic interpretation of estimated parameters or at least parameter ranges.

In the following, we first present the usual procedure for the econometric estimation of infringement effects in the context of a private action for damages. This application is however chosen only for illustrative purposes as our insights and proposed method are not confined to such cases. Rather they should be applicable whenever potentially conflicting econometric evidence is brought to court, i.e. not even confined to either antitrust cases or damage cases.

In particular, we consider the frequently encountered situation in which supposedly contradictory empirical evidence is introduced by the parties to the case. Through reference to a specific (numerical) example, we discuss how statistical test results are usually interpreted in practice, and how this impacts judicial decision-making. After presenting in detail the potential pitfalls in this regard, the article concludes by proposing an overall assessment method for considering apparently contradictory statistical evidence in judicial decision-making. Specifically, we advocate considering the "severity" of the

[1] 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. See also more recently DG Mayo *Statistical Inference as Severe Testing. How to Get Beyond the Statistics Wars* (2018).

[2] B Bönisch and R Inderst 'Overcharge Estimation: Making Statistical Evidence More Meaningful' (2019) 10.8, *Journal of European Competition Law & Practice* 499-504. A German version, again without an application to the reconciliation of different statistical evidence, can be found in B Bönisch and R Inderst 'Zur Interpretation empirischer Evidenz vor Gericht' (2020) 18, *Zeitschrift für Wettbewerbsrecht* pp. 52–68.

Electronic copy available at: https://ssrn.com/abstract=3702906

evidence presented by both parties in order to distinguish between actual or merely apparent contradiction. For this we first discuss this statistical concept and then apply it to the considered illustrative case.

The exposition of our analysis comes with two important caveats. First, we not delve into the details of damage quantification.[3] That said, our analysis is independent of the method by which a particular estimation of the infringement effect is obtained, as long as it comes with standard statistical information relating to its precision, which we detail below. Second, while we first lay out how typically the results, both from the perspective of the plaintiff and that of the defendants, are presented and how these may be interpreted wrongly, ours is not a general introduction into the respective methods of statistical testing. That said, we strongly believe that the subsequent material can be fully digested even without any statistical background.

2. Constructing an example

The considerations presented in this article are illustrated by drawing on the following fictitious example: In the context of a private suit to recover damages that is filed following a legal judgment of hard-core cartel infringement, both parties to the case submit expert testimony that relies on statistical data. In these fictitious proceedings, we further assume that both experts submit comparative market assessments to quantify the damage resulting from the infringement. However, these assessments arrive at divergent conclusions, to be specific due to their use of different data sets.

In fact, when such data sets are introduced as evidence in legal proceedings, they naturally encompass divergent time ranges and/or customer and products segments. This is to be expected, as companies collect data in different ways, and also maintain a unique set of business relationships. Plaintiffs, for example, can usually draw on information from different cartel members (meaning the data stem from a low number of customers, but several manufacturers). By contrast, defendants typically have access to transaction data from various customers, but only from one manufacturer (i.e., their own company). As a result, alternating perspectives on the same empirical phenomenon – i.e. the antitrust infringement – can arise solely by virtue of the data normally available to the parties. In this way, even if the parties rely on equivalent economic models, though this may rarely be the case in practice, they will not arrive at identical estimates of the incurred damage.

Expert witnesses usually address the uncertainty attendant to their estimation results by offering caveats regarding statistical significance. Typically, experts conduct significance tests and may represent the outcome by means of so-called probability values (or p-values).[4] In the following, we present an example of empirical evidence furnished to a court in order to discuss the prevailing methods for assessing statistical findings.

[3] See, for instance, 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.

[4] For a formal introduction, see for instance DG Spanos: Probability Theory and Statistical Inference: Empirical Modeling with Observational Data (2019), p. 553.

Electronic copy available at: https://ssrn.com/abstract=3702906

2.1. Empirical evidence for infringement-related damages

Let us assume by way of example that the plaintiff has provided expert testimony that demonstrates significant damages resulting from the infringement. In our fictitious case a comparative market approach concludes that damages have been suffered in the amount of €8 per item sold. The first step in assessing this finding is to place it in proper context.

The methods of statistical inference that are commonly used by expert witnesses are ultimately based on the idea of a test. The null hypothesis describes the starting point of the empirical investigation. In the present context, the null hypothesis represents the situation that would have existed in the absence of infringement – or if the infringement, while it was confirmed to exist, did not have an effect. Furthermore, the statistical rejection of this hypothesis is used to infer the existence of an infringement effect. The decision to accept or reject the null hypothesis is always subject to a certain probability of error. The respective uncertainty associated with the estimate is typically depicted as shown in figure 1. The distribution of this curve is a measure of the uncertainty that attends the empirical estimation of the infringement effect. We describe this now in more detail.

The horizontal axis of figure 1 shows potential infringement effects. The empirically estimated infringement effect of €8 is represented by the dashed vertical line at €8. The null hypothesis (of no effect) is also shown with a dotted line (at €0) in figure 1. The area shaded red represents the p-value, which is a measure of the compatibility of the null hypothesis with the estimated infringement effect of €8. This p-value indicates the probability of estimating an infringement effect greater than €8 if the null hypothesis is valid (i.e., if there is indeed no effect). This aspect of the test is crucial. Although the effect has been estimated at €8, given the attendant uncertainties, we must consider the probability with which larger (or smaller) values could have been obtained even when the effect was zero.

In our example, the size of the area shaded red indicates that, given the validity of the null hypothesis ("no infringement effect"), the probability of obtaining a damage estimate larger than €8 is approximately 1%. This is usually considered a sufficiently small probability to reject the null hypothesis and thus (with sufficient probability) to assume an infringement effect. In practice, this is usually expressed in terms of statistical significance. Specifically, this measure indicates the probability of *incorrectly* interpreting the estimated infringement effect of €8 as evidence against the null hypothesis. Accordingly, when experts prepare statistical evidence they usually pick a low significance level; 5% is a common figure.

Electronic copy available at: https://ssrn.com/abstract=3702906

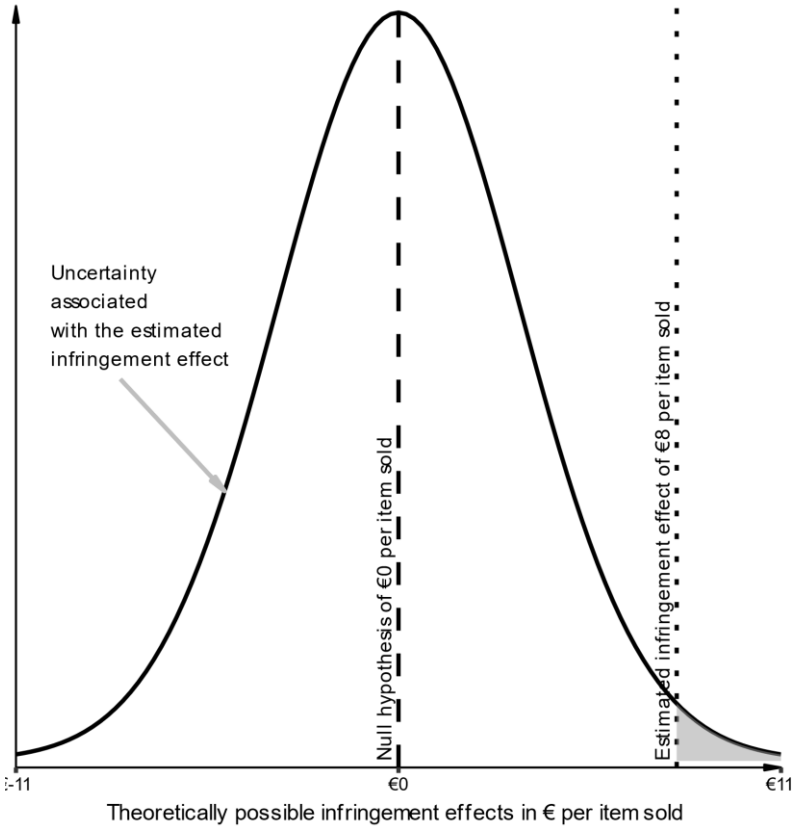

Figure 1: Empirical evidence for an infringement effect (plaintiff's expert)

In our example, the probability of an incorrect conclusion (estimating an effect if the true effect is indeed zero) is approximately 1%, which is well below 5%. The probability of erroneously rejecting the null hypothesis can thus be regarded as sufficiently small. Accordingly, the null hypothesis of "no infringement effect" would be rejected. By extension, the estimated effect is considered to be *significantly different from zero.* Yet what does the rejection of the null hypothesis with an estimated value of €8 mean in concrete terms? Does this mean that the true effect is €8? In answering these questions, there is a risk of drawing incorrect conclusions. We will return to this point later, after having introduced the defendant's evidence.

2.2. Empirical evidence against an infringement effect

Let us assume that the defendant has also submitted empirical evidence, which estimates the damages at €1 per item sold. As discussed, figure 2 shows the uncertainty associated with the estimate of €1. The area shaded grey, or the p-value, expresses a probability of approximately 45%. In other words, the probability of incorrectly interpreting the estimated value as evidence against the null hypothesis of "no infringement effect" is 45%, which is much higher than the commonly used significance level of 5%. Accordingly, the estimation must be considered statistically insignificant or statistically not different from zero.

Electronic copy available at: https://ssrn.com/abstract=3702906

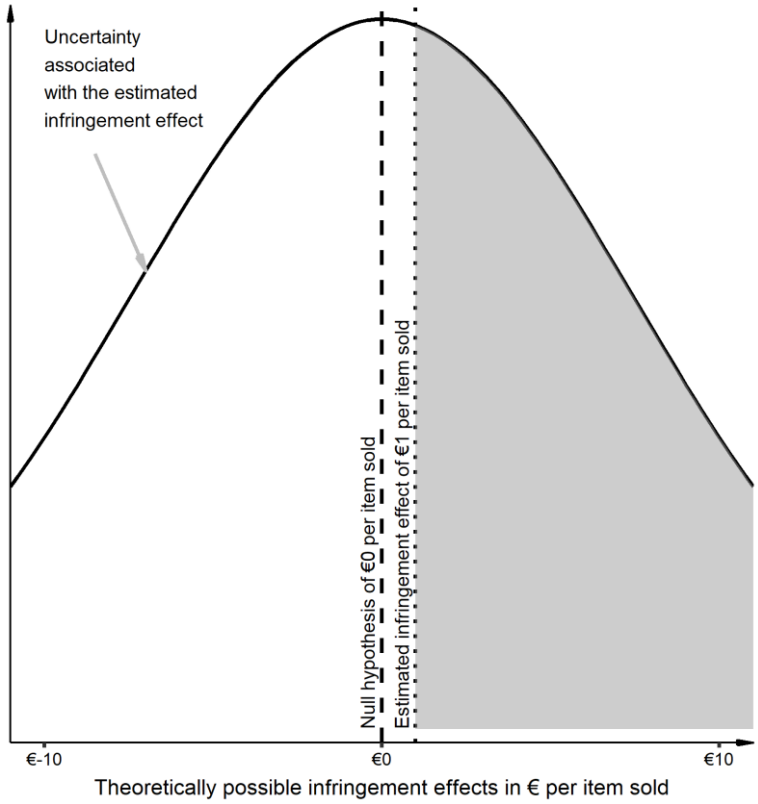

Figure 2: Empirical evidence against an infringement effect (defendant's expert)

2.3. Possible causes of contradictory estimation results in real-world practice

Before comparing the two estimation results, let us briefly discuss possible reasons for these divergent empirical findings.

First of all, it should be recalled that any estimation is subject to a certain degree of uncertainty simply because the underlying data are always incomplete and susceptible to measurement errors. As discussed at the beginning of section 2, conflicting assessments of the matter under litigation are unavoidable in many cases simply because the parties have access to different sets of data. Therefore, depending on the robustness of their estimation models and the size of their samples, their respective conclusions will always differ to a certain extent.

However, there may be other reasons for divergent conclusions. In some cases, the selection of methods or data by one or both parties may also be results-driven, and, in this way, produce sharply divergent conclusions. We discuss this issue separately below.

2.4. Initial conclusions

With a view to the overall assessment of the presented statistical evidence, it is of fundamental concern whether the difference between the estimated damages (in our example, €1 versus €8) differ (a) due to small sample sizes and associated estimation uncertainty or divergent data sets; (b) due to (explicitly or implicitly) contradictory assumptions in the underlying models; or (c) due to the results-driven selection of the estimation results. Based on the methods that are currently used, however, these potential reasons can often not be distinguished. As a result, in judicial practice, apparently or actually contradictory findings may tend to simply cancel each other out.

Electronic copy available at: https://ssrn.com/abstract=3702906

The broader insight offered by this article is that (1) not all cases of seeming contradiction are however irreconcilable, but that (2) certain contradictions point to deeper methodological discrepancies between the analyses in question. We argue that each case of contradiction should be handled differently: In the first case, the presented results should be synthesized in a manner that is viable and expedient. In the second case, a more detailed analysis should be carried out and, if necessary, the underlying problems should be addressed.

In the following, we first present the errors of interpretation that commonly give rise to the unjustified view that statistically significant and insignificant estimation results are incompatible. We then propose a heuristic that enables one to differentiate between situation 1 and 2. Depending on the apparent source of contradiction, different approaches are advised. Finally, we outline a heuristic for conducting an overall assessment of divergent evidence, given a contradiction that is merely apparent.

3. Problems with interpreting empirical evidence in professional practice

Before we assess the empirical evidence presented in our example, let us first address common fallacies that plague the interpretation of statistical estimates.

3.1. Overinterpretating statistically significant results

A common mistake is to overinterpret statistically significant findings. First of all, it should be remembered that statistical analysis is usually carried out in the sense of a test. This involves testing a null hypothesis – in our case, the hypothesis of "no infringement effect." The statistically significant result of €8 thus represents evidence against the null hypothesis of no effect.

However, the rejection of the null hypothesis should not be overinterpreted as evidence for a *specific* alternative hypothesis. In other words, the estimate of €8 provides no indication per se as to the reliability of €8 as an alternative hypothesis.[5] Due to estimation uncertainty, the actual effect could easily be €7, a figure that is highly compatible with the actual estimate of €8. Or it could be €9. Fleshing out this "compatibility" is a key element of the considerations presented in the following.

3.2. Overinterpretating statistically insignificant results

There is a similar potential for error when interpreting the statistically insignificant finding of €1. Here, as well, there is a risk of overinterpretation if one ignores the context of the test. A non-significant finding initially only means that the null hypothesis cannot be rejected at the applied significance level. Indeed, even when an estimate yields a large effect size, statistical insignificance may be attributable to low estimation accuracy (e.g. because of a limited number of data points, or because of a large

[5] This becomes intuitively clear when one considers that an effect of more than €8 was not rejected. Rather, the discrepancy relative to the null hypothesis is merely considered too large.

Electronic copy available at: https://ssrn.com/abstract=3702906

number of other factors that affect prices). In such a case, it is obvious that a statistically insignificant result should not be interpreted per se as evidence for the absence of an infringement effect.[6]

In our example, the estimated result of €1 is also highly compatible with an actual effect of €3. In the following, we discuss this notion of "compatibility" in greater detail.

3.3. Undertaking a critical overall assessment

In light of the estimation uncertainties described above, it cannot be concluded that a statistically significant estimate by one party and a statistically insignificant estimate by another party are mutually exclusive in a categorical sense. In our case, the estimate of €8 does not necessarily prove an infringement effect of exactly €8, and the alternate estimate of €1 does not represent evidence for the general absence of an effect.

It would therefore be wrong to ignore the evidence furnished by both parties on the basis of this discrepancy alone. Yet it would also be misguided to arrive at a compromise by averaging the two values: for example, by averaging €0 (the absence of an infringement effect put forward by one party) and €8 (the estimated value put forward by the other party in our example). This would yield (€8 + €0)/2 = €4. Using this approach, it would not matter whether the non-rejected infringement effect of €1 was €-2 or €3. Regardless of the specific estimate, it would be mistakenly interpreted as "zero" solely on the basis of its statistical insignificance. More rarely, one encounters an approach in which the estimated effects are averaged. In our example, the result would then be (€8 + €1)/2 = €4.50. Such an approach fails to take into account that the results may have divergent levels of uncertainty..

Overall, such averaging also runs the risk, as mentioned in section 2.4, of ignoring potentially deeper methodological problems plaguing the applied models, problems that ultimately substantiate a conclusion of inherent contradiction. Accordingly, we propose performing an overall assessment that considers the "severity" of the furnished evidence. This can help us to determine whether the findings furnished by the two parties are actually compatible or not.

4. Introduction to the concept of "severity"

In this section, we consider each estimate separately while introducing the concept of "severity." While we explain this concept in detail below, we refer also to the references mentioned in the introduction. Subsequently, we take both estimates together, again using the concept of "severity".

4.1. Reinterpreting the evidence for an infringement effect

Let us begin with the statistically significant estimate of €8, which led us to reject the null hypothesis of "no infringement effect." The plaintiff asserts this is evidence for an infringement effect of exactly €8. However, as noted, this is an erroneous conclusion, for only the null hypothesis was tested. No other hypotheses were assessed (e.g. Do the damages amount to exactly €8? Do they amount to at least €8?). Statistically, we have only established that there is a very low probability of no positive

[6] As with all statistical tools, we must consider the extent to which the test and available data actually allow us to determine deviations from the null hypothesis – that is, we must consider the *statistical power* of the test.

Electronic copy available at: https://ssrn.com/abstract=3702906

infringement effect. The evidence thus weighs “severely” against the absence of an infringement effect.

Let us now use the test result to evaluate also different hypotheses: namely, that the infringement effect is “not higher than at most €x,” where x stands for a possible infringement effect of, say, €2 or €6. The key question is: How severely does the evidence presented by the plaintiff weigh against such an alternative null hypothesis, i.e. of an infringement effect of at most €x? We can answer this question again only with a probability of error, which we can estimate – similar to how we asked whether the null hypothesis can be rejected with sufficient certainty.

Specifically, using standard statistical calculations, we can answer the following question for each value of an infringement effect x: If the actual infringement effect is at most x, what is the probability of the observed estimate lying below the actually observed value of €8? We already know the answer when x equals €0: namely, 99%, i.e. the difference between 100% and the previously considered 1% (the probability of error when rejecting the null hypothesis “no infringement effect”). We can now apply this logic to other hypotheses about the magnitude of the damages (i.e. to other values of x). When x = €5, the corresponding probability is, for example, 80%. Therefore, if the actual infringement effect is not higher than €5, then there is an 80% probability of obtaining an estimated value below the observed value of €8. If the value of x is higher – for example, €6 – then the probability would naturally decrease (in this case, to 72%).

While the observed evidence of €8 thus weighs more severely against the hypothesis of an infringement effect of €5 at most, it obviously weighs less severely against the hypothesis of an infringement effect of €6 at most – and the difference can be expressed in terms of the respective probabilities. Figure 3 shows the severity of the evidence against the hypothesis of an infringement effect of €x at most for each estimated value x (each possible infringement effect). Figure 3 also shows a threshold of 80%, which is initially arbitrary and will be discussed below. This can be interpreted as a minimum standard for the required severity of evidence as follows: If an 80% threshold is used as a minimum standard, one would assume that the evidence presented, which led to an estimate of €8, speaks with sufficient severity against an infringement effect of up to €5 – but not against an infringement effect of up to a higher value, say €6.

It should be noted that this statement is not subject to the previously discussed fallacy of overinterpreting the meaning of €8. Indeed, this statement does not contend that the actual value is necessarily equal or close to the estimated value. Rather, we have merely evaluated the reliability of this estimate by considering the probability of observing a different result. This act of interpretation also makes it possible to check the compatibility of the evidence with the estimates produced by the opposing party.

Electronic copy available at: https://ssrn.com/abstract=3702906

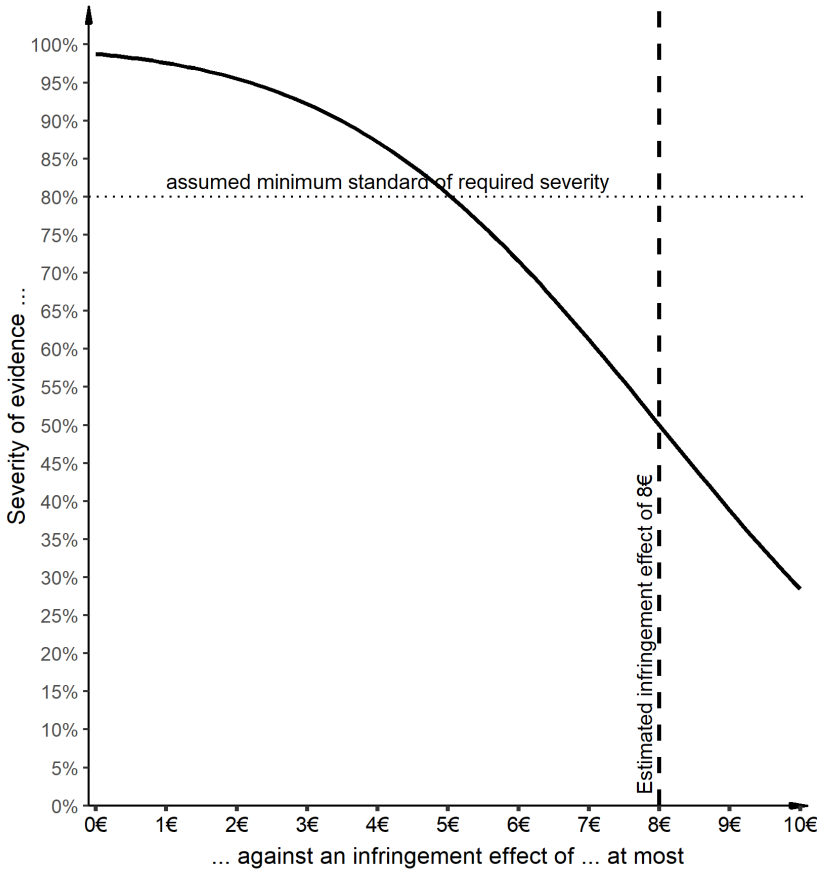

Figure 3: Severity of possible infringement effects (for rejection of null hypothesis) based on plaintiff's evidence

Before applying the concept of severity to the opposing party's estimate, let us briefly consider this concept from a slightly different perspective. Ultimately, the curve in figure 3 can be regarded intuitively as the result of a "question–answer sequence," or, in other words, as a summary of the different assertions that result from different initial hypotheses (different values for x, i.e. potential infringement effects). For example, when x equals €7, the plaintiff is making the following assertion: "The infringement effect is at least €7" or "The evidence speaks against an infringement effect of less than €7." In order to assess this assertion, the judge then asks for an associated probability.[7] Specifically, she examines the plaintiff's assertion by asking the following: Given an infringement effect of up to €7, what is the probability that the estimated infringement effect would have been below the observed effect of €8? If this probability is sufficiently high, then she has reason to consider the plaintiff's claim as sufficiently plausible (or as sufficiently robust statistically). If the judge requires a very high threshold, lower values can be excluded. At a threshold of 90%, for example, even the assertion of an infringement effect of at least €3.50 would be dismissed by the judge as insufficiently robust.

Although a threshold of 80% was used in figure 3, this article does not clarify the level at which the threshold should be set. A generally applicable answer to this question cannot be provided, as it may be reasonable to consider case-specific circumstances, such as the gravity of the infringement or the

[7] In this context, the question of the probability of the infringement effect assuming a certain value would not be meaningful, since *a priori* the probability for each value is zero. The usual test statistics can also not provide information about the probability of the actual value being within a certain interval. This is often a source of misunderstanding or incorrect interpretation. Such statements can be made with Bayesian statistics (which, however, are not usually used – and raise their own problems).

Electronic copy available at: https://ssrn.com/abstract=3702906

persuasiveness of a given theory of harm. Ultimately, the consideration of proximal circumstances should not culminate in the setting of arbitrary thresholds, but rather help to ensure that the development of damage estimates does not become a purely mechanistic exercise, detached from the real world. The citing of concrete grounds for a higher or lower threshold in combination with specific reference to statistical findings, as shown in figure 3, can help to ensure that the overall assessment of the evidence takes place within an objective and rational framework.

4.2. Reinterpreting the evidence against an infringement effect

The concept of severity introduced in the foregoing can now be applied in a similar manner to the estimate furnished by the opposing party. Recall that this party had estimated a (statistically not significant) effect of €1. Thus, the null hypothesis of "no infringement effect" was not rejected. However, as discussed, it would be mistaken to conclude from this estimate that the size of the effect is necessarily zero. Figure 4 illustrates this fact, by applying the concept of severity to different null hypotheses.

Figure 4 provides a graphic depiction of answers to the following question for different values of x: If the actual infringement effect is at least x, what is the probability of the observed estimate being *higher than* the actually observed value of €1? The greater this probability, the more severe the observed evidence weighs against this respective value for x. We already know the answer to x = €0: namely, 45%, as shown in figure 2. Drawing on the p-value, i.e. the probability of error, it was deduced that the null hypothesis of "no infringement effect" cannot be rejected at the chosen significance level. However, it is also evident from figure 2 that the estimate as a whole is subject to considerable inaccuracy, which is why the bell curve is relatively broad (see figure 1). This is now directly reflected in the evaluation of other initial hypotheses, i.e. other possible x values for an infringement effect.

For example, the severity of the evidence (plotted on the vertical axis) only crosses the threshold of 80% at a value of approximately €7.30. Therefore, if the actual infringement effect was at least €7.30, then the probability of estimating a value higher than €1 would be 80%. If this threshold is applied, then only values above €7.30 can be excluded due to estimation uncertainty; the evidence only weighs with sufficient severity against values above €7.30. Figure 4 shows the severity of evidence (vertical axis) for all initial hypotheses (horizontal axis).

Electronic copy available at: https://ssrn.com/abstract=3702906

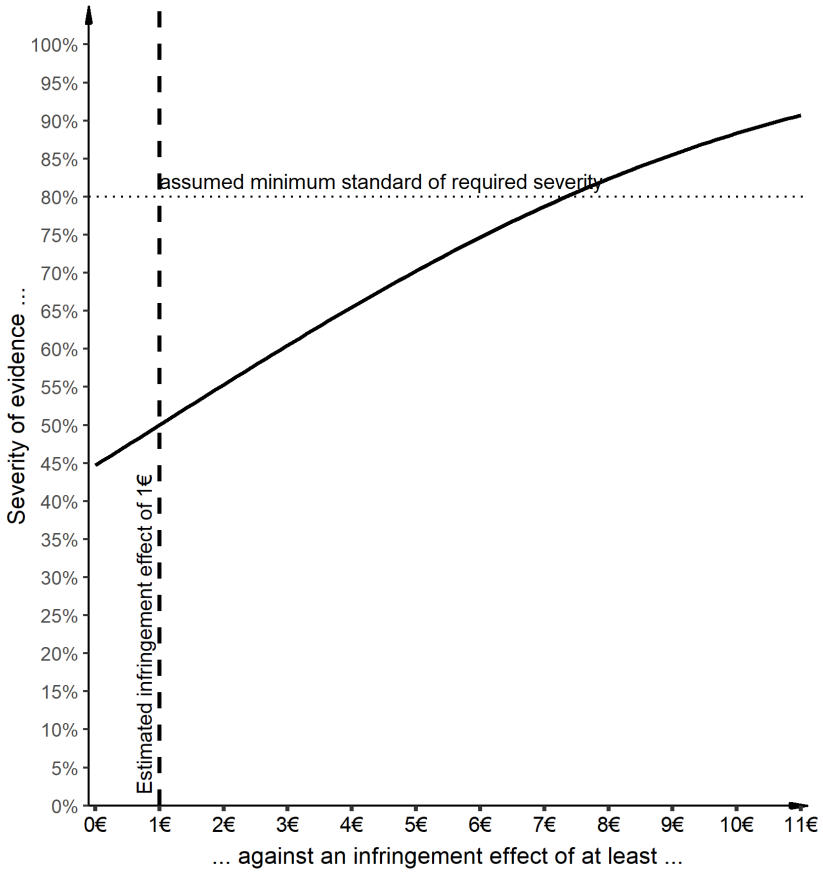

Figure 4: Severity of possible infringement effects (for non-rejection of null hypothesis) based on defendant's evidence

5. A proposal for the overall assessment of seemingly contradictory evidence

Now that we have introduced the concept of severity and have applied it to each of the two estimates, let us now consider the presented seemingly contradictory evidence together. In our example case, this contradiction can only be ostensible, for the rejection of the null hypothesis of "no infringement effect" with an estimated value of €8 is not necessarily evidence for this exact value, and, by the same token, the opposing non-significant estimate of €1 is not necessarily evidence against any effect. Accordingly, as discussed, it would also be a mistake to simply average the two values in an undifferentiated fashion. Instead, the evidence may be synthesized by linking the two severity curves (figures 3 and 4). This allows us to consider the two actual estimates of €1 and €8 while undertaking a differentiated evaluation of divergent potential effect sizes. This is shown in figure 5.

Figure 5 also uses the threshold of 80% as the applied criterion, which we will not elaborate on at present. Using this threshold, we obtain a range of possible infringement effects (the white area between €5 and €7.30) that are sufficiently compatible with both estimates. More precisely, at the severity level considered sufficient, these estimates of infringement effect are rejected neither by the plaintiff's evidence nor by the defendant's. In other words, the figures in this range are neither too low (in relation to the evidence from the statistically significant estimate of €8) nor too high (in relation to the statistically non-significant estimate of €1).

Electronic copy available at: https://ssrn.com/abstract=3702906

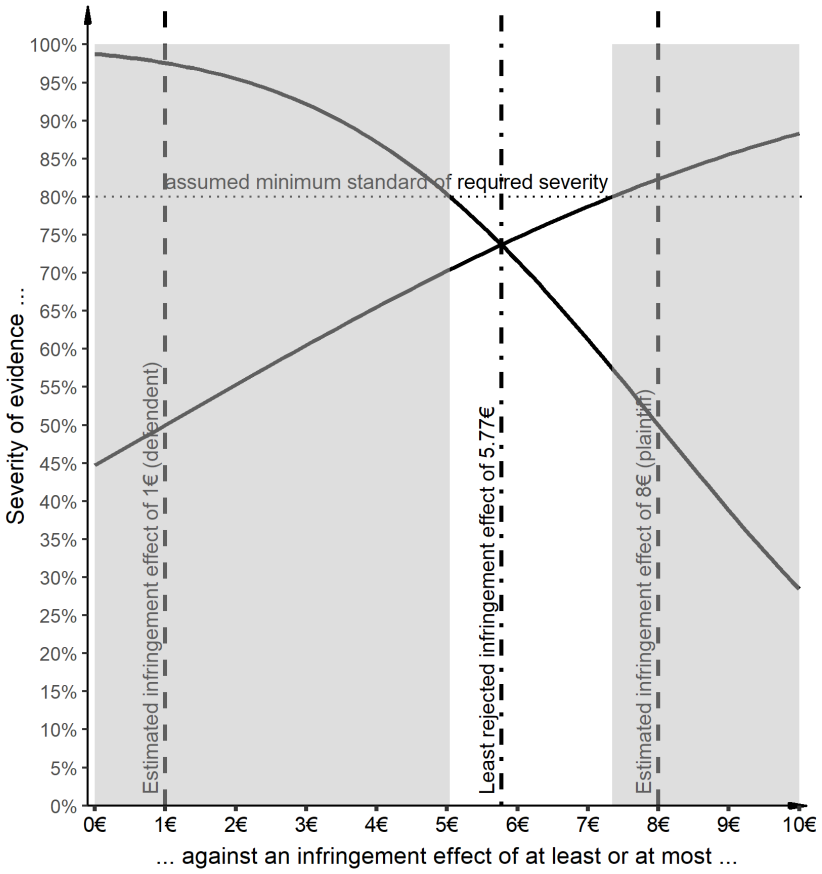

Figure 5: Overall assessment of compatible evidence

If the required severity for rejecting the alternative hypothesis were to be set higher – at, say, 90% – then a wider range of values would be compatible with the two estimates, i.e. there would be both higher and lower possible infringement effects that would not be rejected by any given result with the severity considered necessary. On the other hand, the range narrows if the required severity is reduced. This makes sense intuitively – as does the fact that the range becomes empty when the severity threshold is sufficiently low. This dynamic is not a statistical artifact, but lies in the nature of things. We will demonstrate this briefly with a threshold value of 70%. Given an estimated infringement effect of €8, at a 70% threshold one would be prepared to discard all infringement effects below €7, as the observed estimate would have to be below the observed value of €8 with a probability of 70%. Likewise, the lower threshold of 70% also excludes a wider range of higher infringement effects due to the opposing test result of €1. Finally, in the extreme case of a 50% threshold, each of the two estimates would be taken as sufficient evidence that the actual infringement effect was not below €8 or above €1, which not only leads to an empty interval, but is also not useful, for the required severity would then be equivalent to a coin toss.

The point of intersection of the two curves in figure 5 constitutes the “overall least rejected” infringement effect based on the estimates of €8 and €1. (We will elaborate on this designation later on). Our overall assessment is informed by the logic of falsification, whereby each empirical estimate provides no evidence for a specific value, but only against a specific null hypothesis. Following this logic, the infringement effect at the intersection of the two severity curves is distinguished by the fact that at least one of the two pieces of evidence would weigh more severely *against any other higher or lower*

Electronic copy available at: https://ssrn.com/abstract=3702906

infringement effect. In the present case, as can be seen from the vertical axis at the point of intersection, neither of the two pieces of evidence would weigh more severely than 74% against the choice of €5.77.[8]

This brings us finally to the answers to the questions raised in section 2.4:

1. Is it possible to consolidate the two estimates, which initially appear contradictory; in other words, are the estimates compatible or not (given a required degree of severity)?
2. If they are compatible, at what infringement effect level should this consolidation take place?

The first question is answered by the severity at the point of intersection of the two curves (i.e. 74%, as shown in figure 5). If there is a region of overlap that is below the required severity threshold, then it is possible to undertake a meaningful overall assessment of the opposing estimates. In our case, it can be assumed with a certain degree of probability that the estimates represent different views of the same phenomenon and that they complement each other in a meaningful way. This certainly holds true for the results shown in figure 5, given a defined threshold of 80%. In particular, the intersection of the curves is not rejected by any evidence with a severity higher than 74%. In this way, if a threshold of 80% is selected, then the opposing pieces of evidence are only contradictory in an *ostensible* sense. A contradiction only arises when one interprets the results in an excessively mechanistic way (as discussed in sections 3.1 and 3.2). Indeed, at an 80% severity threshold, there is an interval of possible values that can be considered sufficiently compatible with the evidence furnished by both parties.

Let us now turn to the second question. The intersection of the two severity curves suggests an infringement effect of around €5.77. In this way, our overall assessment does not arrive at a value that is midway between the estimates provided by the two parties. If we had merely averaged the two estimates, then we would have found an infringement effect of €4 (if the non-significant value of €1 was considered zero) or, alternatively, of €4.50 (if the average between €1 and €8 was chosen). The fact that the value at the point of intersection (€5.77) is higher than €4 and €4.50 is due to the fact that the estimate of €1 has a lower accuracy, which, as already shown, is evident from the divergent widths of the bell curves in figures 1 and 2.

The selected point of intersection of the severity curves thus takes into account not only the magnitude of the estimation results, but also their relative accuracy.[9] Since the construction of the severity curves always follows the principle of a hypothesis test, the point of intersection emerges based on the fact

[8] Strictly speaking, the two estimates are interpreted as evidence of an infringement effect of *at least* x on the one hand and *at most* x on the other. In this way, the results do not argue explicitly *for* a certain infringement effect, but rather, argue *the least* against an infringement effect of €5.77.

[9] In principle, there are certainly alternative and possibly even preferable ways to combine two forms of evidence. One method is to aggregate all observations into one data set and perform subsequent re-estimation. This would usually require the involvement of an independent expert, as well as the release of the corresponding data (and additionally presupposes data compatibility). Another option, although not one that is used in professional practice to our knowledge, would be to calculate an average that is weighted according to estimation uncertainty. These alternatives are not discussed in detail in this paper. In any event, even in such cases, the question would have to be answered as to whether an overall assessment of the results is possible at all or whether it would be better to examine the underlying results in more detail (cf. section 6).

Electronic copy available at: https://ssrn.com/abstract=3702906

that neither piece of evidence weighs with disproportionate severity *against* this choice.[10] However, no conclusions can be derived regarding the probability that the actual infringement value lies in a certain range around the point of intersection or around any other point. The point of intersection also does not take into account how inaccurate *both* estimates (jointly) may be. As noted above, the less accurate the two estimates, the larger the white area in figure 5. This is because neither the lower estimate nor the higher estimate weighs with (sufficient) severity against many values between both estimates.

By graphing and comparing the severity curves of damage estimates, we thus obtain supplementary information that can allow us to reconcile seemingly contradictory findings. This is a particularly valuable assessment technique when the robustness of the presented evidence needs to be examined. However, one must be careful to avoid constructing a point of intersection in an excessively mechanistic fashion. This applies in particular when the issue of "whether" an effect has occurred still needs to be clarified, and the court applies an asymmetrical standard, such that a ruling is made in favor of the defendant if neither of the two parties has presented evidence that is considered to be reliable.

6. Closing remarks on incompatible empirical evidence

In our example case, the seemingly inconsistent evidence proved to be compatible when a severity threshold of 80% was applied. Despite the divergence between €1 and €8, it was possible to reconcile these estimates because both figures (especially the defendant's) were subject to sufficient uncertainty. This is immediately apparent in figure 5, as the two severity curves intersect below the applied threshold of 80%.

Now let us change the initial situation as follows. Assume the defendant's estimate is -1 € (instead of €1), while the plaintiff's estimate is more extreme, at €10. The corresponding severity curves are shown in figure 6. Now, the two curves only intersect above the 80% threshold. In concrete terms, this means that the estimates are not compatible at the required severity level. The plaintiff's estimate of €10 weighs with sufficient severity against mid-range values and, in particular, against all values that are sufficiently compatible with the defendant's estimate (and vice versa).

[10] More specifically, there is no evidence to suggest correspondingly higher values, including the point of intersection, or correspondingly lower values, including the point of intersection.

Electronic copy available at: https://ssrn.com/abstract=3702906

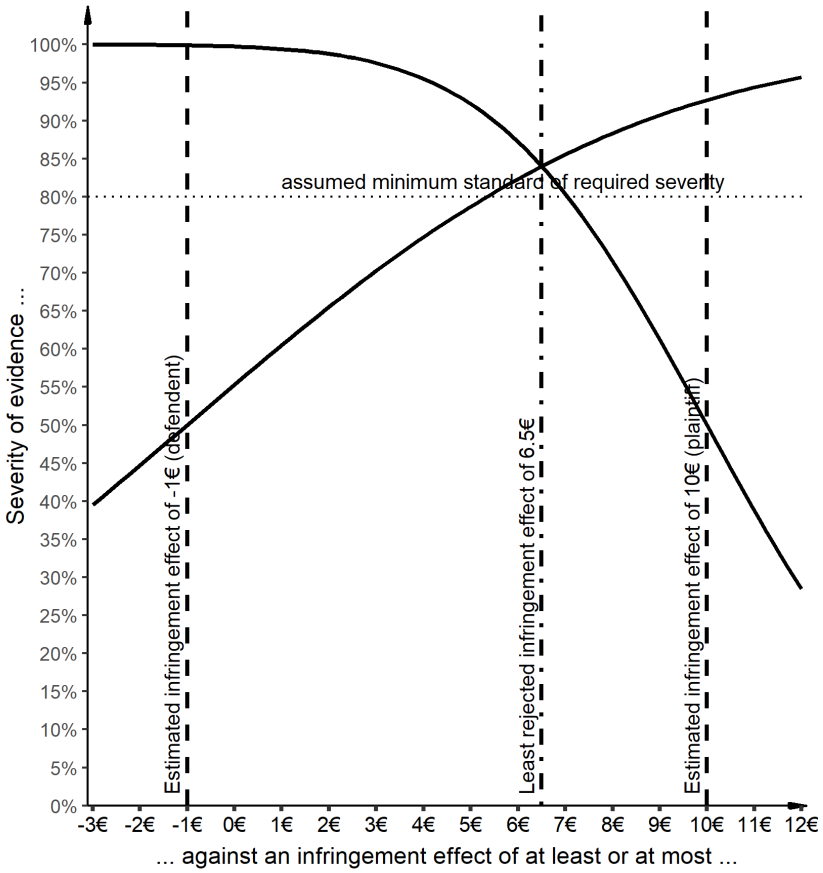

Figure 6: Overall assessment of inconsistent evidence

If no overlapping area exists at a given severity level, then it is necessary to perform a deeper examination of the assumptions underlying the divergent estimations. Specifically, the parties must provide supplementary clarification or an expert must be appointed to investigate this matter for the judge.

Of course, given the uncertainties inherent to estimation, seemingly incompatible discrepancies may arise purely by chance. In other cases, however, the reasons may quickly become obvious upon closer examination. A common cause is the “cherry-picking” of favorable data or estimation techniques. In this connection, biased evidence may not be readily apparent as such if the judge only considers evidence from one party in isolation. Contradictions are more readily apparent when estimates furnished by both parties are compared. Beyond its value as method for reconciling estimations, the concept of severity, as expressed in simple terms in figures 5 and 6, aids in the identification of cases in which evidence from one or both parties is marked by bias and in need of revision. However, other methods are also suitable for screening furnished evidence; we will address this issue in a separate paper.

Electronic copy available at: https://ssrn.com/abstract=3702906