Bönisch, Peter; Inderst, Roman

Working Paper

Ein Vorschlag zur Würdigung vermeintlich widersprüchlicher empirischer Evidenz im kartellrechtlichen Kontext

Ein Vorschlag zur Würdigung vermeintlich widersprüchlicher empirischer Evidenz im kartellrechtlichen Kontext

Peter Bönisch, Compass Lexecon

Roman Inderst, Goethe Universität Frankfurt*

Die Vorlage von in der Regel unterschiedlicher ökonometrischer Evidenz durch die verschiedenen Parteien birgt die Gefahr, dass diese vor Gericht ohne weitere Analyse als widersprüchlich erachtet und damit ggf. sogar gänzlich ignoriert werden. Mittels des Konzepts der „severity" oder „Schwere" einer vorgelegten Evidenz entwickeln wir ein einfaches Vorgehen, durch das zwischen Fällen einer tatsächlichen oder nur vermeintlichen Widersprüchlichkeit unterschieden werden kann. Wir schlagen ebenfalls eine Verfahren vor, wie im Falle einer lediglich vermeintlichen Widersprüchlichkeit eine Gesamtwürdigung erfolgen kann.

* Roman Inderst dankt dem Center for Advanced Studies: Foundations of Law and Finance.

1. Einführung

Nicht zuletzt im Rahmen der steigenden Zahl von Kartellschadensersatzverfahren nimmt die Notwendigkeit der juristischen Würdigung statistisch fundierter Expertengutachten, sei es in Form von Parteigutachten oder in Form von Gutachten gerichtlich bestellter Sachverständiger speziell im kartellrechtlichen Kontext stetig zu. Dabei kommt es manchmal zu der paradoxen Situation, dass mit der zunehmenden Einführung statistisch fundierter Beweismittel in ein Verfahren, die Bedeutung empirischer Evidenz für die richterliche Entscheidungsfindung abnimmt. Vermeintlich widersprüchliche empirische Evidenz tendiert dazu, sich in der Praxis zu neutralisieren. Dadurch entsteht die Gefahr, dass die juristische Gesamtwürdigung in solchen Fällen auf ad hoc Heuristiken ausweicht, welche einen Großteil der vorhandenen Information ignorieren. In solchen Fällen kann dann ein Mehr an statistischer Expertise zu einem Weniger an empirischer Fundierung der resultierenden Entscheidung führen.

Im vorliegenden Beitrag argumentieren wir, dass eine Ursache für diese Entwicklung in der vorherrschenden, teils allzu mechanistischen Interpretation statistischer Testergebnisse zu suchen ist. Wir führen deshalb eine alternative oder zumindest ergänzende Interpretationsweise in die Diskussion ein und schlagen mittels des Konzepts der „severity" bzw. „Schwere" der vorgelegten empirischen Evidenz ein Schema zur Berücksichtigung widerstreitender empirischer Evidenz in der gerichtlichen Praxis vor.

Im Folgenden wird zunächst das übliche Vorgehen bei der ökonometrischen Schätzung von Verstoßeffekten im Rahmen einer privaten Schadensersatzklage im Kontext eines Hardcore-Kartellverstoßes dargestellt. Dabei wird speziell die häufig anzutreffende Situation betrachtet, in der von den prozessführenden Parteien vermeintlich widersprüchliche empirische Evidenz in das Verfahren eingeführt wird. Anhand eines Beispiels werden die in der Praxis vorherrschende Interpretation statistischer Testergebnisse und der daraus üblicherweise folgende Ansatz einer Gesamtwürdigung dargestellt. Nach einer eingehenden Diskussion der dabei möglichen Fehler schließt dieser Beitrag mit einem Vorschlag zur Gesamtwürdigung vermeintlich widersprüchlicher empirischer Evidenz bei der richterlichen Entscheidungsfindung ab. Dabei wird mittels der „Schwere" der vorgelegten Evidenz beider Parteien zwischen tatsächlich oder nur vermeintlich widersprüchlicher Evidenz unterschieden.

2. Konstruktion eines Beispiels

Die folgenden Ausführungen sollen anhand eines Beispiels illustriert werden. Im Rahmen einer privaten Schadensersatzklage im Nachgang zu einem festgestellten Hardcore-Kartell wurden von den prozessführenden Parteien statistisch fundierte Expertengutachten vorgelegt. Ferner sei angenommen, dass beide Gutachten einen Vergleichsmarktansatz verwenden, um auf Basis unterschiedlicher beklagten- bzw. klägerspezifischer Datensätze eine Quantifizierung des durch den Verstoß verursachten Schadens vorzunehmen.

Naturgemäß decken die üblicherweise in ein solches Verfahren eingeführten Datensätze verschiedene Zeitperioden und/oder Kunden- bzw. Produktbereiche ab. Dies liegt einerseits an der unterschiedlichen Organisation der Datenerfassung in verschiedenen Unternehmen,[1] andererseits in der Natur der parteienspezifischen Geschäftsbeziehungen. Während Kläger üblicherweise Informationen verschiedener Kartellanten verwenden können (wenige Kunden, mehrere Hersteller), stehen den Beklagten

[1] Beispielsweise führt die Umstellung von Buchungssystemen zu Brüchen in der Verbuchung von Transaktionsdaten. Gleiches gilt für Änderungen der Rechtsform oder der regulatorischen Rahmenbedingungen.

Transaktionsdaten verschiedener Kunden zur Verfügung (viele Kunden, nur ein Hersteller). Daraus folgt, dass allein die üblicherweise den Parteien zur Verfügung stehenden Daten stets mehr oder weniger verschiedene Perspektiven auf dasselbe empirischen Phänom – hier auf den kartellrechtlichen Verstoß – ermöglichen. Allein dadurch wird es in den seltensten Fällen zu exakt identischen Ergebnissen bei der Schadensquantifizierung kommen, selbst wenn die Parteien ansonsten vergleichbare ökonomische Modelle zugrunde legen würden.

Der daraus resultierenden Unsicherheit einer jeden Schadensquantifizierung wird in der gutachterlichen Praxis üblicherweise durch Aussagen zur statistischen Signifikanz Rechnung getragen. Diese wird im Allgemeinen durch sogenannte Signifikanztests und die dazugehörigen *p-Werte* (engl. Probability Values) ausgedrückt.[2] Im Folgenden stellen wir die in unserem Beispiel von den Parteien in das Verfahren eingeführte empirische Evidenz dar und diskutieren das in der Praxis vorherrschende Vorgehen bei der Gesamtwürdigung derselben.

2.1. Vorgelegte empirische Evidenz für einen Verstoßeffekt

Angenommen in unserem Beispiel habe die klagende Partei im Rahmen eines Expertengutachtens empirische Evidenz für einen substantiellen Schaden vorgelegt. Dieser Schaden betrage im vorliegenden Fall, ermittelt beispielsweise auf Basis eines Vergleichsmarktansatzes, 8 € pro Einheit. Es gilt nun zunächst, dieses Ergebnis richtig einzuordnen.

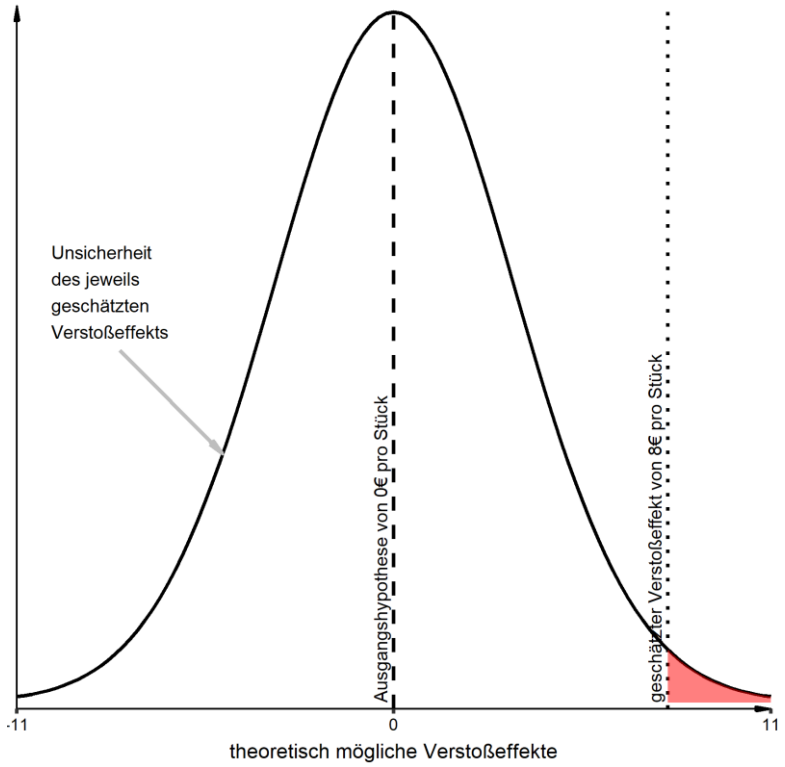
Wie alle in der gutachterlichen Praxis üblicherweise verwendeten Verfahren statistischer Inferenz basiert auch die so vorgelegte Evidenz auf der Idee eines Tests. Dabei beschreibt die sogenannte „Nullhypothese“ den Ausgangspunkt der empirischen Untersuchung, weshalb sie nachfolgend auch als *Ausgangshypothese* bezeichnet werden soll. Im hier betrachteten Kontext stellt sie die Abwesenheit eines Verstoßeffekts dar, sodass aus der statistischen Ablehnung derselben auf das Vorhandensein eines Verstoßes geschlossen wird. Die Entscheidung zwischen Annahme oder Ablehnung der Ausgangshypothese ist dabei stets mit einer gewissen Fehlerwahrscheinlichkeit behaftet. Die mit der vorgelegten Schätzung verbundene Unsicherheit lässt sich anhand der sog. „Glockenkurve“ in Abbildung 1 darstellen. Die Spreizung dieser Glockenkurve ist dabei ein Maß für die mit der empirischen Schätzung des Verstoßeffekts verbundenen Unsicherheit.

Auf der horizontalen Achse von Abbildung 1 werden potentielle Verstoßeffekte dargestellt. Der empirisch geschätzte Verstoßeffekt von 8 € ist durch die gepunktete vertikale Linie bei 8 € dargestellt. Die Ausgangshypothese ist in Abbildung 1 ebenfalls durch eine gestrichelte Linie (an der Null) markiert. Die rot schraffierte Fläche in Abbildung 1 stellt nun den sogenannten *p-Value*, ein Maß für die Vereinbarkeit des geschätzten Verstoßeffekts von 8 € mit der Ausgangshypothese, dar. Diese rot schraffierte Fläche bzw. der p-Value gibt mit den verfügbaren Daten und der daraus resultierenden Schätzunsicherheit die Wahrscheinlichkeit an, dass unter Gültigkeit der Ausgangshypothese ein Verstoßeffekt größer 8 € geschätzt werden würde. Diese Aussage ist wesentlich. Zwar wurde tatsächlich ein Effekt von 8 € geschätzt, allerdings ist zu fragen, mit welcher Wahrscheinlichkeit angesichts der beobachteten Schätzunsicherheit auch größere (oder kleinere) Werte hätten ermittelt werden können.

[2] Für eine ausführlichere Diskussion auch in Abgrenzung zum klassischen t-Test und Konfidenzintervallen siehe Bönisch und Inderst: Zur Interpretation empirischer Evidenz vor Gericht, ZWeR 52 (2020), S. 52-68. Für ein technische Einführung siehe Spanos: Probability Theory and Statistical Inference. Empirical Modeling with Obervational Data (2019), S. 553ff.

Im vorliegenden Fall ist die rot schraffierte Fälle nun gerade so groß, dass es unter Gültigkeit der Ausgangshypothese („kein Verstoßeffekt“) lediglich mit einer Wahrscheinlichkeit von etwa 1 % zu einem geschätzten Verstoßeffekt von größer als 8 € gekommen wäre. Dies wird in der Regel als eine hinreichend kleine Wahrscheinlichkeit erachtet, um die Ausgangshypothese zu verwerfen und damit (mit hinreichender Wahrscheinlichkeit) von einem Verstoßeffekt auszugehen. Dies wird in der Praxis beispielsweise mittels des sog. statistischen Signifikanzniveaus ausgedrückt. Dieses gibt an, mit welcher Wahrscheinlichkeit man den geschätzten Verstoßeffekt von 8 € *fälschlicherweise* als Evidenz gegen die Ausgangshypothese interpretieren würde. Das Signifikanzniveau wird deshalb in der Praxis möglichst klein, in der gutachterlichen Praxis üblicherweise bei 5 % gewählt.

Abbildung 1: Empirische Evidenz für einen Verstoßeffekt



Im vorliegenden Fall liegt die Wahrscheinlichkeit einer Falschinterpretation mit ca. 1 % deutlich unter 5 %, sodass die Wahrscheinlichkeit einer fälschlichen Ablehnung der Ausgangshypothese als ausreichend klein erachtet werden kann. Die Ausgangshypothese „kein Verstoßeffekt“ wäre somit abzulehnen. Das Schätzergebnis wird dann als *statistisch signifikant von Null verschieden* bezeichnet. Zur Frage, was die Ablehnung mit einem geschätzten Wert von 8 € konkret bedeuten soll, kommen wir erst im Anschluss, da es hier zu Fehlinterpretationen kommen kann.

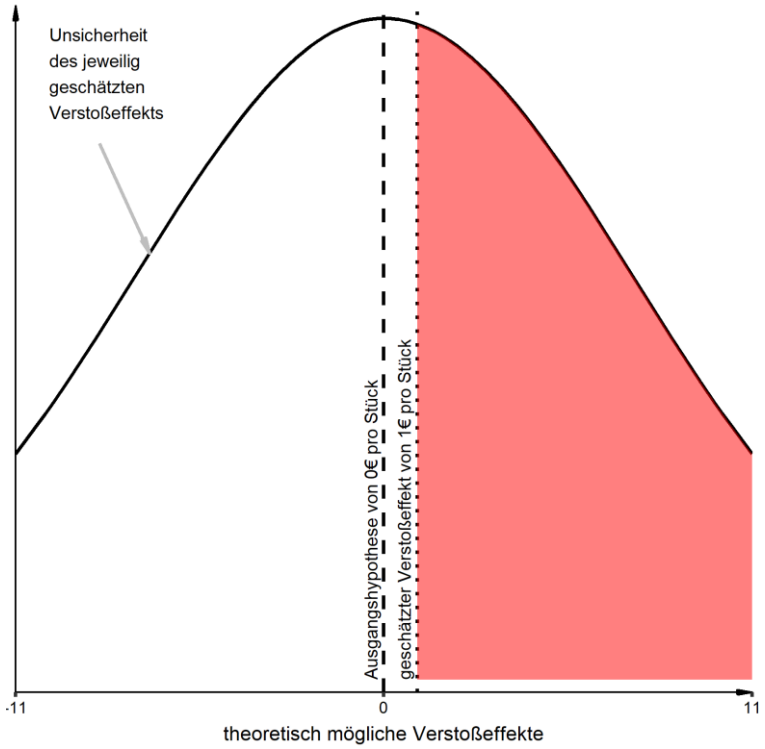
2.2. Vorgelegte empirische Evidenz gegen einen Verstoßeffekt

Nun sei weiterhin angenommen, dass die beklagte Partei ebenfalls empirische Evidenz vorlegt habe. Die Gutachter sollen dabei einen Effekt von 1 € geschätzt haben. Dieses Ergebnis ist nun in Abbildung 2 dargestellt.

Wie bereits diskutiert, stellt auch hier die dargestellte Glockenkurve die mit dem geschätzten Wert von 1 € einhergehende Unsicherheit dar. Die rot schraffierte Fläche (der *p-Value*) entspricht nun einer

Wahrscheinlichkeit von ca. 45 %. Die Wahrscheinlichkeit einer fälschlichen Interpretation dieses Schätzwertes als Evidenz gegen die Ausgangshypothese „kein Verstoßeffekt“ liegt damit deutlich über jedem gängigen Signifikanzniveau (üblicherweise 5 %). Das Schätzergebnis würde als statistisch insignifikant oder als statistisch nicht von Null verschiedenen erachtet.

Abbildung 2: Empirische Evidenz gegen einen Verstoßeffekt



2.3. Mögliche Ursachen widersprüchlicher Schätzergebnisse in der Praxis

Bevor die beiden Schätzergebnisse zusammengeführt werden, soll kurz auf mögliche Gründe für potentiell unterschiedliche empirische Ergebnisse für denselben rechtlichen Sachverhalt eingegangen werden.

Zunächst sei daran erinnert, dass jede Schätzung allein aufgrund der stets unvollständigen und mit Messfehlern behafteten Datenbasis mit einer gewissen Unsicherheit behaftet ist. Wie bereits zu Beginn von Abschnitt 2 diskutiert, ist eine unterschiedliche Perspektive auf das verhandelte Phänomen in den meisten Fällen allein deshalb unvermeidlich, weil die Prozessbeteiligten nur auf verschiedene Datenbasen zurückgreifen können. Deshalb werden die vorgelegten Ergebnisse immer in einem gewissen Umfang auseinanderliegen, abhängig von der Robustheit der vorgelegten Modelle und der Größe der verfügbaren Stichproben.

Möglicherweise haben die Unterschiede jedoch auch andere Gründe, so etwa die selektive Wahl eines spezifischen Regressionsmodells bzw. der darin berücksichtigten Variablen. In manchen Fällen mag auch eine ergebnisgetriebene Auswahl der verwendeten Methoden oder Daten (einer oder) beider Parteien zu eklatanten Diskrepanzen in den ermittelten Ergebnissen führen. Darauf wird nachfolgend noch gesondert eingegangen.

2.4. Zwischenfazit

Ob Unterschiede zwischen den in einem Verfahren eingebrachten Verstoßeffektsschätzungen (im vorliegenden Fall 1€ versus 8€) (a) aufgrund geringer Stichprobengröße oder verschiedener Datenbasis und der damit unvermeidlich verbundenen Schätzunsicherheit, (b) aufgrund (explizit oder implizit) widersprüchlicher Annahmen der zugrundliegenden Modelle oder gar (c) aufgrund ergebnisgetriebener Auswahl der Schätzergebnisse auseinanderliegen, ist eine fundamentale Frage für die Gesamtwürdigung der vorgelegten Evidenz. Mit den derzeit verwendeten Methoden jedoch, kann diese kaum beantwortet werden, was oft zu einer undifferenzierten Neutralisierung (vermeintlich oder tatsächlich) widersprüchlicher Evidenz in der Praxis führt.

Die generelle Einsicht dieses Aufsatzes ist, dass (1) nicht jede vermeintlich widersprüchliche Evidenz tatsächlich einen unvereinbaren Widerspruch konstituiert, allerdings (2) bestimmte Widersprüche auf tieferliegende methodologische Diskrepanzen zwischen den vorgelegten Analysen hindeuten. Wir argumentieren, dass beide Fälle unterschiedlich gehandhabt werden sollten: Im ersten Fall ist eine Konsolidierung der vorgelegten Ergebnisse sinnvoll und möglich. Im zweiten Fall sollte eine weitergehende Analyse erfolgen und gegebenenfalls das tieferliegende Problem adressiert werden.

Im Folgenden stellen wir zunächst die Fehler bei der Interpretation empirischer Evidenz in der gutachterlichen Praxis dar, welche die nicht gerechtfertigte Ansicht begründet, dass signifikante und insignifikante Schätzergebnisse unvereinbar seien. Anschließend schlagen wir eine Heuristik vor, welche die Unterscheidung zwischen den Fällen (1) und (2) erleichtert und ein differenziertes Vorgehen in Abhängigkeit von den vermutlichen Gründen für eine Abweichung ermöglicht. Schließlich skizzieren wir für den Fall (1) eine Heuristik zur Gesamtwürdigung der vorgelegten Ergebnisse.

3. Probleme bei der Interpretation empirischer Evidenz in der Praxis

Bevor wir zu einer Gesamtwürdigung der in einem Verfahren vorgelegten empirischen Evidenz kommen, ist es hilfreich, zunächst mögliche Fehler bei der Interpretation der beiden Schätzungen (sog. „fallacies“) aufzuzeigen.

3.1. Fehler bei der (Über-)Interpretation statistisch signifikanter Ergebnisse

Ein häufig gemachter Fehler liegt in der folgenden (Über-)Interpretation eines statistisch signifikanten Schätzergebnisses. Zunächst sei daran erinnert, dass die statistische Bewertung in der Regel im Sinne eines Testverfahrens durchgeführt wird. Dabei wird eine Ausgangshypothese geprüft – im vorliegenden Fall die Hypothese „kein Verstoßeffekt“. Das statistisch signifikante Ergebnis von 8 € stellt damit Evidenz gegen diese Ausgangshypothese dar.

Die dann erfolgte Ablehnung der Ausgangshypothese kann allerdings nicht (über-)interpretiert werden als Evidenz für eine *konkrete* Alternativhypothese. Dies bedeutet, dass die Schätzung von 8 € an sich nichts aussagt dazu, wie belastbar insbesondere die Alternativhypothese ist, dass der tatsächliche Effekt *genau* 8 € betrage.[3] Bereits intuitiv ist aufgrund der Schätzunsicherheit beispielsweise ein tatsäch-

[3] Dies wird intuitiv deutlich, wenn man bedenkt, dass ja die kein Effekt von über 8 € abgelehnt wurde, sondern lediglich die Diskrepanz zur Ausgangshypothese als zu groß erachtet wird.

licher Effekt von 7 € in hohem Maße vereinbar mit der tatsächlichen Schätzung von 8 €. Die Konkretisierung dieser „Vereinbarkeit“ wird nachfolgend ein zentrales Element der erweiterten Betrachtung sein.

3.2. Fehler bei der (Über-)Interpretation statistisch insignifikanter Ergebnisse

Der mögliche Fehler bei der Interpretation der insignifikanten Schätzung von 1 € ist analog. Auch hier besteht die Gefahr, dass dies deshalb überinterpretiert wird, weil der konkrete Rahmen des Tests der Ausgangshypothese vergessen wird. Das nicht signifikante Ergebnis bedeutet zunächst lediglich, dass die Ausgangshypothese zum angelegten Signifikanzniveau nicht verworfen werden kann. Im Einzelfall kann dies selbst bei einem sehr hoch geschätzten Effekt daran liegen, dass die Schätzungenauigkeit ebenfalls sehr hoch ist, beispielsweise aufgrund nur weniger Datenpunkte oder aber aufgrund einer Vielzahl von anderen Faktoren, die auf die Preise wirken. In einem solchen Fall ist es offensichtlich, dass das insignifikante Schätzergebnis nicht per se als Evidenz für die Abwesenheit jeglichen Verstoßeffekts interpretiert werden darf.[4]

Für unser Beispiel gilt ebenso, dass das Schätzergebnis von 1 € sehr wohl auch in hohem Maße mit beispielsweise einem tatsächlichen Effekt von 3 € vereinbar ist. Wiederum steht die Konkretisierung dieser „Vereinbarkeit“ nachfolgend im Vordergrund.

3.3. Kritische Gesamtwürdigung

Angesichts der dargestellten Schätzunsicherheit kann damit auch nicht gefolgert werden, dass sich eine statistisch signifikante Schätzung einer Partei und die statistisch insignifikante Schätzung der Gegenpartei, die diese als Evidenz gegen einen Verstoßeffekt interpretiert, kategorisch ausschließen. Weder belegt im Beispielsfall die eine Schätzung zwingend einen Verstoßeffekt von genau 8 € noch die andere Evidenz die generelle Abwesenheit jeglichen Effekts.

Es wäre damit fehlerhaft, allein aufgrund dieser Diskrepanz die Evidenz der beiden Parteien zu ignorieren. Und genauso fehlerhaft wäre es, einen naiven Kompromiss etwa durch eine Mittelung der beiden Werte herzuführen:Beispielsweise könnte in diesem Fall als Gesamtwürdigung eine Mittelung aus „Null“ (der Abwesenheit eines Verstoßeffekts) und 8 € (dem geschätzten Wert) vorgeschlagen werden. Damit ergäbe sich (8 € + 0 €)/2 = 4 €. Bei diesem Vorgehen würde es keine Rolle spielen, ob der nicht abgelehnte Verstoßeffekt 1 € wie im Beispiel oder aber -2 € oder 3 € wäre. Er würde in jedem Fall allein aufgrund der statistischen Insignifikanz fälschlicherweise als zwingend Null interpretiert. Seltener begegnet man einen Ansatz, bei dem die geschätzten Effekte gemittelt werden. Das Ergebnis wäre in unserem Beispiel dann (8 € + 1 €)/2 = 4,5 €. Hier fände die ggf. aufgrund unterschiedlicher Datenbasis ebenfalls unterschiedliche Belastbarkeit der verschiedenen Ergebnisse keine Berücksichtigung.

Insgesamt besteht bei solchen Mittelungen auch die Gefahr, dass, wie in Abschnitt 2.4 bereits angesprochen, potentiell tieferliegende methodologische Probleme der vorgelegten Modelle ignoriert werden, die letztlich sogar eine inhärente Widersprüchlichkeit der Evidenz nahelegen. Das nachfolgend eingeführte Konzept der „Severity“ (bzw. „Schwere“) der Evidenz erlaubt dagegen eine begründete

[4] Wie bei jedem Instrumentarium ist auch hier die Frage zu evaluieren, welche Abweichungen von der Ausgangshypothese der verwendete Test mit den vorliegenden Daten überhaupt ernsthaft detektieren kann. Das entsprechende Konzept in der Statistik ist das Konzept der *Power*.

Gesamtwürdigung – und damit auch die Beantwortung der Frage, in wieweit die Ergebnisse der beiden Parteien tatsächlich miteinander vereinbar ist oder nicht.

4. Einführung in das Konzept der „Severity“ („Schwere“)

Wiederum wird zunächst jede der beiden Schätzungen für sich betrachtet, nun unter Heranziehen des Konzepts der „Severity“ („Schwere“ der entsprechenden Evidenz), welches dabei eingeführt wird.[5]

4.1. (Re-)Interpretation der Evidenz für einen Verstoßeffekt

Wir beginnen mit der statistisch signifikanten Schätzung von 8 €, die zur Ablehnung der Ausgangshypothese „kein Verstoßeffekt“ geführt hat und die die Klägerin als Evidenz für einen Verstoßeffekt von genau 8 € interpretieren will. Wie wir allerdings bereits dargestellt haben, ist Letzteres ein Trugschluß: Denn getestet wurde damit lediglich die Ausgangshypothese, nicht aber irgendeine andere Hypothese und insbesondere nicht die Hypothese, dass der Verstoßeffekt genau 8 € oder aber beispielsweise mindestens 8 € betragen soll. Aus statistischer Sicht ist bislang lediglich festgestellt, dass mit einer als hinreichend gering angesehenen Irrtumswahrscheinlichkeit davon ausgegangen wird, dass kein positiver Verstoßeffekt vorliegt. Die Evidenz wiegt damit entsprechend „schwer“ gegen die Abwesenheit eines Verstoßeffekts.

Nun soll das Testergebnis auch dafür benutzt werden, andere Ausgangshypothese zu evaluieren, in folgender Weise: Statt sich lediglich auf den Wert von Null, daher auf die Ausgangshypothese „kein Verstoßeffekt“ zu fokussieren, soll nun die verallgemeinerte Ausgangshypothese betrachtet werden, dass der Verstoßeffekt „*nicht höher als x €*“ sei, wobei für x für eine mögliche Verstoßeffektshöhe steht und beispielsweise 2 € oder 6 € eingesetzt wird. Gefragt wird wiederum, wie schwer die von der Klägerin vorgelegte Evidenz gegen eine solche alternative Ausgangshypothese, d.h. eines Verstoßeffekts unter x €, wiegt. Dies soll durch die Irrtumswahrscheinlichkeit operationalisiert werden – analog zur Frage, ob die Ausgangshypothese mit hinreichender Sicherheit verworfen werden kann.

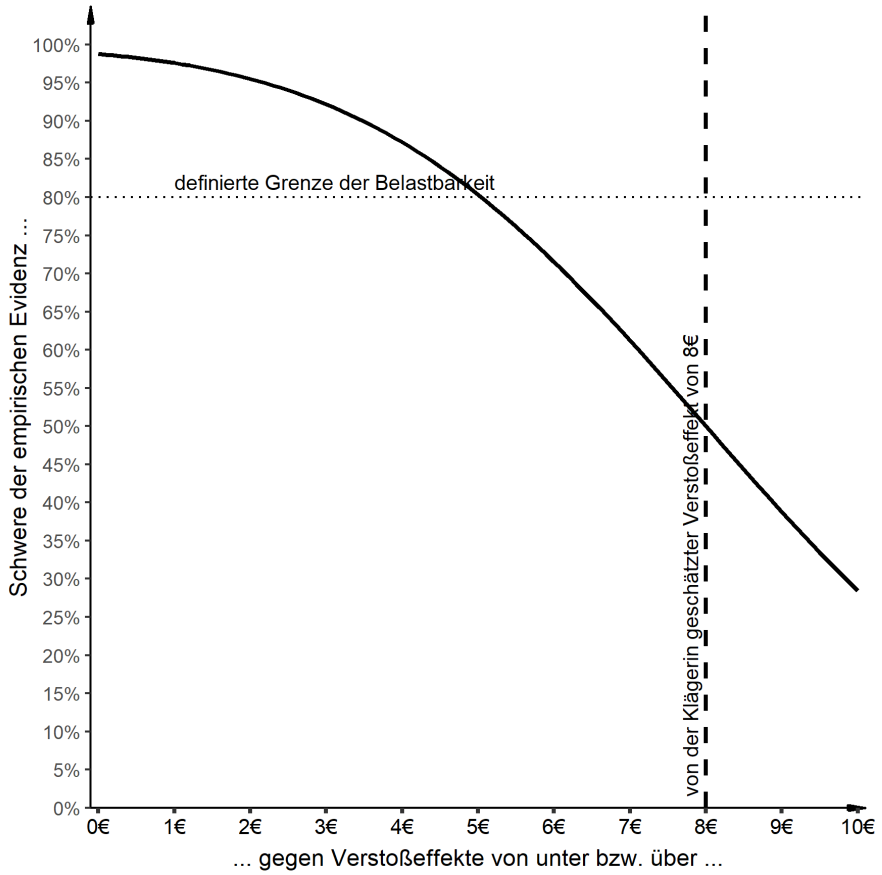
Konkret kann mittels der üblicherweise ermittelten statistischen Werte für jeden Wert x die folgende Frage beantwortet werden: Falls der tatsächliche Verstoßeffekt nicht höher als x wäre, mit welcher Wahrscheinlichkeit hätte dann der beobachtete Schätzwert *unter* dem tatsächlich beobachteten Wert von 8 € liegen müssen? Für x gleich 0 € kennen wir bereits die Antwort: 99 %, also die Differenz aus 100 % und den zuvor betrachteten 1 % (der Irrtumswahrscheinlichkeit bei Ablehnung der Ausgangshypothese „kein Verstoßeffekt“). Wendet man diese Logik nun auch auf andere Hypothesen über den möglichen Verstoßeffekt an (d.h. auf andere Werte x an), so ist beispielsweise die entsprechende Wahrscheinlichkeit für den Wert ca. 5 € gleich 80 %. Daher: Falls der tatsächliche Verstoßeffekt nicht höher als 5 € wäre, so würde man mit einer Wahrscheinlichkeit von 80 % einen niedrigeren Schätzwert als den beobachteten Wert von 8 € ermittelt haben. Bei einem höheren Wert von x – beispielweise 6 € – würde die Wahrscheinlichkeit natürlich sinken, so beispielsweise auf 72 % im Falle von x gleich 6 €.

[5] Siehe dazu Mayo und Spanos: Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction, in: Brit. J. Phil. Sci. 57 (2006), S. 232-357 bzw. Bönisch und Inderst: Zur Interpretation empirischer Evidenz vor Gericht, ZWeR 52 (2020), S. 52-68.

Während damit die Evidenz von 8 € schwerer gegen einen Verstoßeffekt von höchstens 5 € wiegt, so wiegt sie offensichtlich weniger schwer gegen die Hypothese eines Verstoßeffekt von höchstens 6 €. Wie schwer die Evidenz mehr oder weniger gegen die Hypothese von Verstoßeffekten von maximal einem bestimmten Wert von x € wiegt, dies lässt sich für jeden Wert x (jeden möglichen Verstoßeffekt) aus der „severity"-Kurve in Abbildung 3 ablesen. In Abbildung 3 wurde zudem eine Schwelle von 80 % eingezeichnet, die zunächst noch arbiträr ist und erst nachfolgend noch diskutiert wird. Diese kann wie folgt als ein Mindeststandard für die erforderliche Schwere der Evidenz interpretiert werden: Legt man die Schwelle von 80 % für einen solchen Mindeststandard an, so würde man davon ausgehen, dass die vorgelegte Evidenz, die zu einer Schätzung von 8 € geführt hat, mit hinreichender Schwere gegen Verstoßeffekte von unter 5 € spricht.

Hierbei ist zu bemerken, dass diese Aussage gerade nicht dem zuvor dargestellten Fehler einer Überinterpretation des geschätzten Wertes von 8 € unterliegt. Es wird gerade *nicht* festgestellt, dass der tatsächliche Wert zwingend gleich dem geschätzten Wert sein soll oder zwingend in einer bestimmten Umgebung dessen liegen muss. Vielmehr wird neben dem Schätzergebnis selbst auch die Verlässlichkeit desselben im Sinne der Wahrscheinlichkeit evaluiert, gegebenenfalls auch ein anderes Ergebnis beobachtet zu haben. Diese zusätzlich geleistete Interpretation der statistischen Evidenz erlaubt es nachfolgend auch, die Vereinbarkeit mit der Schätzung der Gegenseite zu prüfen.

Abbildung 3: Schwere der Ablehnung verschiedener möglicher Verstoßeffekte bei Ablehnung



Bevor das Konzept der Schwere der Evidenz auch auf die Schätzung der Gegenseite angewendet wird, soll dies zur zusätzlichen Illustration noch einmal alternativ dargestellt werden. Letztlich kann die Kurve in Abbildung 3 intuitiv als Ergebnis einer „Frage-Antwort-Sequenz" bzw. einer Würdigung unterschied-

licher Behauptungen betrachtet werden, die sich aus den unterschiedlichen Ausgangshypothesen (unterschiedliche Werte für x, also Verstoßeffekte) ergeben. So stellt die Klägerin etwa für x gleich 7 € die folgende Behauptung auf: „Der Verstoßeffekt ist mindestens 7 €“ bzw. „Die Evidenz spricht gegen einen Verstoßeffekt, der kleiner als 7 € sein soll“. Um diese Behauptung zu würdigen, fragt der Richter nach Wahrscheinlichkeiten.[6] Konkret prüft er die Behauptung mit der Frage, mit welcher Wahrscheinlichkeit bei einem Verstoßeffekt von maximal 7 € der beobachteten Verstoßeffekt in der Tat niedriger ausgefallen wäre als der beobachtete Effekt von 8 €. Ist diese Wahrscheinlichkeit hinreichend groß, so mag er die Behauptung der Klägerin für hinreichend plausibel (bzw. für hinreichend statistisch belastbar) halten. Erfordert der Richter eine sehr hohe Schwelle, so können weniger Werte ausgeschlossen werden. Bei einer Schwelle von 90 % würde er beispielsweise selbst die Behauptung der Klägerin, der Verstoßeffekt läge angesichts der Evidenz bei mindestens 3,50 €, nicht für hinreichend belastbar halten.

Dieser Beitrag klärt nicht die Frage nach der Höhe der anzuwendenden Schwelle, die in Abbildung 3 auf 80 % gesetzt wurde. Hierfür kann es auch keine eindeutige Antwort geben. So können in diese Schwelle weitere kontextuelle Umstände, so die Schwere des Verstoßes oder die Überzeugungskraft der vorgelegten Schadenstheorie, mit einfließen. Dies darf natürlich nicht in Beliebigkeit ausarten, aber es stellt sicher, dass die Schätzung nicht eine reine statistische Übung, losgelöst vom konkreten Sachverhalt, wird. Durch die konkrete Begründung einer höheren oder niedrigeren Schwelle und durch die dann erfolgende Verknüpfung mit der statistischen Evidenz, wie in Abbildung 3 demonstriert, erfolgt die Würdigung der insgesamt vorgelegten Evidenz immer noch in einem nachvollziehbaren, objektivierten Raster. Dies betrifft auch die Gesamtwürdigung unterschiedlicher Evidenz, wie nachfolgend dargestellt wird.

4.2. (Re-)Interpretation der Evidenz gegen einen Verstoßeffekt

Das damit eingeführte Konzept der Schwere der Evidenz kann nun völlig analog auf die Schätzung der Gegenseite angewendet werden. Zur Erinnerung: Diese Partei hatte einen (statistisch nicht signifikanten) Effekt von 1 € geschätzt. Damit wurde die Ausgangshypothese „kein Verstoßeffekt“ nicht abgelehnt. Wie dargestellt wurde, wäre allerdings die Schlussfolgerung, dass daraus zwingend ein Effekt von Null abzuleiten wäre, fehlerhaft. Gerade dies zeigt auch die Anwendung des Konzepts der Schwere der Evidenz auf unterschiedliche Ausgangshypothesen sowie wiederum die abschließende Darstellung in Abbildung 4.

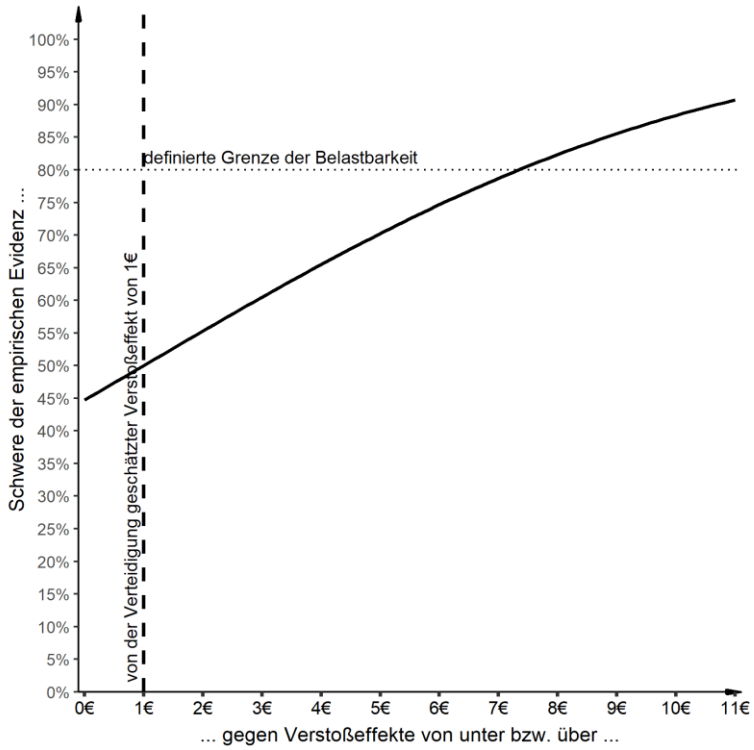
Hierbei wird für unterschiedliche Werte x die folgende Frage gestellt: Falls der tatsächliche Verstoßeffekt mindestens x wäre, mit welcher Wahrscheinlichkeit hätte dann der beobachtete Schätzwert *über* dem tatsächlich beobachteten Wert von 1 € liegen müssen. Je größer diese Wahrscheinlichkeit, desto schwerer wiegt die beobachtete Evidenz gegen den jeweiligen Wert x. Für x gleich 0 € kennen wir wiederum bereits die Antwort: 45 %, wie in Abbildung 2 dargestellt. Dabei wurde auf den Wert p-Wert abgestellt, d.h. auf die Irrtumswahrscheinlichkeit, woraus lediglich abgeleitet wurde, dass die Ausgangshypothese „kein Verstoßeffekt“ nicht auf dem gewählten Signifikanzniveau verworfen werden

[6] Dabei wäre die Frage, mit welcher Wahrscheinlichkeit der Verstoßeffekt einen bestimmten Wert annimmt, nicht sinnvoll, da a priori für jeden Wert die Wahrscheinlichkeit Null ist. Allerdings kann die übliche Teststatistik auch keine Aussagen dazu leisten, mit welcher Wahrscheinlichkeit der tatsächliche Wert in einem bestimmten Intervall liegt. Hier liegt oft ein Missverständnis bzw. eine fehlerhafte Interpretation vor. Solche Aussagen lassen sich mit der Bayesianischen Statistik treffen, die allerdings in der Regel nicht angewandt wird (und eigene Probleme aufwirft).

kann. Allerdings ist bereits aus Abbildung 2 ebenfalls ersichtlich, dass die Schätzung insgesamt mit einer erheblichen Ungenauigkeit behaftet ist, daher die „Glockenkurve“ relativ breit ist (so im Vergleich zu Abbildung 1). Dies spiegelt sich nun unmittelbar in der Bewertung anderer Ausgangshypothesen, d.h. anderer möglicher Verstoßeffekte x, wider.

So ist die Schwere der Evidenz (abgetragen auf der vertikalen Achse) erst ab dem Wert von ca. 7,3 € gleich der Schwelle von 80 %. Daher: Falls der tatsächliche Verstoßeffekt mindestens 7,3 € wäre, so würde man mit einer Wahrscheinlichkeit von 80 % einen höheren Schätzwert als den beobachteten Wert von 1 € ermittelt haben. Legt man diese Schwelle an, so lassen sich aufgrund der Schätzunsicherheit lediglich Werte über 7,30 € ausschließen: Die Evidenz wiegt nur hinreichend schwer gegen Werte über 7,30 €. Abbildung 4 stellt wiederum die entsprechende Schwere der Evidenz (vertikale Achse) für alle Ausgangshypothesen (horizontale Achse) dar.

Abbildung 4: Schwere der Ablehnung verschiedener möglicher Verstoßeffekte bei Annahme



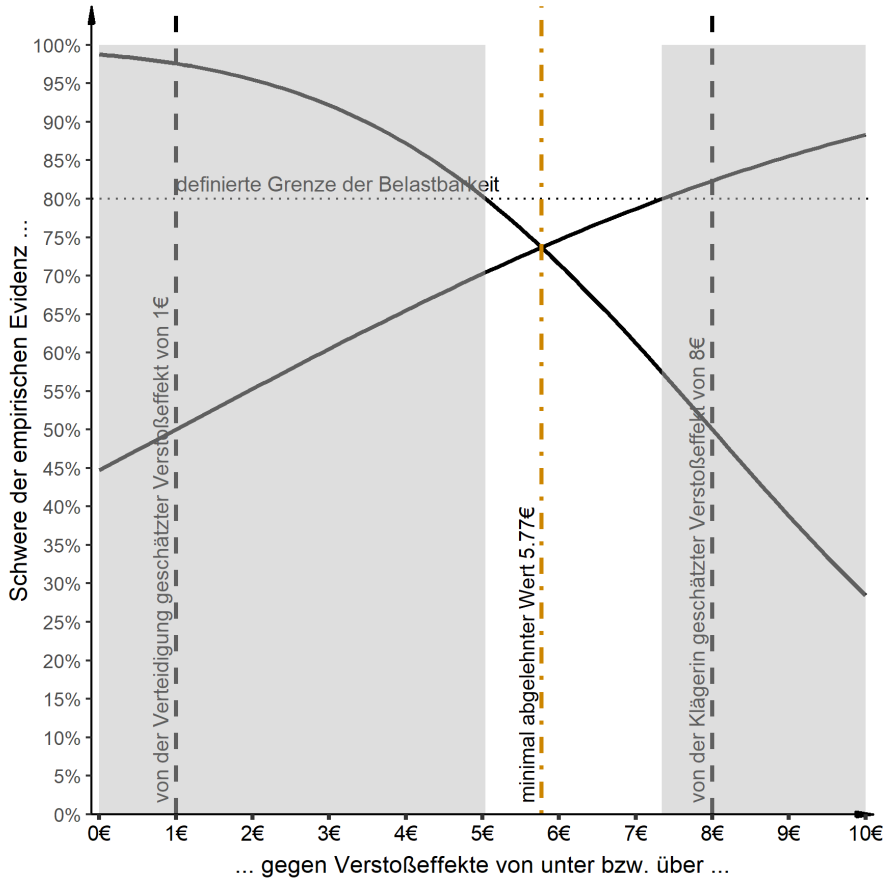
5. Ein Vorschlag zur Gesamtwürdigung (vermeintlich) widersprüchlicher Evidenz

Nachdem das Konzept der „severity“ oder „Schwere“ der Evidenz eingeführt und für jede der beiden Schätzungen angewendet wurde, wollen wir nun die häufig auftretende Situation vermeintlich widersprüchlicher Evidenz betrachten. Lediglich vermeintlich widersprüchliche Evidenz kann hierbei beispielsweise vorliegen, wenn fälschlicherweise die Ablehnung der Ausgangshypothese „kein Verstoßeffekt“ mit einem geschätzten Wert von 8 € zwingend als Evidenz für genau diesen Wert betrachtet wird und ebenso eine nicht signifikante Schätzung, hier von 1 €, zwingend als Evidenz gegen jeglichen Effekt betrachtet wird. Deshalb ist auch, wie dargestellt wurde, eine undifferenzierte Mittelung abzulehnen. Stattdessen wird nun eine Konsolidierung der Evidenz mittels der Verknüpfung der beiden Severity-

Kurven (Abbildung 3 und 4) geleistet. Dadurch ist die Berücksichtigung der beiden tatsächlich beobachteten Schätzwertes von in diesem Fall 1 € bzw. 8 € für eine differenzierte Evaluierung verschiedener potentieller Effekthöhen möglich. Dies ist in Abbildung 5 dargestellt.

In Abbildung 5 ist ebenfalls wieder der Schwellenwert von 80 % eingetragen – als eine zunächst nicht weiter motivierte Grenze der Belastbarkeit. Verwendet man diese Belastbarkeitsgrenze, so ergibt sich ein Intervall (die weiße Fläche zwischen 5 € und 7,3 €) von möglichen Verstoßeffekten, das mit beiden Ergebnissen hinreichend kompatibel ist. Genauer: Diese Werte für den Verstoßeffekt werden weder von der Evidenz der Klägerin noch von der Evidenz der Beklagten mit der als hinreichend erachteten Schwere zurückgewiesen. Daher: Sie werden weder als zu klein (gegenüber der Evidenz aus der statistisch signifikanten Schätzung von 8 €) noch als zu hoch (gegenüber der statistisch nicht signifikanten Schätzung von 1 €) erachtet.

Abbildung 5: Gesamtwürdigung vereinbarer Evidenz



Würde man die erforderliche Schwere für das Zurückweisen der alternativen Hypothese höher ansetzen, etwa auf 90 %, so wäre ein breiteres Intervall von Werten mit den beiden Schätzungen kompatibel, sprich: Es gäbe sowohl höhere als auch tiefere mögliche Verstoßeffekte, die von keinem vorliegenden Ergebnis nicht mit der für erforderlich erachteten Schwere zurückgewiesen werden. Andererseits verengt sich das Intervall, wenn man die erforderliche Schwere herabsetzt. Letzteres ist intuitiv – ebenso wie die Tatsache, dass ab einer hinreichend geringen Schwere das Intervall leer wird. Letzteres ist kein Artefakt, sondern liegt in der Natur der Sache. Dies soll kurz mit einer Schwelle von 70 % demonstriert werden.

In diesem Falle wäre man bereit, aufgrund des Schätzwertes von 8 € bereit alle Verstoßeffekte unter 7 € zu verwerfen, da dann der beobachtbare Schätzwert mit einer Wahrscheinlichkeit von 70 % unter dem beobachteten Wert von 8 € hätte liegen müssen. Ebenso schließt man mit der niedrigeren Schwelle von 70 % aufgrund des anderen Testergebnisses von 1 € ebenfalls ein breiteres Intervall an (höheren) Verstoßeffekten aus. Im Extremfall einer Schwelle von 50 % würde man schließlich jeden der beiden Schätzwerte jeweils als hinreichende Evidenz dafür nehmen, dass der tatsächliche Verstoßeffekt nicht unter 8 € bzw. über 1 € läge, was nicht nur offensichtlich zu einem leeren Intervall führt, sondern ebenso offensichtlich nicht sinnvoll ist: Denn die erforderliche Schwere käme dann einem Münzwurf gleich.

Der Schnittpunkt der beiden Kurven in Abbildung 5 ist der aufgrund der beiden vorgelegten Schätzwerte (1 € versus 8 €) „insgesamt am wenigsten abgelehnte" Verstoßeffekt, wobei diese Bezeichnung noch konkretisiert wird. Die Gesamtwürdigung folgt dabei der falsifikationistischen Logik, dass jede empirische Schätzung keine Evidenz für einen bestimmten Wert, sondern nur gegen eine bestimmte Ausgangshypothese liefert. Dieser Logik folgend, wäre der Verstoßeffekt im Schnittpunkt der beiden Severity-Kurven dadurch ausgezeichnet, dass *gegen jeden anderen höheren oder niedrigeren* Verstoßeffekt zumindest eine der beiden Evidenzen schwerer wiegt. Im vorliegenden Fall würde, wie sich an der vertikalen Achse beim Schnittpunkt ablesen lässt, keine der beiden Evidenzen schwerer als 74 % gegen die Wahl von 5,77 € wiegen.[7]

Damit kommen wir abschließend zur Beantwortung der in Abschnitt 2.4 aufgeworfenen Fragen:

1. Ist eine Konsolidierung der beiden (zunächst widersprüchlich erscheinenden) Ergebnisse möglich bzw. sind diese (gegeben einem geforderten Grad der Schwere) miteinander vereinbar oder nicht?
2. Wenn ja, bei welcher Verstoßeffektshöhe sollte diese Konsolidierung erfolgen?

Die erste Frage wird durch die Schwere (in Abbildung 5 ist diese 74 %) an der Stelle des Schnittpunktes beantwortet. Gibt es eine Überlappungsregion, welche unterhalb der geforderten Belastungsgrenze liegt, dann ist eine sinnvolle Gesamtwürdigung der Ergebnisse möglich. In diesem Fall kann mit gewisser Wahrscheinlichkeit davon ausgegangen werden, dass die vorgelegten Ergebnisse unterschiedliche Perspektiven auf das gleiche Phänomen geben und sich gegenseitig sinnvoll ergänzen. Bei einer angenommenen Belastungsgrenze von 80 % wäre das für das im in Abbildung 5 dargestellten Beispiel der Fall. Insbesondere wird der Schnittpunkt beider Kurven von keiner Evidenz mit einer Schwere höher als 74 % abgelehnt. Wählt man die Schwelle von 80 %, so liegt damit nur eine *vermeintlich* widersprüchlicher Evidenz vor. Der Widerspruch kam dann lediglich durch eine ungerechtfertigt mechanistische Interpretation der Ergebnisse zustande (vgl. Abschnitte 3.1 und 3.2). Tatsächlich ergibt sich ein Intervall von möglichen Werten, die bei Anwendung der 80 % Schwelle als hinreichend kompatibel mit der Evidenz beider Parteien erachtet werden kann.

Geht man nun einen Schritt weiter, um die zweite Frage zu beantworten, so liegt es nahe, so würde in einer Gesamtwürdigung die Verwendung des Schnittpunkt der beiden Severity-Kurven einen Verstoßeffekt von rund 5,77 € nahelegen. Eine solche Gesamtwürdigung der beiden Schätzergebnisse würde damit in der Regel gerade nicht mittig zwischen den Ergebnissen der beiden Parteien liegen. Letztere wäre entweder bei 4 €, falls man aus dem insignifikanten Ergebnis auf Null schließen würde,

[7] Genaugenommen werden die beiden Schätzungen als Evidenz gegen einen Verstoßeffekt von einerseits *mindestens* x und andererseits *höchstens* x interpretiert. In diesem Sinne ist die nun zusammenfassend gewählte Darstellung der Gesamtwürdigung *für* einen bestimmten Verstoßeffekt verkürzt. Tatsächlich wird eher deutlich, dass *gegen* den konsolidierten Verstoßeffekt von 5,77 € „am wenigsten" spricht.

oder aber bei 4,50 €, falls der Mittelwert von 1 € und 8 € gewählt wird. Die Tatsache, dass der Schnittpunkt von 5,77 € strikt über den Werten 4 € und 4,50 € liegt, ist der Tatsache geschuldet, dass die entsprechende Schätzung von 1 € eine geringere Ungenauigkeit aufweist, was, wie bereits dargestellt, an der unterschiedlichen Breite der „Glockenkurven" in den Abbildungen 1 und 2 ersichtlich ist.

Die Wahl des Schnittpunkts der Severity-Kurven berücksichtigt damit nicht nur die Höhe der Schätzergebnisse, sondern auch deren (relative) Genauigkeit.[8] Da die Konstruktion der Severity-Kurven stets dem Prinzip des Hypothesentest folgt, begründet sich die Wahl des Schnittpunkts darin, dass keine Evidenz entsprechend schwer *gegen* diese Wahl wiegt.[9] Es können dagegen aber keine Aussagen hinsichtlich der Wahrscheinlichkeit abgeleitet werden, mit der der tatsächliche Verstoßwert beispielsweise in einem bestimmten Intervall um den Schnittpunkt oder aber um einen beliebig anderen Punkt liegt. Die Betrachtung des Schnittpunkts allein berücksichtigt auch nicht, wie ungenau möglicherweise sogar beide Schätzungen sind. Wie bereits dargestellt wurde, ist das weiße Intervall in Abbildung 5 umso größer, je ungenauer die beiden Schätzungen sind. Denn dann wiegt weder der niedrige geschätzte Verstoßeffekt noch der höhere geschätzte Verstoßeffekt (hinreichend) schwer gegen viele mittlere Werte.

Die Frage nach der Schwere der Evidenz und die Konstruktion des Schnittpunkts der Severity-Kurven generieren damit einerseits nützliche zusätzliche Information und erlauben es, zunächst unter Umständen (vermeintlich) widersprüchliche Evidenz in einer Gesamtwürdigung zu kombinieren. Gerade dann, wenn auch die Belastbarkeit der gesamten vorgelegten Evidenz zu prüfen ist, ist aber andererseits auch hier von einer rein mechanistischen Bildung des Schnittpunktes zu warnen.[10]

6. Abschließende Überlegungen zum Fall unvereinbarer empirischer Evidenz

Im bisher herangezogenen Beispiel erwies sich die zunächst widersprüchlich erscheinende Evidenz als, bei Anwendung der Schwelle von 80 %, hinreichend kompatibel. Trotz der unterschiedlichen Schätzungen von 1 € und 8 € war dies möglich, da beide Schätzungen, insbesondere aber die Schätzung der Beklagten, mit hinreichender Unsicherheit versehen war. In Abbildung 5 war dies dadurch unmittelbar ersichtlich, dass sich die beiden Severity-Kurven unterhalb der angewendeten Schwelle von 80 % schnitten.

Nun ändern wir die Ausgangssituation wie folgt. Die Schätzung der Beklagten soll -1 € (anstatt zuvor 1 €) betragen. Und auch die Schätzung der Kläger soll extremer ausgefallen sein und bei 10 € liegen. Die entsprechenden Severity-Kurven sind bereits zusammenfassend in Abbildung 6 dargestellt. Abgesehen von den abgeänderten Schätzwerten besteht nun der wesentliche Unterschied darin, dass sich die beiden Kurven erst oberhalb der Schwelle von 80 % Schwere scheiden. Konkret bedeutet dies in

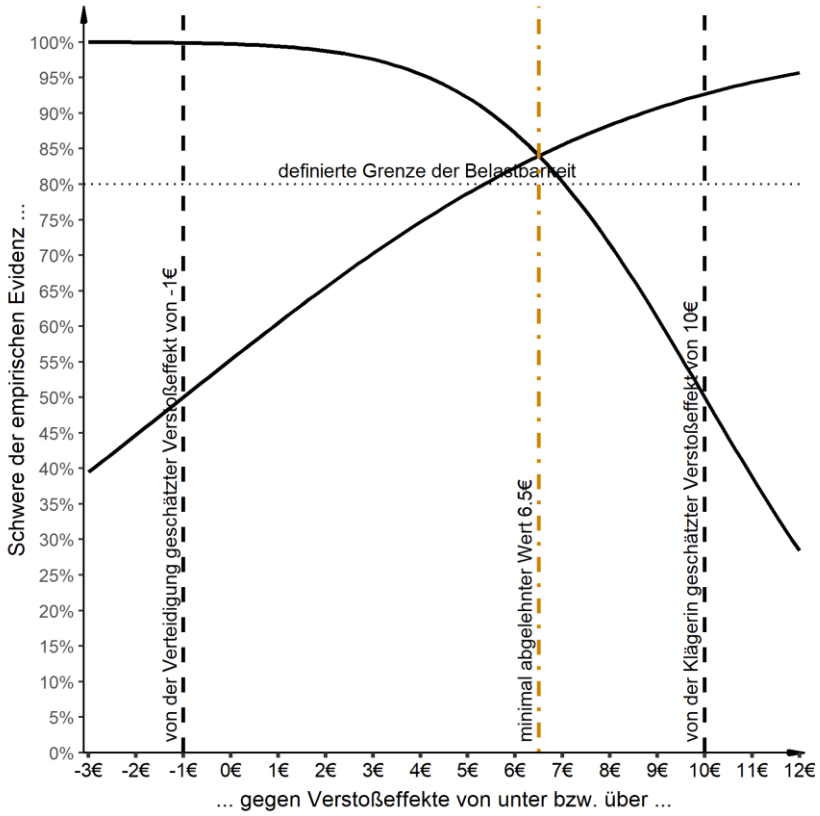
[8] Prinzipiell gibt es mit Sicherheit alternative und ggf. sogar vorzugswürdige Wege, um die unterschiedliche Evidenz zu verbinden, allen voran die Aggregation aller Beobachtungen in einem Datensatz mit anschließender erneuter Schätzung. Dies würde in der Regel den Einbezug eines unabhängigen Sachverständigen voraussetzen, sowie die Herausgabe der entsprechenden Datenpunkte und im einfachsten Fall auch deren Kompatibilität bzw. die Kompatibilität der gesamten Analysen (z.B. in Hinblick auf die zusätzlich verwendeten Variablen). Alternativ, wenn auch nach unserem Wissen in der Praxis so nicht verwendet, würde sich eine Mittelung gewichtet mit einer entsprechend normierten Schätzunsicherheit anbieten. In diesem Beitrag werden diese Alternativen nicht näher beleuchtet. In jedem Fall wäre auch in diesen Fällen die Frage zu beantworten, ob eine Gesamtwürdigung der Ergebnisse überhaupt sinnvoll möglich ist oder besser eine genauere Untersuchung der zugrundeliegenden Ergebnisse erfolgen sollte; siehe nachfolgend Abschnitt 6.

[9] Genauer gesagt spricht keine Evidenz gegen entsprechend höhere Werte, einschließlich des Schnittpunktes, oder entsprechend niedrigere Werte, einschließlich des Schnittpunkts.

[10] Dies gilt auch dann, wenn noch die Frage des „Ob" eines Effektes zu klären ist und das Gericht hierbei einen asymmetrischen Standard anwendet, so dass dann, wenn keine der beiden Parteien eine als belastbar angesehene Evidenz vorgelegt hat, das Gericht zugunsten der Beklagten entscheidet.

der hier vorgelegten Interpretation, dass sie auf dem geforderten Niveau nicht mehr kompatibel sind. Die Schätzung der Kläger von 10 € wiegt hinreichend schwer auch gegen mittlere Werte und insbesondere gegen all diese Werte, die als hinreichend kompatibel mit der Schätzung der Beklagten empfunden wurden (d.h. gegen die diese nicht hinreichend schwer wiegt).

Abbildung 6: Gesamtwürdigung unvereinbarer Evidenz



Wenn bei Annahme einer bestimmten Belastbarkeitsgrenze, sagen wir 80 %, kein nicht-abgelehnter Schnittpunkt verbleibt, dann muss eine tiefere Untersuchung der den Schätzergebnissen zugrundeliegenden Annahmen erfolgen. Es müssen entweder Rückfragen von den Parteien beantwortet oder ein Experte als Gutachter bestellt werden, der diese Fragen stellt bzw. für den Richter untersucht.

Natürlich ist angesichts der jeder Schätzung inhärenten Unsicherheit nie auszuschließen, dass die als unvereinbar empfundene Diskrepanz der Evidenz rein zufällig entstanden ist. Im Einzelfall können jedoch auch andere Gründe nahe liegen, so etwa ein „cherry picking“ bei der Wahl des Datensatzes oder des verwendeten ökonometrischen Modells. Betrachtet man in diesem Fall die Evidenz einer Partei isoliert, so kann dies nicht unmittelbar erkannt werden. Betrachtet man lediglich die Schätzergebnisse beider Parteien, so kann, wie bereits dargestellt, die Widersprüchlichkeit lediglich vermeintlich sein. Das Konzept der Severity bzw. Schwere der Evidenz, kompakt ausgedrückt in den Abbildungen 5 und 6, erlaubt dagegen auch eine Identifikation derjenigen Fälle, bei denen es nahe liegt, dass zumindest die Evidenz einer der Parteien, womöglich aber auch beider, einen entsprechenden „bias“ aufweist und eine Revision erfordert. Für ein solches „screening“ bieten sich allerdings auch andere Methoden an, die der Gegenstand eines weiteren Aufsatzes sein sollen.