ABC economics

ABC Working Paper 2026/04

# Ein signifikantes Missverständnis – Zur richterlichen Gewichtung von Regressionsanalysen in Kartellschadensersatzfällen*

Abstract

*Dieser Beitrag untersucht die Rolle statistischer Signifikanz bei der richterlichen Würdigung von Regressionsanalysen in Kartellschadensersatzfällen. Ausgehend von der Rechtsprechung des BGH zu Lkw II und Lkw III wird gezeigt, dass Signifikanz kein eigenständiges Gütekriterium einer Schätzung ist, sondern lediglich das Verhältnis von Effekthöhe und Präzision abbildet. Maßgeblich für die Aussagekraft einer Regressionsanalyse sind vielmehr die Erwartungstreue des Modells und die Präzision der Schätzung. Vor diesem Hintergrund wird erläutert, weshalb nicht-signifikante Ergebnisse nicht pauschal geringer gewichtet werden dürfen und weshalb Konfidenzintervalle für die Beweiswürdigung informativer sind als das binäre Etikett „signifikant" oder „nicht-signifikant". Ein Münzwurfbeispiel veranschaulicht, wie Fehlinterpretationen der Signifikanz zu einer verzerrten Gewichtung ökonometrischer Evidenz führen können. Abschließend wird ein kurzes Prüfschema für den praktischen Umgang mit divergierenden Regressionsergebnissen entwickelt.*

Keywords

*Kartellschadensersatz; Schadensquantifizierung; ökonometrische Gutachten; Regressionsanalysen; statistische Signifikanz; Konfidenzintervalle*

* Autoren: Simon Block, Dr. Falk Laser, Prof. Dr. Frank Maier-Rigaud und Florian Schimmel. Für Hinweise und Anmerkungen danken wir Dr. Ellen Braun, Prof. Dr. Wolfgang Kirchhoff und Prof. Dr. Daniela Seeliger.

## 1. Hintergrund

Im Urteil zu Lkw II benennt der BGH drei Kriterien zur Einordnung von Regressionsanalysen in Kartellschadensersatzfällen: (i) eine hinreichend zuverlässige Datengrundlage, (ii) die methodische Korrektheit und (iii) die Signifikanz der Ergebnisse.[1] Im Urteil zu Lkw III führt der BGH zum Kriterium der statistischen Signifikanz zudem aus, ein nicht-signifikantes Ergebnis bedeute *„lediglich, dass die Nullhypothese – wonach das Kartell keinen Preiseffekt hatte – nicht mit einer vom Gutachter als notwendig erachteten Sicherheit verworfen werden kann."*[2] Diese auch nach der 9. GWB Novelle noch relevant bleibenden Ausführungen des BGH zur Interpretation der statistischen Signifikanz sind inhaltlich nicht falsch, aber irreführend. Hintergrund ist, dass in jüngster Rechtsprechung, mit Verweis auf die Ausführungen des BGH, die statistische Signifikanz eines Regressionsergebnisses als eigenständiges Gütekriterium verwendet wird. Die statistische Signifikanz wird in diesen Fällen – losgelöst von der geschätzten Effekthöhe – genutzt, um im Rahmen der freien Beweiswürdigung zu entscheiden, welches Gewicht den Ergebnissen einer Regressionsanalyse beigemessen wird. Beispielsweise argumentiert das LG Stuttgart in seinem Urteil in Sachen Lkw-Kartell vom 27.02.2025 explizit, dass die insignifikante Schätzung der Beklagtenseite unter anderem deshalb geringer zu gewichten sei, da sie *„‚nur' ein statistisch nicht signifikantes Ergebnis"*[3] vorbringe.[4]

Die Ausführungen des BGH scheinen zu dem Missverständnis geführt zu haben, dass statistische Signifikanz (im Kontrast zu fehlender Signifikanz) ein eigenständiges Gütekriterium für die Aussagekraft einer Regressionsanalyse sei. Dadurch ist der Eindruck entstanden, signifikanten Ergebnissen komme von vornherein ein höherer Beweiswert zu als nicht-signifikanten Ergebnissen. Eine solche Hierarchisierung, die statistisch signifikanten Ergebnissen *a priori* ein höheres Gewicht beimisst als statistisch nicht-signifikanten Ergebnissen, ist jedoch nicht haltbar. Eine sachgerechte Bewertung von Regressionsanalysen erfordert vielmehr eine Auseinandersetzung mit den beiden übergeordneten Gütekriterien der Erwartungstreue und Präzision. Die statistische Signifikanz eines Schätzwerts steht zwar im Zusammenhang mit der Präzision einer Analyse, darf aber nicht als Maß für diese Präzision missverstanden werden, weil sie neben der Präzision, ganz mechanisch von der Höhe des Schätzwerts selbst abhängt. In diesem Beitrag wird dargestellt, dass *Konfidenzintervalle* das besser geeignete Mittel sind, um die Ergebnisse von Regressionsanalysen – ihrer Präzision angemessen – zu gewichten.

Im Folgenden wird zunächst anhand der beiden übergeordneten Gütekriterien der Erwartungstreue und Präzision einer Schätzung beschrieben, wie die drei BGH-Kriterien

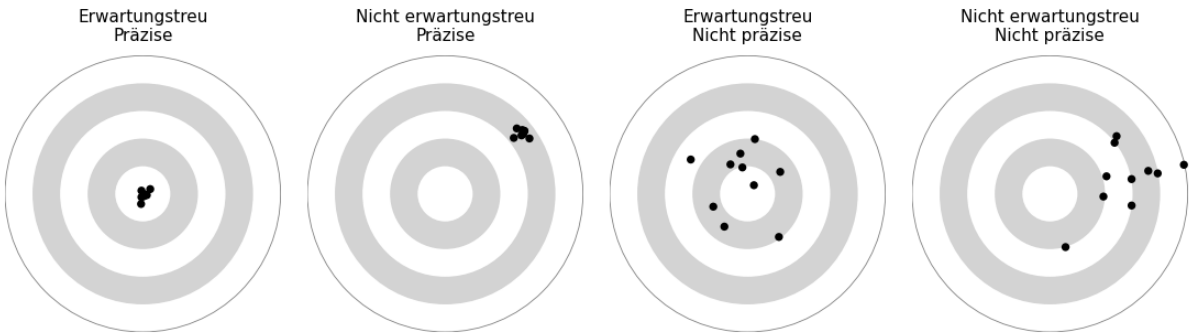
[1] Siehe BGH, Urteil vom 13. April 2021 – KZR 19/20, Rn. 66, Lkw-Kartell II.
[2] BGH, Urteil vom 05.12.2023 - KZR 46/21, Rn. 41, Lkw-Kartell III.
[3] Siehe LG Stuttgart, Urteil vom 27.02.2025 – 30 O 239/17, Rn. 308, Lkw-Kartell. Ein inhaltlich vergleichbares Urteil ist LG Stuttgart, Urteil vom 27.02.2025 – 30 O 235/17.
[4] Siehe zur geringeren Gewichtung nicht-signifikanter gegenüber signifikanten Ergebnissen auch LG Stuttgart, Urteil vom 27.02.2025 – 30 O 239/17, Rn. 316, Lkw-Kartell.

der methodischen Korrektheit, der belastbaren Datengrundlage und der statistischen Signifikanz zusammenhängen. Anschließend wird der Zusammenhang zwischen dem Konzept der statistischen Signifikanz und der Präzision eines Regressionsergebnisses erläutert. Auf dieser Grundlage wird anhand eines Beispiels herausgearbeitet, weshalb Konfidenzintervalle im Vergleich zur Signifikanz das geeignetere Instrument sind, um die Belastbarkeit eines Schätzergebnisses zu bewerten. Die folgenden Ausführungen konzentrieren sich ausschließlich auf die richterliche Würdigung von Regressionsanalysen als Indiz für oder gegen das Eintreten eines Kartellschadens. Die Bewertung anderer in Schadensersatzfällen relevanter Indizien wird aus dieser Betrachtung bewusst ausgeklammert.[5]

## 2. Übergeordnete Gütekriterien: Erwartungstreue und Präzision

Die vom BGH aufgeführten drei Kriterien der methodischen Korrektheit, Datengrundlage und (statistischen) Signifikanz sind nicht unabhängig voneinander und können nicht gegeneinander abgewogen werden. Grundsätzlich wirken sich sowohl die Datengrundlage als auch die Methodik auf die beiden zentralen Gütekriterien einer Regressionsanalyse aus: Erwartungstreue und Präzision. Eine Schätzung ist erwartungstreu, wenn sie bei häufiger Wiederholung mit unabhängigen Daten, im Durchschnitt das richtige Ergebnis erzielt. Die Präzision einer Schätzung gibt an, wie sehr sich die Ergebnisse einer wiederholten Schätzung untereinander ähneln. Der Unterschied zwischen Erwartungstreue und Präzision wird in Abbildung 1 symbolisch dargestellt.

**Abbildung 1: Unterschied zwischen Erwartungstreue und Präzision.**




*Quelle: ABC economics, eigene Darstellung.*

In Abbildung 1 werden die beiden statistischen Eigenschaften Erwartungstreue und Präzision anhand von Treffern auf einer Zielscheibe veranschaulicht. Die einzelnen Treffer symbolisieren jeweils eine Ausprägung eines Schätzers und der Mittelpunkt der Zielscheibe entspricht dem wahren Wert. Auf der linken Zielscheibe werden Ausprägungen eines erwartungstreuen und präzisen Schätzers abgebildet. Die Erwartungstreue des Schätzers ist daran erkennbar, dass sich die Treffer auf der Zielscheibe gleichmäßig um den Mittelpunkt anordnen. Die hohe Präzision des Schätzers

[5] Für eine Diskussion der Bedeutung der Schadenstheorie in Kartellverfahren siehe Bantle, Kaliske-Mielicki, Laser & Maier-Rigaud (2026).

zeigt sich an der geringen Streuung der Treffer. Die zweite Zielscheibe von links stellt die Ausprägungen eines präzisen, aber nicht erwartungstreuen Schätzers dar. Dass der Schätzer präzise ist, zeigt sich wieder an der geringen Streuung. Die Treffer gruppieren sich allerdings nicht um den Mittelpunkt, sondern der Schätzer ist systematisch verzerrt. Auf der dritten Zielscheibe von links werden die Ausprägungen eines unpräzisen, aber erwartungstreuen Schätzers dargestellt. Die Treffer sind gleichmäßig um den Mittelpunkt der Zielscheibe verteilt und die Schätzung führt somit im Mittel zum korrekten Ergebnis. Im Gegensatz zu den Treffern auf der linken Zielscheibe sind die Treffer aufgrund der hohen Streuung allerdings durchschnittlich weiter vom Mittelpunkt entfernt. Der Schätzer ist also weniger präzise. Auf der rechten Zielscheibe werden zuletzt die Ergebnisse eines weder erwartungstreuen noch präzisen Schätzers dargestellt. Die Ergebnisse gruppieren sich in diesem Fall nicht um den Mittelpunkt der Zielscheibe und die Streuung der Treffer untereinander ist hoch.

Sowohl methodische Fehler als auch eine schlechte Datengrundlage können eine Regressionsanalyse systematisch verzerren, also dazu führen, dass das Schätzergebnis nicht erwartungstreu ist. Dies kann zu statistisch signifikanten aber ökonomisch bedeutungslosen Regressionsergebnissen führen. Daher gilt es zunächst zu prüfen, ob derartige Fehler oder Verfälschungen vorliegen. Insbesondere muss die Spezifikation des Regressionsmodells untersucht werden, um sicherzustellen, dass alle notwendigen Kontrollvariablen einbezogen wurden,[6] und keine der einbezogenen Kontrollvariablen eine systematische Verzerrung verursacht.[7] Wenn davon auszugehen ist, dass eine systematische Verzerrung vorliegt, ist eine genauere Betrachtung der Regressionsergebnisse in der Regel nicht zielführend. Ob eine solche Verzerrung vorliegt, muss jedoch anhand grundlegender ökonometrischer Erwägungen und auf Grundlage der fallspezifischen ökonomischen Umstände geprüft werden. Anhand der Regressionsergebnisse selbst kann eine solche Verzerrung hingegen nicht festgestellt werden.

Wenn die Prüfung der Modellspezifikation und Datengrundlage ergibt, dass nicht von einer systematischen Verzerrung auszugehen ist,[8] gilt es den Blick auf die Regressionsergebnisse zu richten. Ins Zentrum der Betrachtung rückt dann die Frage, wie *präzise* die Schätzergebnisse der Regressionsanalyse sind. Die Präzision einer Regressionsanalyse hängt ebenfalls sowohl von der gewählten Methode als auch von der zugrunde liegenden Datengrundlage ab. Es gibt *a priori* viele unverzerrte Modellspezifikationen, die jedoch unterschiedlich gut für die Einflüsse weiterer Faktoren auf den Preis kontrollieren. Je besser ein Regressionsmodell für Preiseinflüsse anderer Faktoren kontrollieren kann, desto geringer ist die verbleibende statistische Unsicherheit

[6] Siehe Wooldridge (2013) zu Verzerrungen aufgrund ausgelassener Kontrollvariablen (omitted variable bias).
[7] Siehe beispielsweise Angrist & Pischke (2009) für eine Auseinandersetzung zu „schlechten Kontrollvariablen“ (bad controls).
[8] Siehe Inderst & Thomas (2021) für eine Unterscheidung zwischen unsystematischen und systematischen Datenfehlern und den jeweiligen Potenzialen, die Ergebnisse einer Regressionsanalyse zu verzerren.

und desto präziser lässt sich der Kartelleffekt selbst isolieren. Ebenso beeinflussen Umfang und Güte der Datengrundlage die Präzision einer Schätzung.

Ein statistischer Hypothesentest setzt schließlich die Höhe eines Schätzergebnisses ins Verhältnis zu dessen Präzision. Dadurch kann untersucht werden, ob ein von der sogenannten Nullhypothese verschiedenes Schätzergebnis auch zufällig entstanden sein könnte, oder ob die Nullhypothese durch das Schätzergebnis mit hinreichender Sicherheit verworfen werden kann. In letzterem Fall spricht man von einem statistisch signifikanten Ergebnis. Im folgenden Abschnitt werden die zentralen statistischen Grundkonzepte – insbesondere Konfidenzintervalle und statistische Signifikanz – genauer erläutert und die Zusammenhänge zwischen diesen Grundkonzepten aufgezeigt.

## 3. Auswertung von Regressionsanalysen

In diesem Abschnitt werden die Zusammenhänge zwischen den für diesen Artikel zentralen statistischen Grundkonzepten erläutert. Zunächst wird die Funktion von Standardfehlern und Konfidenzintervallen als Präzisionsmaße beschrieben. Dann wird erläutert, wie sich die statistische Signifikanz aus dem Schätzwert und der Präzision der Schätzung zusammensetzt. Auf dieser Grundlage wird gezeigt, dass Konfidenzintervalle das besser geeignete Mittel sind, um zu beurteilen, welches Gewicht einer Regressionsanalyse beizumessen ist.

### 3.1 Standardfehler und Konfidenzintervalle

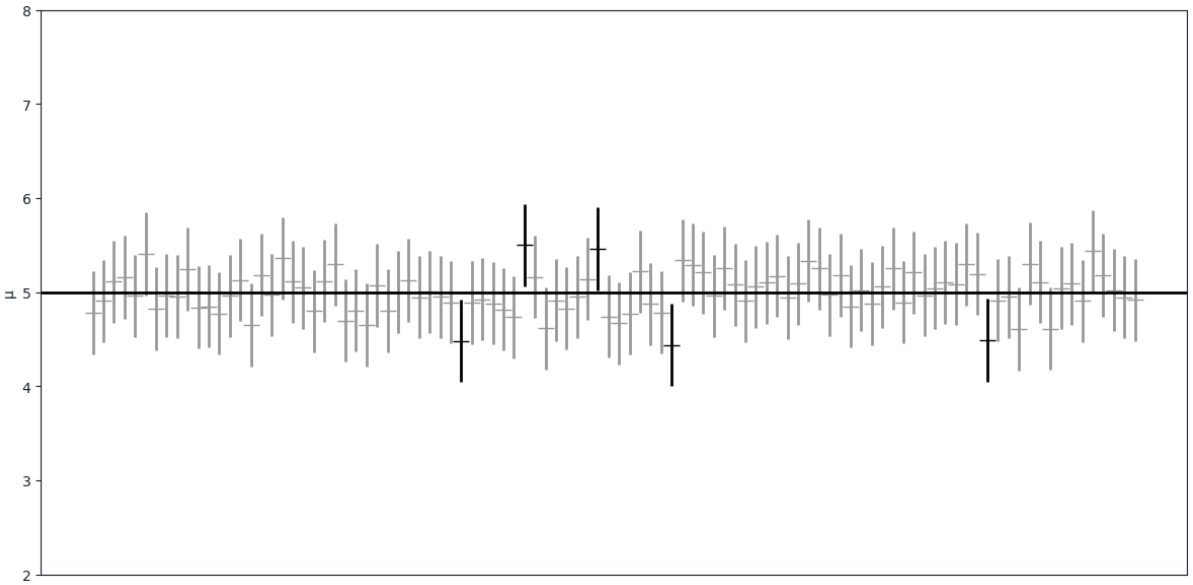
Eine Regressionsanalyse untersucht den Zusammenhang zwischen einer abhängigen Variablen und einer oder mehreren erklärenden Variablen. In Kartellschadensersatzfällen ist in der Regel der Preis die abhängige Variable.[9] Die erklärende Variable von Interesse ist zumeist ein Indikator für die Kartellbetroffenheit von Preisbeobachtungen.[10] Neben dem Kartellindikator werden in der Praxis oftmals eine Reihe weiterer Kontrollvariablen einbezogen, die den Einfluss weiterer Faktoren auf den Preis auffangen sollen. Dadurch soll verhindert werden, dass der Einfluss dieser weiteren Faktoren fälschlicherweise dem Kartellindikator zugeschrieben wird.

Das Regressionsergebnis beinhaltet für jede erklärende Variable einen Punktschätzer und einen Standardfehler. Der Punktschätzer misst den wahrscheinlichsten Effekt der erklärenden Variablen auf den Preis.[11] Der Standardfehler misst die Präzision des Punktschätzers; er ist kleiner – und die Präzision entsprechend höher – wenn Preisschwankungen möglichst eindeutig der betreffenden erklärenden Variablen zugeordnet werden können. Außerdem ist der Standardfehler umso geringer, je besser das Regressionsmodell Preisschwankungen insgesamt erklärt.

[9] In der Praxis wird zumeist der logarithmierte Preis als abhängige Variable verwendet, um den Kartelleffekt nicht als absoluten, sondern als prozentualen Preisaufschlag zu messen.
[10] Dieser Indikator wird auch als Kartell-Dummy bezeichnet und nimmt den Wert 1 an, falls eine Preisbeobachtung kartellbetroffen war und den Wert 0, falls die Beobachtung nicht kartellbetroffen war.
[11] Dies gilt nur, sofern ein korrekt spezifiziertes Regressionsmodell verwendet wird und der Punktschätzer dementsprechend erwartungstreu ist.

Anhand des Standardfehlers können Konfidenzintervalle um den jeweiligen Punktschätzer ermittelt werden. Das besonders häufig betrachtete 95 %-Konfidenzintervall ergibt sich beispielsweise, wenn man den Standardfehler multipliziert mit dem Faktor 1,96 zu beiden Seiten des Punktschätzers abträgt. Der Streckfaktor 1,96 ist so gewählt, dass in wiederholten Stichproben etwa 95 % der ermittelten Konfidenzintervalle den wahren Parameterwert einschließen würden.[12] Diese Interpretation des Konfidenzintervalls wird in Abbildung 2 graphisch veranschaulicht.

**Abbildung 2: 95 %-Konfidenzintervalle aus 100 unabhängigen Stichproben**




*Quelle: ABC economics, eigene Darstellung.*

In Abbildung 2 sind auf der y-Achse mögliche Koeffizienten abgetragen. Der wahre Koeffizient (in Höhe von 5) ist durch die waagerechte Linie gekennzeichnet. Die senkrechten Linien stellen 100 Konfidenzintervalle dar, die auf Basis von 100 unabhängigen Stichproben aus derselben Grundgesamtheit berechnet wurden. Helle Konfidenzintervalle enthalten den wahren Koeffizienten, dunkle hingegen nicht. Dem Erwartungswert entsprechend ist der wahre Wert in der in Abbildung 2 dargestellten Simulation in genau 95 der 100 Konfidenzintervalle enthalten.

Das Konfidenzintervall erlaubt folglich eine Abschätzung, in welchem Wertebereich ein Schaden zu vermuten ist. Je schmaler das Konfidenzintervall ist, desto höher ist die Präzision der Schätzung. Im Gegensatz dazu ist ein breites Konfidenzintervall Ausdruck einer hohen statistischen Unsicherheit.

[12] Dies gilt, sofern genügend Datenpunkte vorliegen, damit die Verteilung des Punktschätzers durch eine Normalverteilung approximiert werden kann.

## 3.2 Hypothesentests und statistische Signifikanz

Das Konzept der statistischen Signifikanz ist stets mit einem zugrunde liegenden Hypothesentest verknüpft. Bei einem Hypothesentest wird untersucht, ob ein Schätzwert – unter Berücksichtigung der Präzision der Schätzung – hinreichend unterschiedlich vom Wert der Nullhypothese ist. Die Nullhypothese eines Hypothesentests ist die Annahme, dass der wahre Koeffizient einem bestimmten Wert entspricht.[13] Unter dieser Annahme kann berechnet werden, wie wahrscheinlich bestimmte Punktschätzer durch reine Zufallsschwankungen zustande kommen können.

Ein Regressionsergebnis wird als statistisch signifikant bezeichnet, wenn der Punktschätzer so weit vom Wert der Nullhypothese abweicht, dass die Wahrscheinlichkeit einer derartigen oder noch höheren Abweichung – unter der Annahme der Nullhypothese – einen bestimmten Schwellenwert unterschreitet. Der verwendete Schwellenwert wird als Signifikanzniveau bezeichnet. Wenn das gewählte Signifikanzniveau unterschritten wird, spricht man auch davon, dass die Nullhypothese verworfen wird. In der Praxis wird häufig ein Signifikanzniveau von 5 % verwendet und als Nullhypothese ein Wert von Null – also die Abwesenheit eines statistischen Zusammenhangs – angenommen.[14]

Es gibt also *zwei* Voraussetzungen für ein statistisch signifikantes Regressionsergebnis: Der geschätzte Effekt muss hinreichend groß und die Schätzung muss hinreichend präzise sein. Diese beiden Voraussetzungen bedingen sich gegenseitig. Je höher der geschätzte Effekt, desto eher ist ein Regressionsergebnis auch bei geringerer Präzision noch statistisch signifikant. Entscheidend ist das Verhältnis zwischen geschätzter Effekthöhe und Präzision.

Ein Ergebnis ist nicht-signifikant, wenn der geschätzte Effekt – gemessen an seiner Präzision – nicht deutlich genug von einem Nulleffekt abweicht. Um ein statistisch nicht-signifikantes Schätzergebnis zu bewerten, ist es daher zwingend notwendig, die Präzision der Schätzung zu betrachten. Ein Schätzergebnis, das *aufgrund* mangelnder Präzision nicht statistisch signifikant ist, hat dementsprechend wenig Aussagekraft. Ein Schätzergebnis, das *trotz* hoher Präzision nicht statistisch signifikant ist – weil es präzise einen Wert nahe der Nullhypothese schätzt – hat hingegen eine hohe Aussagekraft.[15] Die Aussagekraft einer Schätzung hängt logischerweise nicht vom geschätzten Wert selbst ab, sondern von der Qualität der Schätzmethode mit der dieser Wert ermittelt wurde.

Ein signifikantes Ergebnis ist folglich per Definition als mögliches Indiz *für* einen Schaden zu werten. Vor diesem Hintergrund wird bereits deutlich, weshalb es sich bei der Interpretation der Signifikanz als Gütemerkmal eines Schätzergebnisses um ein

[13] Genauer gilt diese Beschreibung für einen *zweiseitigen* Hypothesentest. Bei einem einseitigen Hypothesentest ist die Nullhypothese, dass der wahre Koeffizient größer beziehungsweise kleiner als ein bestimmter Wert ist.
[14] Die Nullhypothese bezieht sich nicht zwingend auf den Wert 0, sondern beschreibt allgemein „keinen Effekt", beziehungsweise den Referenzwert.
[15] Siehe dazu auch Abadie (2020).

Missverständnis handelt. Besonders deutlich kommt dieses Missverständnis in der folgenden Ausführung des BGH im Urteil zu Lkw II zum Ausdruck:

*"Sie [die Regressionsanalyse] stellt damit zugleich - wenn sie auf einer hinreichend verlässlichen Datengrundlage methodisch korrekt und mit signifikanten Ergebnissen durchgeführt worden ist - ein relevantes Indiz für oder gegen den im Rahmen eines Grundurteils zu ermittelnden Umstand dar, dass der klagenden Partei durch den Kartellverstoß wahrscheinlich jedenfalls ein Schaden in irgendeiner Höhe entstanden ist."*[16]

Der BGH scheint die Signifikanz hier also als Qualitätskriterium einzuordnen, das vermeintlich auch bei Indizien *gegen* einen Schaden zum Tragen kommt. Da als Nullhypothese ein Nullschaden angenommen wird, ist ein signifikantes Ergebnis jedoch begriffsnotwendig unvereinbar mit einem Nullschaden. Eine Regressionsanalyse mit signifikantem Ergebnis als Indiz gegen die Entstehung eines Schadens ist daher in sich widersprüchlich.[17] Wie auch Schliffke (2024) anmerkt, würde die ausschließliche Berücksichtigung signifikanter Ergebnisse folglich die ökonomische Verteidigung gegen das „Ob" eines Kartellschadens anhand von Regressionsanalysen *de facto* unmöglich machen.

### 3.3 Zusammenhang von Hypothesentests und Konfidenzintervallen

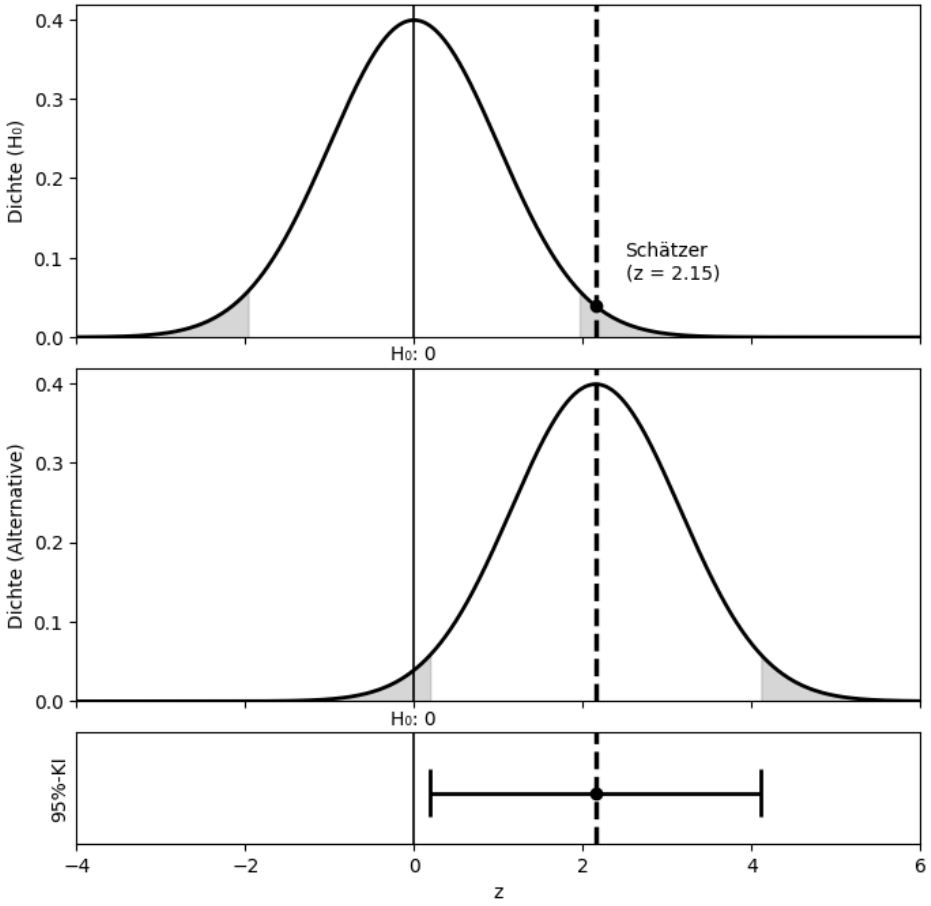
Am Konfidenzintervall lässt sich unmittelbar ablesen, ob ein Schätzergebnis statistisch signifikant ist. Dazu muss lediglich geprüft werden, ob der Wert der Nullhypothese, üblicherweise der Wert Null, im Konfidenzintervall enthalten ist. Ist der Wert der Nullhypothese nicht im Konfidenzintervall enthalten, wird die Nullhypothese abgelehnt.[18]

Da beispielsweise das 95 %-Konfidenzintervall per Konstruktion in 95 % der Fälle den wahren Wert einschließt, gilt im Umkehrschluss, dass *alle* Werte außerhalb des Konfidenzintervalls als Nullhypothese eines Hypothesentests zu einem Signifikanzniveau von 5 % abgelehnt werden. Das Konfidenzintervall erlaubt folglich, auf einen Blick eine Vielzahl von Hypothesentests durchzuführen. Dieser Zusammenhang zwischen Konfidenzintervall und Signifikanz wird in Abbildung 3 veranschaulicht. In der Abbildung

[16] BGH, Urteil vom 13.04.2021 - KZR 19/20, Rn. 66a), Lkw-Kartell II; eigene Hervorhebung.
[17] Im Urteil zu Lkw III stellt der BGH zwar klar, dass mit der zitierten Ausführung *„allgemein der Umstand angesprochen [ist], dass die statistische Signifikanz bei der Bewertung ökonometrischer Studien in den Blick zu nehmen ist und je nach den Umständen des Einzelfalls Einfluss auf die Bewertung der Schadensschätzung haben kann"* und auch *„ein statistisch insignifikantes Ergebnis bei der erforderlichen Gesamtbetrachtung einen Beitrag zur Interpretation der vorhandenen Daten und sonstigen qualitativen Indizien und damit zu einer Annäherung an die Wirklichkeit im Sinne einer Schätzung leisten"* kann. Dadurch wird der Widerspruch, eine signifikante Schätzung könne einen Nullschaden argumentativ stützen, allerdings nicht aufgelöst. Vielmehr bleibt der falsche Eindruck bestehen, es könne auch signifikante Regressionsergebnisse geben, die im Vergleich zu nicht-signifikanten Ergebnissen als „stärkeres" Indiz gegen den Eintritt eines Schadens streiten könnten.
[18] Genauer gilt: ist der Wert der Nullhypothese nicht im $X$ %-Konfidenzintervall enthalten, wird die Nullhypothese auf dem $(1 - X)$ %-Signifikanzniveau abgelehnt.

wird die Wahrscheinlichkeitsverteilung des Punktschätzers unter der Nullhypothese dem Konfidenzintervall gegenübergestellt.

**Abbildung 3: Zusammenhang von Hypothesentest und Konfidenzintervall**




*Quelle: ABC economics, eigene Darstellung.*

Im oberen Panel von Abbildung 3 ist die Verteilung der Teststatistik unter der Annahme der Nullhypothese dargestellt. Der beobachtete Schätzwert liegt im grau schattierten Randbereich dieser Verteilung, der unter der Nullhypothese nur mit geringer Wahrscheinlichkeit erreicht wird; die Nullhypothese wird daher verworfen. Das untere Panel zeigt denselben Befund in Form des 95 %-Konfidenzintervalls um den Punktschätzer. Da der Nullwert nicht im Intervall enthalten ist, liegt ein signifikantes Ergebnis vor. Das Konfidenzintervall enthält somit bereits alle notwendigen Informationen darüber, ob ein Schätzergebnis signifikant ist oder nicht.

Konfidenzintervalle enthalten somit mehr Informationen als das Etikett „*signifikant*“ oder „*nicht-signifikant*“. Da alle Werte außerhalb des Konfidenzintervalls als Nullhypothese abgelehnt werden und alle Werte innerhalb des Konfidenzintervalls nicht abgelehnt werden, beinhaltet das Konfidenzintervall die Informationen aus einer ganzen Bandbreite an möglichen Hypothesentests. Das Etikett „*signifikant*“ oder „*nicht-signifikant*“ bezieht sich hingegen immer nur auf einen einzelnen Hypothesentest. Auch Romer (2020) empfiehlt zur Interpretation ökonometrischer Ergebnisse die verstärkte

Auseinandersetzung mit Konfidenzintervallen. Schließlich zeigen diese, mit welchen Werten die jeweilige Schätzung vereinbar ist und gegen welche sie Indizien liefert.

## 4. Vom Münzwurf zum Missverständnis

In diesem Abschnitt wird das Zusammenspiel der zuvor beschriebenen statistischen Grundkonzepte veranschaulicht. Anhand eines Münzwurfbeispiels wird gezeigt, wie eine falsche Interpretation der statistischen Signifikanz als eigenständiges Gütekriterium zu einer falschen Gewichtung von Indizien führen kann.

### 4.1 Münzwurfbeispiel

Vier Personen sollen untersuchen, ob eine gegebene Münze fair oder verfälscht ist.[19] Dazu wird dieselbe Münze von jeder Person mehrmals geworfen und jeweils erfasst, wie häufig die Münze *„Kopf“* zeigt.

Person A wirft die Münze 100-mal und es erscheint 70-mal „Kopf“. Dieses Ergebnis legt bereits intuitiv die Vermutung nahe, dass die Münze verfälscht sein könnte. Es ist aber auch klar, dass das Ergebnis kein unumstößlicher Beweis für eine verfälschte Münze ist, da auch eine faire Münze zufallsbedingt bei 100 Würfen 70-mal „Kopf“ zeigen kann. Durch einen Hypothesentest kann überprüft werden, wie stark die Indizienwirkung der Beobachtungen von Person A ist. Die Nullhypothese ist in diesem Fall die Annahme einer unverfälschten Münze, also dass die Münze bei jedem einzelnen Wurf mit einer Wahrscheinlichkeit von 50 % „Kopf“ zeigt. Unter dieser Annahme liegt die Wahrscheinlichkeit, bei 100 Münzwürfen mindestens 70-mal dieselbe Seite zu erhalten, bei weniger als 0,01 % und damit weit unter den in den Wirtschaftswissenschaften gängigen Signifikanzniveaus von 1 %, 5 % oder 10 %. Person A kommt dementsprechend zu einem statistisch signifikanten Ergebnis, das dafürspricht, die Annahme einer unverfälschten Münze zu verwerfen.

Person B wirft die Münze 10-mal und es erscheint 7-mal „Kopf“. Die durchschnittliche Häufigkeit von „Kopf“ ist also bei Person A und Person B identisch und beträgt jeweils 70 %. Dennoch ist intuitiv klar, dass das Ergebnis von Person B statistisch weniger aussagekräftig ist als das von Person A. Die Wahrscheinlichkeit, dass eine unverfälschte Münze bei 10 Würfen mindestens 7-mal dieselbe Seite zeigt liegt mit ca. 34 % deutlich oberhalb aller gängigen Signifikanzniveaus. Demnach ist das Ergebnis von Person B nicht statistisch signifikant und würde für sich genommen nicht dafürsprechen, die Annahme einer unverfälschten Münze zu verwerfen. Das nicht-signifikante Ergebnis von Person B zieht den Befund von Person A jedoch auch nicht in Zweifel, sondern bestätigt ihn sogar im Punktschätzer, da in beiden Fällen die relative Häufigkeit, mit der „Kopf“ geworfen wird, bei 70 % liegt. Es liefert lediglich eine weniger präzise Evidenz aufgrund der kleineren Stichprobe.

[19] Eine Münze ist fair, wenn sie mit 50 % Wahrscheinlichkeit *„Kopf“* und mit 50 % Wahrscheinlichkeit *„Zahl“* zeigt. Andernfalls ist die Münze verfälscht.

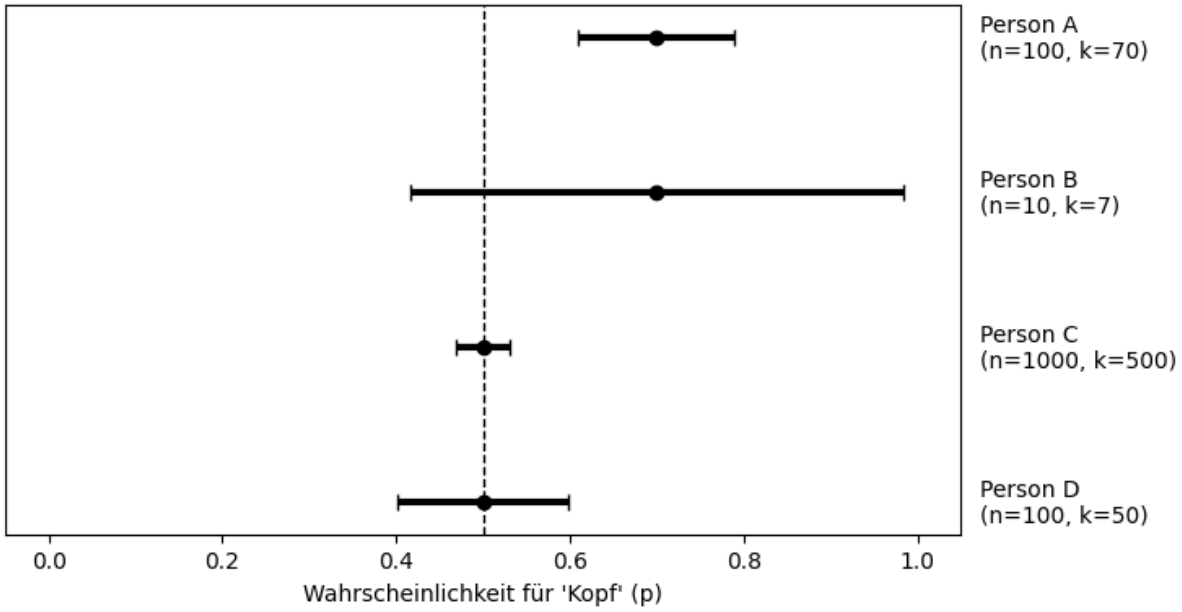
Person C wirft die Münze 1000-mal und es erscheint 500-mal „Kopf“. Die Wahrscheinlichkeit, dass eine Münze bei 1000 Würfen mindestens 500-mal dieselbe Seite zeigt liegt bei 100 %, mithin deutlich über den gängigen Signifikanzniveaus. Das Ergebnis von Person C ist demnach statistisch „maximal“ insignifikant. Es ist naheliegend, dass das statistisch nicht-signifikante Ergebnis von Person C in einem deutlichen Spannungsverhältnis zu dem statistisch signifikanten Ergebnis von Person A steht. Ferner ist intuitiv klar, dass das Ergebnis von Person C durch die höhere Anzahl an Beobachtungen eine höhere Präzision als das Ergebnis von Person A aufweist – es aufgrund seiner statistischen Nicht-Signifikanz geringer zu gewichten wäre offensichtlich verfehlt.

Person D wirft die Münze 100-mal und es erscheint 50-mal „Kopf“. Das Ergebnis von Person D ist somit ebenso wie das Ergebnis von Person C maximal nicht-signifikant. Es sollte jedoch nicht genauso gewichtet werden wie das Ergebnis von Person C, da es aufgrund der geringeren Stichprobengröße weniger präzise ist.

## 4.2 Übertragung auf Schadensersatzfälle

Die Annahme einer unverfälschten Münze entspricht, in Analogie zur Schadenquantifizierung, der Nullschadenshypothese in Schadensersatzfällen. Gleichermaßen entspricht eine etwaige Verfälschung der Münze einem positiven Preiseffekt des Verstoßes, jeweils feststellbar durch Abweichungen von der Nullhypothese. Zur weiteren Veranschaulichung werden in Abbildung 4 die Fallbeispiele mit den entsprechenden Konfidenzintervallen dargestellt. Analog zur Darstellung unterschiedlicher Ergebnisse bei Preisaufschlagsschätzungen bilden die Konfidenzintervalle im Münzwurf-Beispiel neben dem Punktschätzer auch den Präzisionsgrad der Schätzung ab.

**Abbildung *4*: Konfidenzintervalle der im Beispiel beschriebenen Fälle.**




*Quelle: ABC economics, eigene Darstellung.*

Im Münzbeispiel ließe sich eine Integration der jeweiligen Schätzergebnisse vergleichsweise einfach durch die Zusammenlegung der Beobachtungen erreichen. Würden etwa die Würfe der Personen A und B gemeinsam ausgewertet, ergäbe sich weiterhin ein hochsignifikanter Schätzwert von 0,7.[20] Eine gemeinsame Auswertung der Beobachtungen von Person A und Person C würde hingegen zu einem Schätzwert nahe 0,5 führen, der statistisch nicht-signifikant wäre.[21] Gerade hierin zeigt sich jedoch der Unterschied zwischen der anschaulichen Einfachheit des Münzbeispiels und der Komplexität in der Schadensquantifizierung. In der Praxis beruhen Regressionsanalysen nicht auf identischen und frei kombinierbaren Datensätzen, sondern auf unterschiedlichen Modellannahmen und Datenquellen. Eine schlichte „Zusammenlegung" wäre daher in der Praxis weder möglich noch sachgerecht. Die Entwicklung einer statistisch angemessenen Gesamtbetrachtung erfordert vielmehr einen Vergleich der jeweiligen Schätzungen unter Berücksichtigung ihrer Präzision – wofür sich insbesondere ein Abgleich der Konfidenzintervalle anbietet.

Ein Abgleich der Schätzungen von Person A und Person B zeigt, dass beide Ergebnisse sich im Punktschätzer gleichen. Die Konstellation gleicht jenem Beispiel, das von Hürten (2022) abgebildet und vom BGH aufgegriffen wird, bei dem „*ein statistisch insignifikantes Ergebnis bei der erforderlichen Gesamtbetrachtung einen Beitrag zur Interpretation der vorhandenen Daten […] leisten [kann], indem es beispielsweise das Ergebnis einer statistisch signifikanten Schätzung eines Preiseffekts bestätigt*."[22] Würden nicht-signifikante Schätzungen stets eine solche Form annehmen, wären sie im Rahmen der freien Beweiswürdigung tatsächlich von geringem Beweiswert. Dem ist jedoch mitnichten so.

Dass diese Konstellation – die in der Praxis der Kartellschadenschätzung einen Ausnahmefall darstellt – zu kurz greift, zeigt die Hinzunahme des Ergebnisses von Person C. Wie Abbildung 4 zeigt, ist die Präzision der Schätzung von Person C deutlich größer als die der Schätzungen der Personen A und B. Die Konstellation zwischen Person A und Person C – die in der Praxis häufig in ähnlicher Weise anzutreffen ist – zeigt beispielhaft, weshalb eine pauschale Mindergewichtung nicht-signifikanter Ergebnisse eindeutig verfehlt wäre.[23] Aufgrund seiner größeren Präzision käme der Punktschätzung von Person C in der Beweisführung richtigerweise sogar ein deutlich größeres Gewicht zu als den jeweiligen Punktschätzungen von Person A und Person B.

Dieses Argument hängt nicht an den Stichprobengrößen der Schätzungen. *A priori*, also vor Kenntnis der jeweiligen Präzision, sind signifikante und nicht-signifikante Punktschätzungen zunächst gleichermaßen zu gewichten. Dies wird am Vergleich der

[20] Aus einer solchen Zusammenlegung würde sich eine Beobachtung von 110 Münzwürfen mit 77 Mal „Kopf" ergeben. Der Punktschätzer einer gemeinsamen Schätzung läge bei genau 0,7.

[21] Aus einer solchen Zusammenlegung würde sich eine Beobachtung von 1100 Münzwürfen mit 570 Mal „Kopf" ergeben. Der Punktschätzer einer gemeinsamen Schätzung läge bei etwa 0,52.

[22] BGH, Urteil vom 05.12.2023 - KZR 46/21, Rn. 41, Lkw-Kartell III in Bezugnahme auf Hürten (2022).

[23] Auch Frank et al. (2019) weisen darauf hin, dass die Nicht-Signifikanz eines Ergebnisses allein nicht informativ ist, sondern stets mit Blick auf die Präzision der Schätzung zu betrachten ist. An einem knappen Beispiel erläutern die Autoren, dass ein aufgrund mangelnder Präzision nicht-signifikanter Preisaufschlag von 20% in der Regel anders zu interpretieren sei als ein nicht-signifikanter Aufschlag von 2%.

Schätzungen von Person B und Person D deutlich. Wie Abbildung 4 zeigt, gibt es, abgesehen von der Lage des jeweiligen Konfidenzintervalls, keinen grundlegenden Unterschied zwischen der signifikanten Schätzung von Person A und der nicht-signifikanten Schätzung von Person D. Insbesondere ist die Präzision bei beiden Schätzungen identisch. Eine unterschiedliche Gewichtung käme schlicht einer vorweggenommenen Diskriminierung einer Schätzung aufgrund ihrer Schätzhöhe gleich. Wie die Schätzung von Person A die Nullhypothese (hier 0,5) verwirft, so verwirft die Schätzung von Person D umgekehrt den Punktschätzer von Person A (hier 0,7). Aus statistischer Sicht ergibt sich eine Hierarchisierung von gleichermaßen validen bzw. erwartungstreuen Schätzungen jedoch allein aus den jeweiligen Präzisionsgraden, keinesfalls aber aus der Signifikanz oder Nicht-Signifikanz der Ergebnisse.

Es liegt daher nahe, dass die Interpretation des LG Stuttgart, wonach statistisch nicht-signifikante Ergebnisse grundsätzlich geringer zu gewichten seien als signifikante Ergebnisse, das Resultat einer fälschlichen Verallgemeinerung eines Ausnahmefalls ist. Solche Ausnahmefälle sind möglich, wie die Konstellation zwischen Person A und Person B zeigt; den praktischen Regelfall beschreibt diese Konstellation jedoch nicht. Eine Erklärung, die – wie Hürten (2022) – allein auf einen solchen Sonderfall abstellt, zeichnet folglich ein einseitiges Bild und kann Fehlinterpretationen der statistischen Signifikanz befördern. Eine mögliche Erklärung für den wissenschaftstheoretischen Ursprung des zugrunde liegenden Missverständnisses wird im folgenden Abschnitt diskutiert.

## 5. Ein möglicher Ursprung des Missverständnisses

Der Ursprung des Missverständnisses liegt möglicherweise in einer Übertragung des aus der Wissenschaftstheorie bekannten Falsifikationsschemas, wie es insbesondere von Karl Popper geprägt wurde.[24] Nach dieser Logik lassen sich allgemeine empirische Aussagen – etwa die Behauptung *„alle Schwäne sind weiß“* – nicht endgültig verifizieren, wohl aber durch ein einziges Gegenbeispiel falsifizieren. Wird ein schwarzer Schwan beobachtet, ist die Hypothese logisch widerlegt; werden hingegen ausschließlich weiße Schwäne beobachtet, bleibt sie lediglich *vorläufig* bestehen. In diesem Zusammenhang ist der Nachweis einer Abweichung von der ursprünglichen Behauptung also durchaus gewichtiger als eine Bestätigung der ursprünglichen Behauptung. Überträgt man diese Denkfigur auf statistische Ergebnisse, könnte der Eindruck entstehen, eine signifikante Schätzung falsifiziere die Nullhypothese, während eine nicht-signifikante Schätzung lediglich zeige, dass eine solche Falsifikation vorläufig nicht gelungen ist. So schreibt beispielsweise das LG Stuttgart mehrfach, dass die Nullhypothese durch ein nicht-signifikantes Regressionsergebnis „nicht falsifiziert“ werde.[25]

Eine solche Übertragung des Falsifikationsschemas wäre jedoch verfehlt. Hypothesentests operieren nicht auf der Ebene logischer Behauptungen, sondern auf der Grundlage probabilistischer Modelle. Schließlich finden sie ihre Anwendung nicht in

[24] Popper (1934/1994).
[25] LG Stuttgart, 27.02.2025 – 30 O 239/17, Lkw-Kartell, Rn. 315 – 316.

Bezug auf universell-deterministische Aussagen wie *„alle beobachteten Werte sind gleich Null“* – eine solche Aussage würde eine Falsifikation durch Gegenbeispiel erlauben, wäre im Zusammenhang mit Kartellschadensersatzfällen allerdings unsinnig. Stattdessen bieten sie eine statistisch fundierte Entscheidungsregel für probabilistische Aussagen wie *„die Daten entstehen aus einer Verteilung mit Erwartungswert Null“*, die Ausreißer zulassen und dementsprechend nicht logisch zu widerlegen sind. Folglich ist das Verwerfen einer Hypothese im Sinne einer Entscheidungsregel klar von der Falsifikation einer Hypothese im Sinne einer logischen Schlussfolgerung zu unterscheiden.[26]

Im Kontext von Hypothesentests wird eine Nullhypothese auf Grundlage der Daten und des Regressionsmodells entweder verworfen oder nicht verworfen. Diese Entscheidung hängt davon ab, ob die beobachteten Daten unter Annahme der Nullhypothese hinreichend unwahrscheinlich sind oder nicht, und liefert somit wertvolle Indizien für oder gegen einen Effekt. Begriffe wie „Beweis“ und „Falsifikation“ gehören hingegen in erster Linie in den Bereich der Logik. Der BGH weist also richtigerweise darauf hin, dass ein nicht-signifikantes Ergebnis die Nullhypothese lediglich nicht verwerfe und dementsprechend keinen Nachweis dafür darstelle, dass kein Schaden eingetreten ist. Es bedeutet lediglich, dass die Daten hinreichend vereinbar mit der Annahme der Nullhypothese sind. Doch auch ein statistisch signifikantes Ergebnis, das folglich die Nullhypothese verwirft, bedeutet nicht, dass die Nullhypothese logisch falsifiziert wäre. Vielmehr besagt es lediglich, dass die beobachteten Daten unter der Annahme der Nullhypothese hinreichend unwahrscheinlich erscheinen.

Die Entscheidung eines Hypothesentests – Verwerfen oder Nicht-Verwerfen der Nullhypothese – ist daher keine logische Falsifikation, sondern eine statistische Entscheidungsregel unter Unsicherheit. Aus dieser Struktur folgt insbesondere keine grundsätzliche Hierarchie zwischen signifikanten und nicht-signifikanten Schätzungen. Beide stellen lediglich Punktschätzungen mit unterschiedlicher Lage relativ zur Nullhypothese und unterschiedlicher Präzision dar. Die statistische Signifikanz beschreibt daher kein eigenständiges Gütemerkmal einer Schätzung, sondern lediglich das Verhältnis zwischen Effekthöhe und statistischer Unsicherheit.

## 6. Zum Umgang mit Regressionsergebnissen in der Praxis

Wie sollten Gerichte Regressionsergebnisse als Beweismittel vor dem Hintergrund statistischer Unsicherheit und unterschiedlicher Präzisionsgrade einordnen? Diese Frage stellt sich insbesondere dann, wenn voneinander abweichende Ergebnisse verschiedener Gutachter vorliegen. Auf Grundlage der bisherigen Ausführungen stellen wir ein Prüfschema vor, das die Bewertung der Evidenz unterschiedlicher Regressionsergebnisse erleichtern soll.

[26] Auch ökonomische Beiträge machen diese Unterscheidung allerdings nicht immer deutlich und haben so möglicherweise zur Entstehung des Missverständnisses beigetragen. Beispielsweise schreiben Frank et al. (2019): *„Dabei bauen die Hypothesentests auf dem Falsifikationsprinzip auf.“* (S. 52) Diese Aussage intendiert wohl keinen Bezug auf logische Falsifikation im Popper‘schen Sinne, kann aber leicht so verstanden werden.

Mit Blick auf die gerichtliche Praxis ist an dieser Stelle die Unterscheidung zwischen der Frage nach dem „Ob“ und nach der Höhe eines Schadens bedeutsam. Die binäre Logik der statistischen Signifikanz ist dabei besonders anschlussfähig an die Frage nach dem „Ob“. Gleichwohl erschöpft sich die statistische Aussage von Regressionsanalysen nicht in dieser Dichotomie. Das Konfidenzintervall zeigt sowohl, ob ein Nullschaden mit den Daten vereinbar ist, als auch, in welchem Bereich sich die Höhe eines möglichen Schadens in Abhängigkeit von der Präzision der Schätzung vermutlich bewegt. Es bietet damit eine integrierte Grundlage für die Bewertung von „Ob“ und Höhe des Schadens.

## 6.1 Vorgelagerte Prüfung der Erwartungstreue anhand der Plausibilität der Modellannahmen und der Belastbarkeit der Datengrundlage

Damit Regressionsergebnisse sinnvoll interpretierbar sind, muss zunächst gewährleistet sein, dass sie sich nicht auf zweifelhafte Modellannahmen stützen oder auf Grundlage fehlerhafter Daten ermittelt wurden. Würde beispielsweise eine so große Diskrepanz wie zwischen den Schätzungen von Person A und Person C im Münzwurfbeispiel in einem Schadensersatzfall zwischen verschiedenen Parteigutachten auftreten, so wären zunächst die jeweiligen Modellannahmen und Datengrundlagen zu prüfen. Schließlich sind die Schätzungen nur aussagekräftig, sofern sie die realen Gegebenheiten des Einzelfalls hinreichend sachgerecht abbilden. In einer solchen Konstellation wäre es also unwahrscheinlich, dass beide Schätzungen auf plausiblen Modellannahmen beruhen und auf belastbaren Daten aufsetzen.

Sollte sich bei näherer Untersuchung herausstellen, dass eine der beiden Schätzungen auf Modellannahmen beruht, die dem vorliegenden Fall nicht gerecht werden, sollte diese Schätzung nicht weiter berücksichtigt werden.[27] Gleichfalls gilt, dass Schätzergebnisse unberücksichtigt bleiben müssen, sollten sie auf nicht belastbaren Daten beruhen.[28] Eine gleichwertige Betrachtung beider Schätzungen ist hingegen nur dann geboten, wenn die Modellannahmen beider Schätzungen in ähnlich plausibler Weise die fallspezifischen Gegebenheiten abbilden, und wenn sie auf hinreichend belastbaren Daten durchgeführt werden. So ist es beispielsweise vorstellbar, dass beide vorgetragenen Modelle sich der komplexen Realität des Falls aus unterschiedlichen Richtungen annähern, sodass keines der Modelle eindeutig zu bevorzugen ist. Nur dann sind beide Schätzungen bei der Gesamtbetrachtung zu berücksichtigen.

Im Einzelfall müssen sowohl die Plausibilität der Modellannahmen als auch die Eignung der Datengrundlage stets nachgewiesen werden. In vielen Fällen dürfte das Resultat sein, dass zumindest eine Vorgehensweise durch sorgfältige Untersuchung der Modellannahmen und der Daten seitens der Gerichte (ggf. mit Unterstützung eines Sachverständigen) verworfen werden muss und die Ergebnisse von der weiteren Betrachtung ausgeschlossen werden. In solchen Fällen entfällt die Notwendigkeit,

[27] Zu dieser zentralen Stellschraube siehe Bantle, Kaliske-Mielicki, Laser & Maier-Rigaud (2026).
[28] Nicht belastbare Daten sind nicht unbedingt Daten die fehlerbehaftet sind, sondern können auch einfach Daten auf einem falschen Aggregationsniveau sein. Siehe dazu Bantle, Kaliske-Mielicki, Laser & Maier-Rigaud (2026).

widersprüchliche Schätzungen abzugleichen. Wie in Abschnitt 2 ausgeführt, kann nur bei sachgerechter Modellierung und verlässlicher Datengrundlage davon ausgegangen werden, dass die Schätzergebnisse eine hinreichende Erwartungstreue aufweisen. Schließlich bietet selbst eine noch so präzise Schätzung keinen Erkenntnisgewinn, wenn sie systematisch verzerrt ist.

### 6.2 Konfidenzintervalle als Maßstab für die Präzision einer Schätzung

Erst wenn sichergestellt wurde, dass die Regressionsergebnisse interpretierbar sind, da sie auf plausiblen Modellannahmen sowie einer belastbaren Datenbasis beruhen, kann die Präzision der Schätzergebnisse sinnvoll eingeordnet werden. Diese lässt sich anhand der Konfidenzintervalle abbilden. Schätzungen mit schmalen Konfidenzintervallen sind präziser, während breite Konfidenzintervalle auf eine geringere Präzision hinweisen. Die Breite der Konfidenzintervalle stellt daher – im Gegensatz zum bloßen Signifikanzkriterium – ein zentrales Gütemerkmal einer Schätzung dar.[29]

Zudem setzt dieses Bewertungskriterium erstrebenswerte Anreize für Parteigutachter. Die Berücksichtigung der Präzision einer Schätzung als Qualitätsmerkmal nimmt den Parteien den Anreiz, künstlich eine breite Streuung der Ergebnisse zu erzeugen – etwa um dadurch nicht-signifikante Schätzergebnisse herbeizuführen. Eine solche Streuung könnte beispielsweise dadurch erreicht werden, dass selektiv Daten oder Kontrollvariablen gewählt werden, die die Präzision der Schätzung verringern.[30] Mit anderen Worten: Die Einbeziehung eines Präzisionskriteriums sanktioniert unpräzise Analyseansätze und schafft stattdessen Anreize für präzisere Schätzungen.

### 6.3 Konsolidierung valider Regressionsergebnisse unter Berücksichtigung der Präzision

Nach Überprüfung der Modellannahmen und Datenqualität sowie Bewertung der Schätzergebnisse anhand ihrer Präzision verbleiben in einigen Fällen weiterhin widersprüchliche Regressionsergebnisse, die gegeneinander abzuwägen sind. Maßgeblich ist dabei nicht das Etikett „signifikant“ oder „nicht-signifikant“, sondern die jeweilige Präzision der Schätzung, wie sie sich im Konfidenzintervall widerspiegelt.

Mithilfe von Konfidenzintervallen lassen sich voneinander abweichende Schätzergebnisse nachvollziehbar vergleichen. Dabei kann insbesondere geprüft werden, ob und in welchem Umfang sich die von den jeweiligen Schätzungen eröffneten

[29] Wie in Abschnitt 3 beschrieben, hängen Konfidenzintervalle unmittelbar mit den Standardfehlern einer Schätzung zusammen. Diese unterliegen Modellierungsannahmen. Beispielsweise können Standardfehler geclustert werden, um für Abhängigkeiten zwischen den Fehlertermen zu korrigieren. Damit die Konfidenzintervalle unterschiedlicher Schätzungen vergleichbar sind, müssen die gleichen plausiblen Modellierungsannahmen getroffen werden.

[30] Andersherum darf die statistische Präzision einer Schätzung nicht unzulässig erhöht werden, indem Daten gezielt ausgewählt oder Modelle so angepasst werden, dass signifikante Ergebnisse bzw. Ergebnisse mit schmalen Konfidenzintervallen auf Kosten der Erwartungstreue erzeugt werden (sogenanntes p-Hacking). Um dem vorzubeugen, müssen Datenauswahl und Modellierung transparent und nachvollziehbar begründet werden. Dadurch wird sichergestellt, dass die Grundgesamtheit der verfügbaren Daten nicht willkürlich eingeschränkt oder Modellspezifikationen nicht gezielt angepasst werden, um „gewünschte“ Ergebnisse zu erzielen.

Wertebereiche überschneiden. Erstreckt sich etwa das Konfidenzintervall der einen Schätzung von -2 % bis 2 % und das der anderen von 1 % bis 10 %, so besteht eine Überschneidung im Bereich von 1 % bis 2 %. In diesem Bereich liegen Werte, die mit beiden Schätzungen vereinbar sind. Zugleich zeigt der Vergleich der Konfidenzintervalle, welche Schätzung den möglichen Schadensbereich enger eingrenzt und damit präziser ist. Gerade darin liegt ihr Mehrwert für die Beweiswürdigung: Sie erlauben nicht nur eine Aussage darüber, ob ein Ergebnis signifikant ist, sondern auch darüber, wie eng oder weit der mit der Schätzung vereinbare Bereich gezogen ist. Eine methodisch weiterführende Ausarbeitung dieses Gedankens findet sich bei Bönisch & Inderst (2020) mit dem Konzept der „Schwere" („severity") ökonometrischer Evidenz.

Ob ein rein statistischer Abgleich von Regressionsergebnissen in der Praxis sachgerecht ist, erscheint jedoch zweifelhaft, da Regressionsanalysen nur ein Indiz unter mehreren darstellen. Zu Beginn dieses Artikels wurden andere Indizien für die Beweisführung im Kartellschadensersatz bewusst ausgeklammert, um eine Fokussierung auf die Klärung statistischer Konzepte im Rahmen von Regressionsanalysen zu ermöglichen. In der Beweisführung würde eine bloße Betrachtung von Regressionsanalysen jedoch zu kurz greifen. Aus ökonomischer Sicht ist zunächst überhaupt eine Schadenstheorie notwendig, um die gewählten Modellannahmen zu stützen und die empirisch ermittelten Preisaufschläge zu plausibilisieren.[31] Von gerichtlicher Seite sollten die vorgetragenen Modelle und Schätzergebnisse stets im Lichte ebenjener Schadenstheorien betrachtet und entsprechend eingeordnet werden.[32] Das hier vorgestellte Prüfschema kann dabei helfen, die Präzision ökonometrischer Befunde mithilfe von Konfidenzintervallen transparent zu machen und ihren jeweiligen Stellenwert in der Gesamtwürdigung einzuordnen.

## 7. Fazit

Der Hinweis des BGH, dass ein nicht-signifikantes Ergebnis einen Nullschaden nicht beweisen kann, sondern lediglich die Nullhypothese nicht verwirft, ist für sich genommen richtig. Umgekehrt ist allerdings ebenso richtig, dass ein signifikantes Ergebnis einen Nullschaden nicht logisch falsifizieren kann, sondern lediglich im Sinne einer statistischen Entscheidungsregel auf einen Schaden hindeutet. Eine weitergehende Auslegung dieser Grundkonzepte als Rechtfertigung für eine geringere Gewichtung nicht-

[31] Damit die Schadensschätzung nachvollziehbar und empirisch tragfähig ist, muss die ökonomische Theorie zwei unterschiedliche, aber eng miteinander verknüpfte Funktionen erfüllen: Die Entwicklung einer fallspezifischen Schadenstheorie und einer Theorie des wettbewerblichen Verhaltens. Erst das Zusammenspiel beider theoretischen Perspektiven schafft das Fundament, auf dem eine empirische Schadensschätzung sinnvoll durchgeführt und interpretiert werden kann. Nur so wird vermieden, dass zufällige oder irrelevante Zusammenhänge fälschlicherweise als Schaden interpretiert werden. Eine theorielose empirische Betrachtung führt dagegen regelmäßig zu falschen Ergebnissen. Siehe Bantle, Kaliske-Mielicki, Laser & Maier-Rigaud (2026).

[32] Kann die Klägerin keine stichhaltige Theorie für eine Schädigung darlegen, also keinen Mechanismus aufzeigen, wie unter den konkreten Umständen eines Verstoßes auf dem entsprechenden Markt der Preissetzungsmechanismus beeinflusst worden sein könnte, darf ihrem Vortrag nicht allein mit dem pauschalen Hinweis auf die generell hohe Wahrscheinlichkeit eines Schadens bei Kartellrechtsverstößen besonderes Gewicht beigemessen werden. Anders verhält es sich, wenn die Klägerin eine überzeugende Schadenstheorie vorgelegt hat und die Beklagte diese nicht widerlegen kann. In diesem Fall kann beispielsweise selbst eine präzise Schätzung, die einen Nullschaden ergibt, in der Gesamtschau den Vortrag der Klägerin nicht zwangsläufig entkräften.

signifikanter Ergebnisse ist aus statistischer Sicht hingegen nicht nachvollziehbar. Nicht-Signifikanz bedeutet lediglich, dass der Nullwert im Konfidenzintervall der Schätzung enthalten ist – ebenso wie Signifikanz bedeutet, dass er außerhalb dieses Intervalls liegt. Dieses Kriterium kann also eine Ungleichbehandlung signifikanter und nicht-signifikanter Schätzungen nicht rechtfertigen.

Anstatt sich auf das Signifikanzkriterium zu versteifen, sollten Gerichte Regressionsergebnisse vorrangig anhand von Konfidenzintervallen bewerten. Diese enthalten neben Informationen zur Signifikanz auch Informationen über eine plausible Schadenshöhe und die Präzision einer Schätzung. Daher sind sie insbesondere auch das geeignetere Instrument, um widerstreitende Regressionsergebnisse in den Gesamtzusammenhang einzuordnen, der in der Regel auch qualitative Indizien beinhaltet. Diese qualitativen Indizien beeinflussen auch maßgeblich, ob es überhaupt notwendig ist, widerstreitende Regressionsergebnisse gegeneinander abzuwägen. In der Praxis dürften gleichwohl die Plausibilität der Modellannahmen, die Belastbarkeit der Datengrundlage und die ökonomische Einbettung der jeweiligen Schadenstheorie häufig bereits größere Schwierigkeiten bereiten als die korrekte Einordnung einzelner Signifikanzaussagen.

## Literaturverzeichnis

Abadie, A. (2020). Statistical nonsignificance in empirical economics. *American Economic Review: Insights*, *2*(2), 193-208.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Bantle, M., Kaliske-Mielicki, M., Laser, F., & Maier-Rigaud, F. (2026). Kartellschadensersatzprozesse im Blindflug – Zur zentralen Rolle der ökonomischen Theorie, *ABC Working Paper* 2026/03.

Bönisch, P., & Inderst, R. (2020). Zur Interpretation empirischer Evidenz vor Gericht. *Zeitschrift für Wettbewerbsrecht*, *18*(1), 52-68.

Frank, N., Inderst, R. & Oldehaver, G. (2019). Zur Diskrepanz zwischen gerichtlichen Beweisfragen in Kartellschadensersatzverfahren. *Zeitschrift für Wettbewerbsrecht, 17*(1), 39-61.

Hürten, J. (2022). Signifikanz signifikant überschätzt? – Anforderungen an ökonomische Parteigutachten. *Neue Zeitung für Kartellrecht,* 9/2022, 499-503.

Inderst, R., & Thomas, S. (2021). Zum Umgang mit Regressionsanalysen in Kartellschadensersatzfällen. *Zeitschrift für Wettbewerbsrecht*, *19*(4), 432-459.

Popper, K. (1934/1994). *Logik der Forschung*. 10. Auflage, Mohr Siebeck, Tübingen.

Romer, D. (2020). In praise of confidence intervals. *AEA Papers and Proceedings* 2020, 110: 55–60.

Schliffke, P. (2024). Zur (Un-)Möglichkeit einer ökonomischen Verteidigung gegen das „Ob“ eines Kartellschadens nach Lkw III. *Wirtschaft und Wettbewerb,* 5/2024, 255-262.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). South-Western Cengage Learning.