Discussion Papers of the
Max Planck Institute for
Research on Collective Goods
2021/2

Der Umgang mit Empirie beim Nachweis von Diskriminierung

Emanuel V. Towfigh

MAX PLANCK
SOCIETY

Discussion Papers of the
Max Planck Institute
for Research on Collective Goods 2021/2

Der Umgang mit Empirie beim Nachweis von Diskriminierung

Emanuel V. Towfigh

January 2021

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
https://www.coll.mpg.de

Der Umgang mit Empirie beim Nachweis von Diskriminierung

Emanuel V. Towfigh[*]

erscheint in:

Mangold/Payandeh, Handbuch Antidiskriminierungsrecht
Tübingen 2021 (Mohr Siebeck)

[*] Inhaber des Lehrstuhls für Öffentliches Recht, Empirische Rechtsforschung und Rechtsökonomik an der EBS Law School in Wiesbaden und Research Affiliate am Max-Planck-Institut zur Erforschung von Gemeinschaftsgütern in Bonn sowie Rechtsanwalt in Frankfurt am Main.

Für vielfältige Unterstützung und Hinweise bei der Erstellung dieses Manuskript danke ich *Christoph Engel*, *Anna Katharina Mangold*, *Mehrdad Payandeh*, *Katharina Towfigh* und *Katharina Wommelsdorf* sowie *Jan Keesen* und meinem Lehrstuhl-Team. Alle verbleibenden Unzulänglichkeiten und Fehler sind selbstverständlich allein mir zuzuschreiben.

Inhaltsverzeichnis

A. Ziele des Einsatzes empirischer Methoden im Antidiskriminierungsrecht

Bei der rechtlichen Bewertung von Ungleichbehandlungen stellt sich unweigerlich die Frage des Nachweises von Diskriminierung – ganz gleich welcher Art die Diskriminierung sein mag (ob mittelbar oder unmittelbar), weswegen diskriminiert wird (sei es etwa wegen des Geschlechts, der sexuellen Orientierung, der Religion oder Herkunft oder durch das intersektionale Zusammenwirken von Diskriminierungskategorien) oder wer diskriminiert (ob Private oder der Staat).[1] Denn nicht jeder Ungleichbehandlung oder Benachteiligung liegt eine rechtlich relevante Diskriminierung zu Grunde, und nicht jede subjektiv empfundene Diskriminierung ist objektiv als solche einzustufen;[2] umgekehrt können scheinbar harmlose Differenzierungen scharf diskriminieren, und gerade vage und deutungsoffene Konzepte bieten Raum für diskriminierende Unterscheidungen, die in der Folge glaubhaft bestritten werden können (***plausible deniability***).[3] Neben Diskriminierungen, die sich unmittelbar in einem konkreten Einzelfall durch bewusstes oder unbewusstes Handeln von Individuen offenbaren, ist der Nachweis solcher Diskriminierungen von Bedeutung, die systemisch sind – Konstellationen also, in denen äußere Gegebenheiten und Umstände oder besondere Verfahren diskriminierendes Verhalten strukturell begünstigen und zu systematischer Benachteiligung führen können.[4] Mit wachsender gesellschaftlicher Vielfalt treten solche diskriminierenden Strukturen stärker zu Tage, und es zeigt sich ein Bedarf an Werkzeugen, diese sichtbar zu machen. Nur wenn es gelingt, mit Hilfe sozialwissenschaftlicher Methoden auch die diskriminierenden *Mechanismen* zu identifizieren, könne Gesellschaft und Rechtstaat in ihrem Verantwortungsbereich gegensteuern.

Hierzu bedarf es evidenzbasierter, intersubjektiv vermittelbarer – also methodischer – Maßstäbe und Werkzeuge, die es ermöglichen, Diskriminierung nachzuweisen oder zu widerlegen; die Empirik hält solche vor: Diskriminierungen lassen sich mit Hilfe

---

1 *Makkonen,* Data Collection and EU Equality Law, S. 28 ff..

2 Zu Begriff und Konzept der Diskriminierung → *Mangold/Payandeh*, § 1.

3 *van Reenen,* Maintaining Plausible Deniability: Detecting Mechanisms of Subtle Discrimination in a South African Higher Education Institution*,* International Journal for Educational Sciences 13 (2016) S. 18 ff.; *Liu/Mills,* Modern Racism and Neoliberal Globalization: The Discourses of Plausible Deniability and their Multiple Functions, Journal of Community and Applied Social Psychology, 16 (2006) S. 83 ff; *Foblets,* Freedom of religion and belief in the European workplace: Which way forward and what role for the European Union?, International Journal of Discrimination and the Law, 13 (2-3) S. 240 (250).

4 → *Sacksofsky*, § 15.

statistischer Testverfahren nachweisen: Die Analyse einer **Stichprobe** erlaubt unter strengen Voraussetzungen den induktiven Schluss auf alle denkbaren Fälle (**Population**), wobei die statistischen Methoden Auskunft darüber geben, unter welchen Bedingungen und mit welcher Sicherheit wir diesen Schluss ziehen können.[5]

Dem Antidiskriminierungsrecht ist die Einbeziehung empirischer Methoden nicht fremd. Das AGG beispielsweise ermöglicht einen Rückgriff auf statistische Analysen und Quotenvergleiche (also bspw. den Vergleich des Frauenanteils in Führungspositionen mit dem Anteil der im Unternehmen insgesamt beschäftigten Frauen),[6] um Verstöße gegen das Benachteiligungsverbot nachzuweisen. Die Modifikationen der Beweislastregeln für Diskriminierungsfälle und die explizite Erwähnung von *Testing*-Verfahren zum Beweis einer Diskriminierung im Rahmen des Gesetzgebungsverfahrens zum AGG zeigen die Offenheit des Gesetzgebers für die Heranziehung empirischer Methoden bei der Gesetzesanwendung.[7]

Dabei ist die Anwendung empirischer Methoden in juristischen Kontexten nicht neu.[8] Die **Gesetzgebung** etwa macht sich quantitative Methoden bereits im Rahmen der Gesetzesfolgenabschätzung und der Begründung verhaltenslenkender regulatorischer Maßnahmen zunutze; der Schritt hin zu einer Gesetzgebung, die mit Hilfe empirischer Einsichten faktische Diskriminierung bereits im Gesetzgebungsprozess antizipiert und aktiv gegensteuert ist da nicht mehr weit.

Empirische Forschung – und die durch sie eröffnete Möglichkeit eines Blicks auf die Mechanismen von Diskriminierung – kann darüber hinaus ein wichtiges Instrument zu ermessensfehlerfreier Kompetenzausübung in der **Exekutive** sein.[9] Hier sei an die Debatte um *Racial Profiling* durch Polizeibehörden erinnert: Polizeibehörden (und Innenminister) weisen diskriminierendes polizeiliches Handeln mit den Mitteln des *Racial Profiling* weit von sich,[10] während individuelle Erfahrungen von Bürger*innen mit

---

5 *Goerg/Petersen,* in: *Towfigh/Petersen,* Ökonomische Methoden im Recht, Rn. 394.

6 BT-Drs. 16/1780, S. 47; *Bayreuther,* „Quotenbeweis" im Diskriminierungsrecht?, NJW 2009, 806; kritisch hierzu BAG, NZA 2011, 93.

7 Hierzu auch unten B.VIII.

8 *Engel,* Empirical Methods for the Law, Journal for Institutional and Theoretical Economics 174, S. 5 (15).

9 *Christensen,* Statistik vor Gericht – eine empirische Bestandsaufnahme, ASta Wirtschafts- und Sozialstatistisches Archiv (2014) 8, (81), (87); *Ihden,* Die Relevanz statistischer Methoden in der Rechtsprechung und mögliche Implikationen für die juristische Ausbildung, S. 18 (32) ff.

10 *Deutsches Institut für Menschenrechte*, Racial Profiling: Bund und Länder müssen polizeiliche Praxis überprüfen https://www.institut-fuer-

Migrationshintergrund ein anderes Bild zeichnen und an der Einschätzung der Behörden zweifeln lassen. Eine nach allen Regeln der wissenschaftlichen Kunst durchgeführte empirische Studie könnte dieser Debatte versachlichen und zur Befriedung des Diskurses beitragen.

Empirische Forschung kann schließlich der **Rechtsprechung** dienlich sein, wenn diese bei der Rechtsfindung auf statistische Befunde zurückgreifen kann, etwa wenn sie das Verhalten der Streitparteien bewertet oder die Ermessensausübung der Verwaltung kontrolliert.[11] So hat etwa das Oberverwaltungsgericht Münster eine Identitätsfeststellung für rechtswidrig erklärt, bei der die Polizei den Gefahrenverdacht unter anderem mit der „Hautfarbe" des kontrollierten Mannes begründet hatte.[12] Die Behörde behauptete, am Ort der Kontrolle – einem Bahnhof – würden Straftaten überwiegend von Männern „nordafrikanischer" oder „schwarzafrikanischer" Herkunft begangen.

Aber auch die **teleologische Auslegung**, **Verhältnismäßigkeitsprüfungen** im Verfassungsrecht sowie die **Normkonkretisierung** durch Gerichte können von einem Rückgriff auf Empirie profitieren.[13] Die Einschätzung der Schwere einer Grundrechtsbeeinträchtigung in einem konkreten Fall ist in der Regel eine Tatfrage, der Rückgriff auf empirische Einsichten also angebracht.[14] Damit dieser Rückgriff gelingen kann, ist empirische **Methodenkenntnis** unabdingbar, um nicht Gefahr zu laufen, eine juristische Argumentation auf statistische Befunde zu stützen, die die Schlussfolgerung nicht umfassend tragen, wie dies etwa bei der Rauchverbot-Entscheidung des Bundesverfassungsgerichts der Fall war.[15]

---

menschenrechte.de/fileadmin/user_upload/Publikationen/Stellungnahmen/Stellungnahme_Racial_Profiling_Bund_Laender_muessen_polizeil_Praxis_ueberpruefen.pdf (zuletzt abgerufen am 30. August 2020);

11 BAG Urt. V. 21.06.2012 - 8 AZR 364/11, NZA 2012, 1345; OVG Münster, Urteil v. 7. August 2018, Az.: 5 A 294/16, NVwZ 2018, 1497, 1501.

12 OVG Münster, Fn. 11, NVwZ 2018, 1497, 1501.

13 *Petersen,* Braucht die Rechtswissenschaft eine empirische Wende? Der Staat 2010, S. 435 (446); *Gloeckner/Towfigh,* Geschicktes Glücksspiel, JZ 2010, 1027 ff. *Towfigh,* Empirical arguments in public law doctrine: Should empirical legal studies ma a "doctrinal turn"? International Journal of Constitutional Law (I•CON) 2014, S. 670 (676 f.).

14 Vgl. BVerfGE 121, 317, 355 und 364 [2008]; *Petersen,* Fn. 13, S. 435 (449) ff.

15 *Petersen,* Fn. 13, S. 451; s. auch unten B.VII. – Die Vermittlung eines minimalen Grundverständnisses ist auch unentbehrlich, will man völlig absurde Äußerungen von Juristen wie jüngst jene des Trump-Lagers bei der Anfechtung der US-Präsidentschaftswahlen 2020 vermeiden, die einen ganzen Berufsstand in Misskredit zu bringen und das Vertrauen in die Rechtspflege zu stören geeignet sind, vgl.

Freilich wohnt allen empirischen Befunden die **epistemische Unsicherheit** inne, die jedes Wissen relativiert; auch Daten und statistische Analysen stellen keine objektive Wahrheit dar. Dennoch ist **statistische Evidenz**, sofern sie methodenstreng gewonnen wird, im Hinblick auf Diskriminierung zuverlässiger als intuitive Plausibilisierungen aufgrund individueller Erfahrungen (**anekdotische Evidenz**). Das Antidiskriminierungsrecht sollte daher die empirischen Methoden als wichtige Verbündete erkennen, mit deren Hilfe sich die Konturen dieses Rechtsgebiets schärfen und Ungleichbehandlungen methodenstreng nachweisen lassen.

I. Empirie zur Feststellung von Diskriminierung anhand bestimmter Merkmale

Um empirische Argumente führen oder widerlegen und um Diskriminierung mit ihrer Hilfe nachweisen zu können, sind Grundkenntnisse der entsprechenden Methoden und Wachsamkeit im Umgang mit Fallstricken und Hürden für all jene unabdingbar, die sich mit Diskriminierung auseinandersetzen müssen. In diesem Zusammenhang ist es auch für Jurist*innen notwendig, die Eignung empirischer Befunde für die Beantwortung einer Tat- oder Rechtsfrage einschätzen, statistische Ergebnisse richtig „lesen“ und interpretieren sowie methodische Grenzen oder Fehler erkennen zu können; erforderlich sind mithin spezifisch juristisch ausgeprägte empirische Fähigkeiten („**Rechtsempirie**“), die sich in Teilen etwa von der empirischen Sozialforschung unterscheiden. Das gilt nicht nur für Spezialist*innen im Bereich des Antidiskriminierungsrechts: Rechtsprechung, Verwaltung und Gesetzgebung sind immer wieder mit Diskriminierungsfällen oder Diskriminierungsprävention konfrontiert und sollten zumindest in der Lage sein, empirische Befunde kompetent zu rezipieren.[16]

Das Argument, das mithilfe empirischer Methoden geführt werden soll, hat dabei folgende Grundstruktur: Man betrachtet eine zufällig gezogene Stichprobe und erzeugt anhand aller zu diesen Personen vorliegenden Merkmale und Kontrollvariablen gleichsam „statistische Zwillinge“, die sich am Ende aufgrund statistischer Kontrollen nur mehr hinsichtlich „diskriminierender“ Merkmale unterscheiden, so dass man abgesehen von den untersuchten Merkmalen Gleiches mit Gleichem vergleicht. Nun wird untersucht, ob ein statistisch signifikanter Zusammenhang zwischen der zu erklärenden

---

https://www.supremecourt.gov/DocketPDF/22/22O155/163048/20201208132827887_TX-v-State-ExpedMot%202020-12-07%20FINAL.pdf, dort S. 8 (zuletzt abgerufen am 23. Dezember 2020).

16 *Petersen,* Fn. 13, S. 447 f.; *Towfigh,* Fn. 13, S. 672.

(„abhängigen“) Variable (etwa Gehalt) und einer oder mehreren erklärenden („unabhängigen“) Variablen (etwa Geschlecht oder Umsatz) besteht – ob mit anderen Worten die „statistischen Zwillinge“, die sich nur hinsichtlich der unabhängigen Variablen unterscheiden, ungleiche Werte bei der abhängigen Variablen aufweisen. Beispielsweise könnte in einem Unternehmen die Variable „Gehalt“ bei Vorliegen des Merkmals „weiblich“ – *ceteris paribus*, also etwa bei statistischer Kontrolle für die Höhe des erwirtschafteten Umsatzes – signifikant niedriger sein, wenn die Stichprobe anhand der Variablen „Geschlecht“ geteilt wird. Wenn rechtsempirisch ein Unterschied gezeigt werden kann, für den es keine rechtlich zulässige **Rechtfertigung** gibt, dann ist eine Diskriminierung nachgewiesen.[17]

Ferner erlauben sog. **Interaktionseffekte**, auch intersektionale Diskriminierung sichtbar zu machen:[18] Von Interaktionseffekten spricht man, wenn das gemeinsame Auftreten zweier oder mehrerer erklärender Variablen einen Effekt auf die zu erklärende Variable hat oder die Kombination die Art und Größe des Effekts verändert. Das ist freilich begrifflich zunächst etwas anderes als die wichtige und im Antidiskriminierungsrecht vieldiskutierte Intersektionalität (etwa eine synergistische Wirkung zweier Diskriminierungsmerkmale);[19] gleichwohl können statistische Interaktionseffekte ein sehr wirkungsvolles Instrument sein, um intersektionale Diskriminierung nachzuweisen. Um beim Beispiel des Zusammenhangs zwischen Gehalt und Geschlecht/Umsatz zu bleiben: Das gemeinsame Auftreten der erklärenden Variablen Geschlecht und Alter könnte einen Effekt im Hinblick auf die Höhe des Gehalts (als zu erklärende Variable) haben, der bei einer Betrachtung nur des Geschlechts oder nur des Alters der Mitarbeiter*innen nicht auftritt.

Die Diskussion darum, ob ein Diskriminierungs-Befund tatsächlich als nachgewiesen gelten kann, hat in der Regel zwei Stoßrichtungen: Einerseits kann hinterfragt werden, ob das **Studiendesign** die Regeln der Kunst beachtet. Andererseits können im Studiendesign nicht abgebildete „**Alternativrechtfertigungen**“ für die beobachtete Ungleichbehandlung plausibilisiert werden; dann ist der Nachweis der Diskriminierung nicht vollständig erbracht. Dabei verbleibt ein gewisses Risiko, dass trotz aller

---

17 S.u. B.VI.3.b).

18 Zu intersektionaler Diskriminierung → *Holzleithner*, § 13; s. unten B.VI.3.b) zu Interaktionseffekten.

19 → *Holzleithner*, § 13.

Bemühungen, dem idealen Studiendesign nahezukommen, nicht ganz ausgeschlossen werden kann, dass beobachtete Effekte (auch) auf eine nicht berücksichtigte Variable zurückzuführen sind (sog. *omitted variable)*, die rechtspolitisch weniger problematisch ist.[20]

II. Empirie zur Feststellung bestimmter Merkmale zur Diskriminierung

Wenn Empirie also ein Werkzeug ist, das genutzt werden kann, um Diskriminierung sichtbar zu machen oder nachzuweisen, so hat diese Macht doch auch eine dunkle Seite: Statistische Methoden lassen sich gleichsam in umgekehrter Richtung auch dazu nutzen, Merkmale zu identifizieren, anhand derer unterschieden oder diskriminiert werden „soll". Das liegt etwa im Marketing auf der Hand: Unternehmen versuchen jene Merkmale zu identifizieren, die mit besonders hoher Wahrscheinlichkeit auf ein Kaufinteresse der Merkmalsträger*innen schließen lassen: Im Kosmetik-Geschäft ist empirisch etabliert, dass Frauen für Hygieneprodukte eine deutlich höhere Zahlungsbereitschaft aufweisen als Männer. Und wenn die Demokratische Partei in den USA feststellt, dass unentschlossene Wähler*innen, die in einem 20-Meilen-Radius um eine Filiale der Bio-Supermarkt-Kette *WholeFoods* leben, sich besonders häufig für die Wahl demokratischer Präsidentschaftskandidat*innen gewinnen ließen, so kann auch dieses Wissen für den gezielten Einsatz von auf dieses Milieu zugeschnittene Wahlwerbung genutzt werden.[21]

Vor diesem Hintergrund ist wichtig zu sehen, dass empirische Methoden auch dazu beitragen können, Diskriminierung zu ermöglichen oder zu perpetuieren. Dies kann auf zweierlei Weise geschehen.

1. Dilemma der Differenz

Erstens tritt bei empirischer Forschung zu Diskriminierung ein Dilemma zu Tage, das in der feministischen Rechtswissenschaft und im Antidiskriminierungsrecht als „**Dilemma der Differenz**" bekannt ist:[22] Zur Untersuchung und Feststellung von

---

20 Das hängt damit zusammen, dass demographische Merkmale nicht zufällig „zugeteilt" werden können, was für ein über jeden Zweifel erhabenes Studiendesign erforderlich wäre.

21 Vgl. https://www.nytimes.com/interactive/2020/02/27/upshot/democrats-may-need-to-break-out-of-the-whole-foods-bubble.html mit weiteren Zahlen zu zahlreichen zusätzlichen Marken-Stores (zuletzt abgerufen am 5. Januar 2021).

22 *Holzleithner,* Rechtskritik der Geschlechterverhältnisse, KJ 2008, 250, 252, *Lembke/Liebscher,* Postkategoriales Antidiskriminierungsrecht? – Oder: Wie kommen Konzepte der Intersektionalität

Diskriminierungen ist gerade eine Differenzierung anhand jener Merkmale erforderlich, für die doch eine Differenzierung ausgeschlossen werden soll. Eine Diskriminierung aufgrund des Geschlechts lässt sich nicht feststellen, wenn die betrachteten Fälle nicht nach Geschlecht unterschieden werden; damit wird aber die Vorstellung, die Geschlechter unterschieden sich, gerade perpetuiert, die Differenz akzentuiert. Nicht zufällig sind in der Statistik insofern die Begriffe „**diskriminierende Variable**“ und „**statistische Diskriminierung**“ als *termini technici* gebräuchlich.[23]

Im Antidiskriminierungsrecht sind die Merkmale, hinsichtlich derer eine Anknüpfung unterbunden werden soll, in der Regel die in § 1 AGG genannten: Rasse oder ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter und sexuelle Identität. Schon dieser Katalog ist Gegenstand zahlreicher Diskussionen und wirft mehr Fragen auf, als er beantwortet; das gilt umso mehr für die **Operationalisierung** dieser Merkmale, also die Übersetzung in empirisch erfassbare Konzepte und messbare Variablen,[24] die gleichwohl unabdingbar ist, weil empirische Forschung der Einordnung von beobachtbaren sozialen Phänomenen in Kategorien und Gruppen bedarf. Das Dilemma besteht dabei darin, dass schon die Auswahl der erhobenen Daten (insbesondere persönlicher Merkmale) unter Umständen zur Entstehung und Festigung sozialer Konstrukte beitragen kann.[25] Außerdem birgt die Fokussierung auf bestimmte Merkmale die Gefahr, essentialistischem Denken Vorschub zu leisten, also der Auffassung, dass Menschen sich anhand „natürlicher“ Merkmale unterscheiden und unterscheiden lassen, und dass eine auf diesen Merkmalen beruhende Ungleichbehandlungen stets gerechtfertigt sei.[26] Dies kann ferner dazu führen, dass der Blick

---

in die Rechtsdogmatik?, in: Philipp/Meier/Apostolovski/Starl/Schmidlechner (Hrsg.), Intersektionelle Benachteiligung und Diskriminierung, 2014, S. 261 ff.; → *Baer*, § 5.

23 Während „Diskriminierung“ allgemeinsprachlich einen moralischen Vorwurf der Intentionalität beinhaltet, beschreibt etwa der Begriff „diskriminierende Variable“ als *terminus technicus* in der empirischen Forschung lediglich nachweisbare Unterschiede ohne eine darüberhinausgehende Wertung. Vgl. *Schauer,* Statistical (and non-statistical) discrimination, in: Lippert-Rasmussen (Hrsg.), Routledge Handbook of the Ethics of Discrimination, 2018, S. 42, 43.

24 S. §§ 5-13 Diskriminierungskategorien; § 7 Rasse und ethnische Herkunft als Diskriminierungskategorien; *Feldmann/Hoffmann/Keilhauer/Liebold,* „Rasse“ und „ethnische Herkunft“ als Merkmale des AGG, RW 2018, 23 ff; *Beigang/Fetz/Kalkum/Otto* in: Antidiskriminierungsstelle des Bundes (Hrsg.), Diskriminierungserfahrungen in Deutschland. Ergebnisse einer Repräsentativ- und einer Betroffenenbefragung, S. 15.

25 *Supik,* Statistik und Diskriminierung, in: Scherr/El-Mafaalani/Yüksel (Hrsg.), Handbuch Diskriminierung, S. 196 f.; *Lembke/Liebscher,* Fn. 22, S. 261 ff; *Liebscher/Naguib/Plümecke/Remus,* Wege aus der Essentialismusfalle: Überlegungen zu einem postkategorialen Antidiskriminierungsrecht, KJ 2012, 204, 206; *Beigang/Fetz/Kalkum/Otto*, Fn. 24, S. 15 ff.

26 → *Baer*, § 5 D.III.

von Diskriminierung und diskriminierenden Strukturen weggelenkt wird, sich auf die Diskriminierten und ihre Eigenschaften und Merkmale fokussiert und (allein) dort Grund und Ursache der Diskriminierung sucht.

Ebenso problematisch ist die Tatsache, dass die Herausbildung der Merkmale unter Umständen **Fremdzuschreibungen** enthält, die nicht nur für die mit dem Merkmal beschriebene Person unpassend sind, sondern unter Umständen auch eine **Stigmatisierung** beinhalten bzw. ihrerseits auf (z.B. rassistischen) **Stereotypen** beruhen können.[27] Durch die vom methodischen Erfordernis abgrenzbarer und eindeutiger Kategorien getriebene Operationalisierung menschlicher Merkmale wird die Vorstellung einer klaren Trennung von und einer eindeutigen Zuordnung zu – z.B. „ethnischen" – Gruppen geweckt und bestärkt, die jedenfalls in dieser Eindeutigkeit nicht besteht.[28]

So ist häufig die Diversität *innerhalb* einer Gruppe größer als *zwischen* den Gruppen, wie etwa in der Genetik beobachtet werden kann: Die meisten genetischen Unterschiede in der DNA-Sequenz des Menschen finden sich *innerhalb* einer geografischen Population. Die genetischen Unterschiede *zwischen* geografischen Populationen machen demgegenüber nur einen sehr geringen Teil aus.[29] Gleichzeitig verschleiert die Kategorisierung regelmäßig intersektionale bzw. mehrdimensionale Diskriminierung, da sie eben nicht alle Dimensionen erfassen kann.[30]

Auflösen lässt sich dieser Konflikt in letzter Konsequenz nicht, empirische Forschung ist jedenfalls mit den heute bekannten Methoden ohne Kategorienbildung nicht möglich und muss sie daher in Kauf nehmen. Das Bewusstsein für diesen Konflikt vermag aber dazu beizutragen, bei der Auswahl der Kategorien bzw. bei der Zuschreibung von Merkmalen umsichtig vorzugehen, Essentialisierung zu vermeiden, Selbstzuschreibungen und Selbstkategorisierungen mit einzubeziehen und der Versuchung unbotmäßiger und zu Unterkomplexität führender Vereinfachung zu widerstehen.

---

27 *Nett/Durrough/Jekel/Glöckner,* Perceived biological and social characteristics of a representative set of German first names, Social Psychology, 51, S. 17.

28 *Supik,* Fn. 25, S. 201.

29 *Kattmann,* Reflections on "race" and science and society in Germany, Journal of Anthropological Sciences, Vol. 95 (2017), 1, 6; *Serre/Pääbo,* Evidence for Gradients of Human Genetic Diversity Within and Among Continents, Genome Res. 2004 (14), 1679, 1683.

30 → *Holzleithner*, § 13.

Gelingt dies, so bietet empirische Forschung die Chance, auch intersektionale bzw. mehrdimensionale Diskriminierung zu erfassen und sichtbar zu machen.[31]

2. Diskriminierung durch Statistik

Zweitens wird mit dem Begriff „**statistischer Diskriminierung**" – neben der Diskriminierung im statistischen Sinne, also der Differenzierung nach unterschiedlichen Merkmalen – etwas missverständlich und unpräzise häufig auch ein anderes Phänomen beschrieben, nämlich die **Diskriminierung *durch* Statistik**.

Diskriminierung durch Statistik ist das Resultat einer aus Statistiken gewonnenen Überzeugung bzw. einer Zuschreibung von Merkmalen, die auf (tatsächlichen oder angenommenen) Durchschnittswerten einer Gruppe beruhen. Der Rückzug auf diese durchschnittlichen Gruppenmerkmale soll über Unsicherheiten bezüglich der Merkmalsausprägungen einer einzelnen Person hinweghelfen. Nach diesem Prinzip funktioniert beispielsweise die Schufa, wenn die Kreditwürdigkeit zum Teil unabhängig von der individuellen wirtschaftlichen Lage unter anderem vom Wohnort abhängig gemacht wird oder wenn jungen Mitarbeiterinnen ein höheres Ausfallrisiko aufgrund zu erwartender Schwangerschaften attribuiert wird. Hier wird jeweils ein Merkmal – Wohnort bzw. Geschlecht – als Indikator für etwas völlig anderes – Kreditwürdigkeit bzw. mögliche Ausfallzeiten der Mitarbeiterin – genutzt.[32]

Ein weiteres Modell zur Erklärung von Diskriminierung ist die Theorie der *taste-based-discrimination.*[33] Sie erklärt Diskriminierung auf dem Arbeitsmarkt mit der Neigung von Arbeitgeber*innen, personelle Auswahlentscheidungen an ihren Vorurteilen und Abneigungen orientiert zu treffen – und in der Folge solche Arbeitnehmer*innen nicht auszuwählen, die einer ihnen fernstehenden Minderheit angehören, mit denen sie nicht interagieren möchten, selbst wenn von ihnen höhere Produktivität zu erwarten wäre. Hierdurch verfestigt sich der Eindruck, bestimmte Gruppen seien produktiver als andere; das Resultat ist auch hier Diskriminierung durch Statistik.[34]

---

31 *Lembke/Liebscher,* Fn. 22, S. 261, 269; s.u. B.VI.3.b).

32 *Arrow,* What has Economics to Say about Racial Discrimination?, The Journal of Economic Perspectives, Vol. 12, No. 2, S. 91, 96; *Schauer,* Fn. 23, S. 48.

33 *Becker,* The Economics of Discrimination.

34 *Neilson/Ying,* From taste-based to statistical discrimination, Journal of Economic Behavior & Organization 129 (2016) S. 116.

Daten hierfür stehen inzwischen dank „**Big Data**“ in großer Zahl zur Verfügung und können nun über die klassischen konsumorientierten Bereiche hinaus herangezogen und ausgewertet werden.[35] Sie eröffnen damit auch ein neues Feld potentieller Diskriminierung durch Statistik: Die Diskriminierung durch Algorithmen.[36] Sehr große und komplexe Datensätze („Big Data“[37]) werden heute typischerweise durch mit statistischen Methoden operierende Algorithmen ausgewertet, die oftmals ihre statistischen Vorhersagen anhand der aus den ausgewerteten Daten gewonnenen Einsichten immer weiter verfeinern (was dann gern als „künstliche Intelligenz“ bezeichnet wird).[38] Solche Algorithmen sind damit aber abhängig von den ihnen zugrunde liegenden Daten und keineswegs neutral;[39] die Zuverlässigkeit ihrer Ergebnisse ist abhängig von der Qualität der Daten, mit denen sie gespeist werden.

Werden beispielsweise Personalauswahlentscheidungen (oder eine Vorselektion) an Algorithmen delegiert, so werden diese auf der Grundlage der dem Algorithmus zur Verfügung stehenden Daten getroffen. Ein auf den Einstellungen und Ablehnungen der letzten zehn Jahre gründendes *Recruiting Tool* könnte z.B. die Lebensläufe von Bewerber*innen mit denen der bisher eingestellten Mitarbeiter*innen vergleichen und sie entsprechend bewerten. Sind in der Vergangenheit überwiegend Männer eingestellt worden, scheitern Bewerbungen von Frauen an den vom Algorithmus identifizierten Auswahlkriterien – selbst bei tatsächlich bestehender Qualifikation für die Stelle würden sie nicht als geeignet markiert. Der Algorithmus orientiert sich ausschließlich an in der Vergangenheit gültigen Kriterien und Einschätzungen, korrelierte hier das männliche Geschlecht mit Leistungsfähigkeit (und sei es aufgrund früherer

---

35 Dazu unten B.IV.3.

36 Dazu → *von Ungern-Sternberg*, § 30.

37 Mit dem Attribut „Big Data“ werden Datensätze versehen, die nicht manuell oder mit Mitteln der herkömmlichen Datenverarbeitung analysiert werden können. Dabei wird auf fünf „v“ abgestellt: *volume* (Datenmenge), *velocity* (Geschwindigkeit der Datengenerierung), *variety* (Vielfalt der Datentypen und der Datenquellen), *veracity* (Echtheit), *validity* (Qualität). Weil die Möglichkeiten der Datenverarbeitung sich stetig entwickeln, handelt es sich bei Big Data um einen dynamischen Begriff.

38 *Knorre,* Big Data im öffentlichen Diskurs, in: Hindernisse und Lösungsangebote für eine Verständigung über den Umgang mit Massendaten, in: Knorre/Müller-Peters/Wagner, Die Big-Data-Debatte, S. 1 (6 f.).

39 *Fröhlich/Spiecker genannt Döhmann,* Können Algorithmen diskriminieren?, VerfBlog 2018/12/26, https://verfassungsblog.de/koennen-algorithmen-diskriminieren/, DOI: 10.17176/20190211-224048-0 (zuletzt abgerufen am 30. April 2020); *Altenburger/Ho,* When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions, JITE 175, S. 98 ff..

Diskriminierung oder struktureller Benachteiligung), so ist der Algorithmus nicht in der Lage, tatsächliche Veränderungen bei der Auswahlentscheidung zu berücksichtigen.[40] Das Problem der Zuschreibung von Gruppenzugehörigkeit kann sich also durch den Einsatz von Algorithmen verstärken: Ein Algorithmus, der Frauen mit Betreuungspflicht pauschal als auf dem Arbeitsmarkt schwer vermittelbar qualifiziert, suggeriert zum einen, dass Frauen mit Betreuungspflicht eine homogene Gruppe darstellten und begründet zum anderen mit der Zugehörigkeit zu dieser Gruppe pauschal eine Benachteiligung.[41]

Die tatsächliche oder nur zugeschriebene Zugehörigkeit zu einer statistischen Gruppe führt also zu Schlüssen über die Eigenschaften oder das Verhalten der Einzelnen. Auf dieser Logik beruht auch (diskriminierendes) ***racial profiling***, welches dazu führt, dass verdachtsunabhängige Kontrollen der Polizei häufiger Menschen treffen, die aufgrund ihres „Phänotyps" – etwa ihrer Haut- oder Haarfarbe – oder wegen eines sichtbar getragenen religiösen Symbols die Aufmerksamkeit der Polizei erregen.[42] Die Diskriminierung wird mit empirischen Befunden gerechtfertigt, die beim *racial profiling* zum Ausgangspunkt polizeilichen Handelns werden.[43]

Eine andere Form der Diskriminierung durch Statistik zeigt sich in Rückkopplungseffekten, die dadurch entstehen können, dass eine Person, die sich einer diskriminierten Gruppe zugehörig fühlt, aus dieser Gruppenzugehörigkeit Schlüsse über die eigenen Fähigkeiten oder Chancen zieht („**Andorra-Effekt**"). Das statistische Wissen kann damit zum Hindernis werden und das Verhalten und/oder die Motivation dieser Person im Sinne einer ***self-fulfilling prophecy***[44] beeinflussen: Die Jurastudentin, die der

---

40 *Kleinberg/Ludwig/Mullainathan/Sunstein,* Discrimination in The Age of Algorithms, Journal of Legal Analysis, Vol. 10 (2018) S. 113, 162.

41 *Fröhlich/Spiecker genannt Döhmann,* Fn. 39, S. 1 ff. – Hier steckt natürlich ein normatives Problem: Der Verwender*in des Algorithmus – bspw. Arbeitgeber*in – mag es vor allem darum gehen, falsch-negative Entscheidungen zu vermeiden, so dass sie solche Diskriminierungen in Kauf zu nehmen bereit sein könnte. Die Verfügbarkeit (relativ) präziser Vorhersagen macht solche Entscheidungen einfacher. Auch wenn sie diskriminierend und die zugrundeliegende Heuristik erwiesen falsch ist, kann der Einsatz des Algorithmus aus subjektiver Sicht konsistent sein.

42 *Liebscher,* „Racial Profiling" im Lichte des verfassungsrechtlichen Diskriminierungsverbots, NJW 2016, 2779; → *Tischbirek*, § 27.

43 Das ist jedenfalls in all jenen Fällen ohne Zweifel rechtswidrig, in denen die Empirie, auf die sich eine polizeiliche Auswahlentscheidung stützt, den Schluss objektiv nicht zulässt: OVG Münster, Urteil v. 7. August 2018, Az.: 5 A 294/16; s. auch *Behr,* Diskriminierung durch Polizeibehörden, in: Scherr/El-Mafaalani/Yüksel (Hrsg.), Fn. 25, S. 301 ff.

44 *Merton,* Die self-fulfilling prophecy, in: *Merton,* Soziologische Theorie und soziale Struktur, S. 399; *Word/Zanna/Cooper,* The nonverbal mediation of self-fulfilling prophecies in interracial interaction,

Presse entnommen hat, dass Frauen in juristischen Staatsprüfungen schlechter abschneiden als Männer, kann die Diskriminierung antizipieren, ist möglicherweise in der eigenen Prüfung hierdurch verunsichert und daher weniger leistungsstark (***stereotype threat***[45]). Andersherum kann das Wissen um statistische Befunde auch entlastende Wirkung haben, wenn etwa die gleiche Jurastudentin ihr schwächeres Abschneiden in der Prüfung nicht ihren eigenen Schwächen zuschreibt, sondern einer sie benachteiligenden Prüfungssituation.

B. Grundlagen empirischer Methoden

I. Forschungsfrage

Empirische Sozial- und Rechtsforschung kann einerseits als deskriptive Forschung soziale Zustände beschreiben – beispielsweise ermitteln, dass das durchschnittliche Einkommen von Frauen etwa 21% geringer ist als das durchschnittliche Einkommen von Männern[46]. Andererseits versucht sie, **Korrelationen** und **Kausalzusammenhänge** zu identifizieren, die diesen Befunden zu Grunde liegen, letztlich im Bestreben, den sozialen **Mechanismus** zu identifizieren, der die Befunde gleichsam „erzeugt" – also die Frage zu beantworten, *warum* das durchschnittliche Einkommen von Frauen niedriger ist als das von Männern. Hierzu bedienen sich (Rechts-)Empiriker*innen der Methoden der Statistik.[47]

Korrelationsuntersuchungen sind dabei ein wichtiges Werkzeug, mit dessen Hilfe eine Aussage darüber getroffen werden kann, ob Variablen in einem Zusammenhang stehen und wie stark dieser Zusammenhang ist.[48] Dass zwei Variablen korreliert sind, bedeutet jedoch nicht, dass zwischen ihnen auch ein Kausalzusammenhang besteht. [49] Der unmittelbare Schluss von einer Korrelation auf einen

---

Journal of Experimental Social Psychology, 10 (1974), 109-120; *Supik,* Fn. 25, S. 204; *Fröhlich/Spiecker genannt Döhmann,* Fn. 39, S. 2.

45 *Spencer/Steele/Quinn*, Stereotype Threat and Women's Math Performance, Journal of Experimental Social Psychology 35 (1999) 4-28; *Lusher/Campbell/Carrell,* TAs like me: Racial interactions between graduate teaching assistants and undergraduates, Journal of Public Economics 159 (2018) 203-242.

46 https://www.destatis.de/DE/Themen/Arbeit/Arbeitsmarkt/Qualitaet-Arbeit/Dimension-1/gender-pay-gap.html (zuletzt abgerufen am 30. August 2020).

47 *Goerg/Petersen,* Fn. 5, Rn. 394; *Handl/Kuhlenkasper,* Einführung in die Statistik, S. 3 f.

48 Ausführlich dazu unten, B.VI.3.a).

49 *Kühnel/Dingelstedt,* Kausalität, in: Baur/Blasius (Hrsg.), Handbuch Methoden der empirischen Sozialforschung, S. 1401 ff.

Kausalzusammenhang ist nur ausnahmsweise möglich, etwa wenn die Kausalität nur in eine Richtung verlaufen kann: So kann die Körpergröße der Eltern nicht von jener ihrer Kinder beeinflusst sein (gerichtete Korrelation: Kausalverlauf Eltern → Kind); und die Feststellung, dass mehr Männer als Frauen eine Führungsposition bekleiden, erlaubt die Hypothese, dass Männer leichter Führungspositionen erreichen, nicht aber die umgekehrte Schlussfolgerung, dass zu einem Mann wird, wer eine Führungsposition erlangt hat (gerichtete Korrelation: Kausalverlauf Mann → Führungsposition).[50]

Verschiedene Faktoren können aber auch miteinander korrelieren, ohne dass eine Kausalität besteht. Dann handelt es sich entweder um eine zufällige Korrelation (**Koinzidenz**), oder es liegt eine sog. **Scheinkorrelation** vor. Im Falle einer Scheinkorrelation besteht zwar statistisch ein signifikanter Zusammenhang zwischen zwei Variablen, Ursache hierfür ist aber ein dritter Faktor, der nicht berücksichtigt wurde – und damit letztlich ein Fehler im empirischen Modell. Das paradigmatische Lehrbuchbeispiel hierfür ist die Korrelation zwischen der Anzahl von Storchenpaaren und der Geburtenrate bei Menschen: Eine Auswertung der Daten mehrerer europäischer Länder zu Geburten und Storchenpopulation ergab eine positive Korrelation zwischen diesen Variablen: Je höher die Storchenpopulation desto höher die Geburtenrate.[51] Bei diesem Beispiel ist auch ohne vertiefte statistische Kenntnisse offenkundig, dass es keinen tatsächlichen Zusammenhang zwischen der Storchenpopulation und der Geburtenrate gibt. Beide Variablen mögen zwar positiv korrelieren, stehen aber nicht in dem Zusammenhang, der insinuiert wird. Der Zusammenhang ergibt sich vielmehr erst durch eine hinzutretende dritte Variable als gemeinsamem Treiber beider Variablen, etwa die Fläche eines Landes oder den Grad der Industrialisierung des Landes.[52] Ein anderes Beispiel ist die Korrelation von Körpergröße und Gehalt: Großgewachsene Männer verdienen mehr als ihre kleineren Geschlechtsgenossen.[53] Allerdings besteht zwischen der Körpergröße und dem Gehalt gerade keine unmittelbare Kausalität, sondern es kommen zusätzliche Variablen ins Spiel, wie etwa die Gene oder der

---

50 *Matthews,* Teaching Statistics, 22(2), 2000, S. 36, 37; *Supik,* Fn. 25, S. 203.

51 *Handl/Kuhlenkasper,* Fn. 47, S. 172.

52 *Matthews,* Fn. 50, S. 37; *Goerg/Petersen,* Fn. 5, Rn. 465.

53 *Spanhel,* Der Einfluss der Körpergröße auf Lohnhöhe und Berufswahl: Aktueller Forschungsstand und neue Ergebnisse auf Basis des Mikrozensus, Wirtschaft und Statistik 2 (2010) S. 170.

Lebensstandard der Eltern;[54] auch ein Zusammenhang mit einem stärkeren Selbstbewusstsein oder weiteren Ursachen lässt sich nicht ausschließen.

Um das Risiko einer Scheinkorrelation zu reduzieren, ist einerseits das Forschungsdesign so zu gestalten, dass mögliche **Alternativmechanismen** (etwa der Grad der Industrialisierung oder die Herkunft eines Arbeitnehmers) durch die Untersuchung bereits ausgeschlossen werden, ferner muss im Rahmen statistischer Testverfahren für mögliche **Störfaktoren** kontrolliert werden.[55]

Sorgfalt beim Forschungsdesign ist auch deshalb erforderlich, da eine empirische Studie nur solche Effekte sichtbar machen kann, die sich aus den in der Studie einbezogenen Faktoren (Variablen) ergeben. Nur eine Studie, die alle relevanten Faktoren berücksichtigt, kann zu fundierten Ergebnissen gelangen. Ist das nicht der Fall, kommt es zu einer Verzerrung der Ergebnisse durch ausgelassene Variablen (***omitted variable bias*).** Eine Statistik, mit der Diskriminierung nachgewiesen werden soll, die aber wichtige Faktoren zur Unterscheidung nicht einbezieht, ist nicht geeignet, Diskriminierungen nachzuweisen. Wird bei einer Erhebung zur Diskriminierung auf dem Wohnungsmarkt nur der Geburtsort der Bewerber*innen um eine Wohnung erhoben, so wird sich höchstwahrscheinlich eine Diskriminierung von Wohnungssuchenden der zweiten Generation von Migrant*innen nicht identifizieren lassen.

II. Theorie, Modell und Hypothesen

1. Die theoretisch-empirische Wissensspirale

Ausgangspunkt empirischer Forschung sollte grundsätzlich eine **Theorie** sein, die den empirisch zu untersuchenden Zusammenhang erklärt.[56] So gibt es etwa in der sozialpsychologischen Vorurteilsforschung eine Reihe von Theorien, die abstrakt Diskriminierung zugrundeliegende soziale Mechanismen zu beschreiben und zu erklären beanspruchen.[57] Eine dieser Theorien ist die *Theorie der sozialen Identität*[58], welche für

---

54 *Case/Paxson* (2008): Stature and Status: Height, Ability, and Labor Market Outcomes, in: Journal of Political Economy, H. 3, 116. Jg., S. 499.

55 *Goerg/Petersen,* Fn. 5, Rn. 406.

56 *Schäfer,* Methodenlehre und Statistik, S. 9.

57 *Zick,* Sozialpsychologische Diskriminierungsforschung, in: Scherr/El-Mafaalani/Yüksel (Hrsg.), Fn. 25, S. 60 ff.; *Charles/Guryan,* Studying Discrimination: Fundamental Challenges and Recent Progress, in: Annual Review of Economics, Vol. 3 (2011), S. 479, 494 ff.

58 *Tajfel/Turner,* The social identity theory of intergroup behavior, in: *Worchel/Austin (Hrsg.),* Psychology of intergroup relations, S. 7 ff.

das Verständnis der kollektiven Dimension von Diskriminierung herangezogen wird, da sie den Einflussfaktor „Gruppe“ in den Blick nimmt.[59] Konkrete Fragestellungen beantwortet eine Theorie nicht. Empirische Forschung, die sich mit Fragen von Diskriminierung aufgrund einer Gruppenzugehörigkeit befasst, wird sich aber der Annahmen dieser Theorie bedienen und den abstrakt beschriebenen Mechanismus in einem Modell operationalisieren, um sie empirisch überprüfen zu können. Eine Theorie schlägt also eine sorgfältig begründete, dennoch vorläufige Antwort auf die gestellte Forschungsfrage – ein Modell – vor und setzt damit den Rahmen, aus dem die **Hypothese** abgeleitet wird, die in der Folge anhand von Daten aus der „Wirklichkeit“ empirisch überprüft werden soll. Das Modell, das sich aus miteinander verbundenen Ideen, Annahmen und Hypothesen über einen Sachverhalt zusammensetzt, begrenzt die Freiheitsgrade bei der Erklärung der Daten, schränkt die Erklärungsansätze ein und führt damit zu Methodenstrenge. Empirische Forschung, die nicht theoriegeleitet ist, verleitet leicht zu ***ad-hoc*- und *ex-post*-Erklärungen**, die aufgrund der Suggestionskraft von Daten eine starke Plausibilität aufweisen können, ohne dass ihnen eine solche methodisch zukommt.[60]

Theorie, Modell, Hypothesen und Empirie bilden gemeinsam gewissermaßen eine **Wissensspirale**: Aus der Theorie wird ein Modell entwickelt und aus diesem werden Hypothesen abgeleitet, die idealerweise in einem steten Kreislauf durch neue, auf ihnen gründende Empirie verbessert werden: Das Modell und die auf diesem gründenden Hypothesen werden empirisch überprüft; und die Ergebnisse der empirischen Überprüfung werden wiederum genutzt, um Theorie, Modell und Hypothesen zu verfeinern. Die solcherart modifizierte Theorie wird dann wiederum mit neuen Daten getestet.

Existiert (noch) keine Theorie, so können wir in diese Wissensspirale indessen auch unmittelbar mit empirischer Forschung einsteigen: Man spricht dann von **explorativer Empirie**. In diesem Fall bilden zunächst (Feld-)Daten den Ausgangpunkt der wissenschaftlichen Betrachtung.[61] Auf der Grundlage ihrer empirischen Analyse wird eine

---

59 Für einen Überblick über die zentralen Theorien aus der Sozialpsychologie s. *Zick*, Fn. 25, S. 67 ff.; *Beigang/Fetz/Kalkum/Otto* in: Antidiskriminierungsstelle des Bundes (Hrsg.) Diskriminierungserfahrungen in Deutschland. Ergebnisse einer Repräsentativ- und einer Betroffenenbefragung, S. 28 ff.

60 *Häder,* Empirische Sozialforschung, S. 55.

61 S.u. B.IV.2.

Theorie mit Modell und Hypothesen entwickelt, die diese Daten erklären könnte. Sodann wird diese Theorie mit neuen Daten getestet. Insbesondere in wenig erforschten Bereichen finden sich explorative Studien, die zur Theorienbildung beitragen sollen.[62] Um die oben erwähnten Probleme einer *ad-hoc*- und *ex-post*-Plausibilisierung zu vermeiden, ist entscheidend, dass nach der Theoriebildung wiederum eine empirische Überprüfung erfolgt. Die oftmals fehlende, weitergehende empirische Überprüfung stellt insbesondere bei explorativen Studien, die sich auf „Big Data" stützen, einen Mangel dar:[63] Häufig ist bei solchen Studien die Darstellung des Problems – anders als beim deduktiven Ansatz der theoriengeleiteten Forschung – nicht Ausgangspunkt sondern Ende des Prozesses. Die Daten werden analysiert, aus der Analyse wird eine zu den Daten passende *ad-hoc*-Theorie entwickelt und abschließend das Problem benannt, auf das die Untersuchung eine Antwort gefunden zu haben vorgibt.[64] Darüberhinausgehende Theorienbildung ist hingegen nicht das Ziel.

2. Frequentistisch-probabilistische Aussagen

Deterministische Aussagen – also solche, die einen universalen kausalen Zusammenhang beschreiben – sind für soziale Mechanismen und damit in der Sozialwissenschaft nur selten möglich, denn wir können in aller Regel nicht den Mechanismus selbst, sondern lediglich Signale beobachten, die Rückschlüsse auf den (angeblich deterministischen) Mechanismus erlauben. Daher werden sozialwissenschaftliche Hypothesen in der Regel frequentistisch-probabilistisch formuliert: Ihre Aussage trifft mit einer gewissen Wahrscheinlichkeit auf eine Vielzahl von Fällen zu.[65] So ist beispielsweise die Hypothese, (alle) Frauen würden in der juristischen Staatsprüfung gegenüber Männern (immer) benachteiligt, in ihrer Universalität nicht haltbar: Schon ein einziger Fall einer Bevorzugung einer Frau genügt, um diese Hypothese zu widerlegen. Daher wird die Hypothese probabilistisch formuliert: Die Wahrscheinlichkeit der

---

62 *Döring/Bortz,* in: Döring/Bortz, Forschungsmethoden und Evaluation, S. 192; s. etwa die qualitative Analyse bei *Alidadi/Foblets,* Framing Multicultural Challenges in Freedom of Religion Terms – Limitations of Minimal Human Rights for Managing Religious Diversity in Europe, Netherlands Quarterly of Human Rights, Vol. 30/4, S. 388 (392).

63 Zu Big Data s.u. B.IV.3.

64 *Mahrt,* Mit Big Data gegen das „Ende der Theorie?", in: Maireder/Ausserhofer/Schumann/Taddicken (Hrsg.), Digitale Methoden in der Kommunikationswissenschaft, S. 23 (24); *Mayerl,* Bedeutet Big Data das Ende der sozialwissenschaftlichen Methodenforschung? https://soziopolis.de/beobachten/wissenschaft/artikel/bedeutet-big-data-das-ende-der-sozialwissenschaftlichen-methodenforschung/ (zuletzt abgerufen am 30. April 2020).

65 *Engel,* Fn. 8, S. 7.

Benachteiligung einer Frau in der juristischen Staatsprüfung ist höher als die Wahrscheinlichkeit der Benachteiligung eines Mannes.[66]

3. Theorien und Modelle in der juristischen Antidiskriminierungsforschung

Das menschliches Verhalten am umfassendsten erklärende Modell ist das des ***homo oeconomicus*** (zutreffender als ***rational actor model*** bezeichnet),[67] das in der rechtsökonomischen Forschung vielfach zu Grunde gelegt wird und das bei all seinen Vorzügen mittlerweile auch hinsichtlich seiner Schwächen und Begrenzungen gut erforscht ist (***behavioral law and economics*** sowie Forschung zu ***biases***).[68] Mit Hilfe des *rational actor models* werden Anreize und Motive für *typisiertes* menschliches Verhalten erforscht und erklärt; es geht davon aus, dass menschliches Verhalten nicht vollkommen zufällig ist, sondern auf einen – im Einzelnen näher zu definierenden – Nutzen ausgerichtet ist. Das Modell kann damit auch dafür eingesetzt werden, um Diskriminierung zu erklären, wenn sie sich als Resultat von Rahmenbedingungen darstellt, die dazu führen, dass es für den Einzelnen „rational" im Sinne des *rational actor models* ist zu diskriminieren. Auch *taste-based discrimination*, kann in diesem Sinne „rational" sein, wenn die Präferenz zu diskriminieren bei den Akteur*innen hinreichend ausgeprägt ist.[69]

Für empirische Forschung zu Diskriminierung sind daneben vor allem psychologische (Verhaltens-)Modelle hilfreich, die auf abgegrenzte Bereiche menschlichen Verhaltens beschränkt sind, damit weniger menschliches Verhalten insgesamt erklären (sollen), sondern vielmehr Erklärungen für spezielle und isolierte Verhaltens*effekte* liefern; solche Befunde finden sich häufig etwa in der sozialpsychologischen Diskriminierungsforschung.[70]

---

66 Vgl. *Towfigh/Traxler/Glöckner,* Geschlechts- und Herkunftseffekte bei der Benotung juristischer Staatsprüfungen, ZDRW 2018, S. 115 ff.

67 *Towfigh,* in: *Towfigh/Petersen,* Ökonomische Methoden im Recht, § 2, Rn. 69 ff.

68 *Englerth/Towfigh,* in: *Towfigh/Petersen,* Ökonomische Methoden im Recht, § 8. Die Begriffe *nicht-rational* und *biases* dürfen dabei nicht in ihrer alltagssprachlichen Bedeutung verstanden werden, vielmehr bezeichnen sie als termini technici der Rechtsökonomik Abweichungen vom Rationalmodell.

69 S.o. A.II.2.

70 *Zick,* Fn. 25, S. 67 ff.

III. Operationalisierung der Forschungsfrage

Die für die Beantwortung der Forschungsfrage erforderlichen Daten können jedenfalls in den Sozialwissenschaften in der Regel nicht unmittelbar in der Welt beobachtet werden; vielmehr sind die Konzepte, anhand derer Modelle erstellt und Hypothesen entwickelt werden, zu **operationalisieren**. Das bedeutet, dass Variablen konstruiert werden müssen, bevor sie gemessen und für die Falsifizierung einer Hypothese und letztlich für die Beantwortung der Forschungsfrage herangezogen werden können. Wenn etwa der Zusammenhang zwischen dem Bestehen eines Antidiskriminierungsrechts in einer Rechtsordnung, einem hohen Maßes an Gleichberechtigung und Demokratie untersucht werden soll, muss zunächst festgelegt werden, was für Regelungen erlassen sein müssen, um das „Bestehen eines Antidiskriminierungsrechts" annehmen zu können, welche Anforderungen an ein „hohes Maß an Gleichberechtigung" zu stellen sind und welche Regeln eine Ordnung aufweisen muss, damit ihre Staatsform als „Demokratie" zu qualifizieren ist.[71]

Ist die Operationalisierung plausibel, misst also der empirische Test das Merkmal, das er zu messen vorgibt, so spricht man von der **Validität des Tests** (nicht zu verwechseln mit der **Validität der Schlussfolgerungen**, die aus diesem Test gezogen werden;[72] insofern ist die etablierte Begrifflichkeit leider missverständlich). Wenn also etwa ein Test zur Ermittlung der Sprachkompetenz von Kindern von Migrant*innen so ausgestaltet ist, dass keine Aussagen über eben diese Sprachkompetenz getroffen werden können, weil die Durchführung des Tests *Kenntnisse* einer Sprache voraussetzt, die nicht Muttersprache ist, so stellt der Test kein zur Beantwortung der Forschungsfrage valides Instrument dar, da mit ihm gerade nicht die Sprach*kompetenz*, sondern Sprach*kenntnisse* gemessen werden.[73] Die Forschungsfrage ist nicht valide operationalisiert. Auf der Grundlage solcher nicht-validen Tests gezogene Konsequenzen etwa zur Förderung der getesteten Kinder sind im besten Fall wirkungslos, im schlimmsten Fall begründen sie eine Jahre anhaltende Benachteiligung aufgrund einer fehlerhaften Einschätzung etwa des Verhältnisses von Sprachkenntnissen und

---

71 Vgl. etwa *Petersen,* Antitrust Law and the Promotion of Democracy and Economic Growth, in: Journal of Competition Law and Economics 9 (2013), S. 593 (604), der die Konzepte „Wettbewerbsrecht", „Wachstum" und „Demokratie" operationalisiert.

72 S.u.B.VII.3.

73 *Hormel,* Diskriminierung von Kindern und Jugendlichen mit Migrationshintergrund im Bildungssystem, in: Hormel/Scherr, Diskriminierung, S. 183.

Intelligenz.[74] Eine Messung der Einstellung zu Ausländern muss sicherstellen, dass auch (nur) diese Haltung und nicht andere, verwandte Einstellungen (etwa Nationalismus) getestet werden; insofern besteht eine Nähe zur internen Validität.[75]

Bei der Operationalisierung der Forschungsfrage bestehen **Freiheitsgrade**, die für die Beurteilung der Validität des Tests bedeutsam sind. So mag man etwa das Konzept „Demokratie" entweder an der Verwirklichung des Mehrheitswillens festmachen (regelmäßige Wahlen, Geltung des Mehrheitsprinzips) oder zusätzlich auch Minderheitenrechte mit in den Blick nehmen (etwa die Durchsetzbarkeit von Grundrechten).[76]

Die Validität eines Tests kann unter anderem dadurch geprüft werden, dass nach Korrelationen[77] zwischen dem zu messenden Merkmal und anderen, aus bereits validierten Tests gewonnenen Merkmalen geschaut wird.[78] Das können zum Beispiel andere (neue) Tests für dieselben Variablen und Konzepte sein, die mit dem ursprünglichen Test korrelieren.[79]

IV. Datenerhebung

Sind die Forschungsfrage und die ihr zugrundeliegenden Konzepte operationalisiert, so können Messungen vorgenommen werden; zu diesem Zwecke werden Daten erhoben und ausgewertet. Empirische Forschung gewinnt ihre Einsichten aus Daten, die idealtypisch entweder als (1) **Experimental-** oder als (2) **Felddaten** gewonnen werden; in jüngerer Zeit ist die Verwendung von (3) **Big Data** zunehmend in den Blick geraten.[80] Methodisch gibt es hier eine große Vielfalt an Forschungsdesigns, die aus einer Kombination aus Feld- und Experimentaldaten bestehen oder zwischen diesen beiden Prototypen angesiedelt sind (beispielsweise Quasi-Experimente oder experimentelle Vignetten-Studien).

---

74 *Center for Intersectional Justice (Hrsg.),* Intersektionalität in Deutschland, S. 32 ff; s. auch LG Köln, Urteil v. 17.7.2018, Az.: 5 O 182/16.

75 S.u. B.VII.3.b)

76 *Petersen,* Fn. 71, S. 593 (605).

77 S.u. B.VI.3.a).

78 *Glöckner/Towfigh,* Fn. 146, S. 708.

79 *Krebs/Menjold,* Fn. 146, S. 497.

80 Zur Datenerhebung in der empirischen Sozialforschung ausführlich *Kromrey,* Empirische Sozialforschung, S. 309 ff; *Schäfer,* Fn. 56, S. 33 ff.

1. Experimentaldaten

Die Erhebung von Experimentaldaten erfolgt in der Regel im Labor: Die Teilnehmer*innen werden in mindestens zwei Gruppen – Kontrollgruppe und Versuchsgruppe(n) – mit Aufgaben oder Entscheidungssituationen konfrontiert, wobei sich das Szenario jeder Versuchsgruppe in einem einzigen Punkt von jenem der Kontrollgruppe unterscheidet: Es wird also die unabhängige Variable variiert, um den Effekt auf die abhängige Variable zu messen.[81] Lässt sich ein Effekt zeigen, so kann dieser *kausal* auf die Manipulation des Faktors in der Versuchsgruppe zurückgeführt werden.

Da Störvariablen – also solche Merkmale einer Person oder der Situation, die möglicherweise die abhängige Variable beeinflussen können – weitgehend ausgeschlossen werden können,[82] haben Experimente den entscheidenden Vorteil, dass (anders als bei der Analyse von Felddaten) ein gerichteter Zusammenhang zwischen Variablen hergestellt werden kann, dass mithin **Kausalzusammenhänge identifiziert** werden können.

2. Felddaten

Für Felddaten wird häufig auf bestehende Datensätze zurückgegriffen, die z.B. vom Statistischen Bundesamt[83] oder dem Forschungsdatenzentrum der Statistischen Landesämter[84], aber auch von Behörden zur Verfügung gestellt werden.[85] Ferner stellen auch wissenschaftliche Institutionen Daten bereit, beispielsweise das Leibniz-Institut für Sozialwissenschaften die Daten der *Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften* (ALLBUS) oder das Deutsche Institut für Wirtschaftsforschung (DIW) das *Sozio-oekonomische Panel* (SOEP).[86]

Sind entsprechende Daten noch nicht verfügbar, so können diese auch eigens für die geplante Untersuchung erhoben werden, etwa durch Fragebögen (auch Online-

---

81 *Schäfer,* Fn. 56, S. 38.

82 *Döring/Bortz*, Fn. 62, S. 206.

83 www.destatis.de (zuletzt abgerufen am 30. September 2019).

84 www.forschungsdatenzentrum.de (zuletzt abgerufen am 30. September 2019).

85 *Hartmann/Lengerer,* Verwaltungsdaten und Daten der amtlichen Statistik, in: Baur/Blasius, *(Hrsg.)*, Fn. 49, S.1223 ff.

86 Zu den Herausforderungen der Erhebung von Diskriminierungsdaten vgl. *Ahyoud/Aikins/Bartsch/Bechert/Gyamerah/Wagner*, Wer nicht gezählt wird, zählt nicht. Antidiskriminierungs- und Gleichstellungsdaten in der Einwanderungsgesellschaft – eine anwendungsorientierte Einführung; *Supik,* Fn. 25, S. 194 f.

Befragungen) oder Interviews. Fragen werden als offene oder geschlossene Fragen formuliert. Offen gestellte Fragen, also solche, die keine Antwortmöglichkeiten vorgeben, sondern eine individuelle Antwort des Befragten fordern,[87] werden im Nachhinein **codiert**, das heißt die Antworten werden zuvor aufgestellten Kategorien zugeordnet. Geschlossene Fragen geben Antwortmöglichkeiten vor, die Antworten sind somit unmittelbar vergleichbar und leichter statistisch auszuwerten. Eine Codierung ist auch dann angezeigt, wenn die Datenerhebung mittels Beobachtung sozialer Tatsachen erfolgt; diese Erhebungsart kann etwa gewählt werden, wenn ein Zusammenhang zwischen der Religionszugehörigkeit bzw. Weltanschauung und dem Abstimmungsverhalten von Parlamentsabgeordneten untersucht werden soll: Die Informationen für eine solche Untersuchung lassen sich aus öffentlich zugänglichen Quellen erheben und müssen dann für die Untersuchung entsprechend codiert werden.

Felddaten können in der Regel nur Korrelationen (also ungerichtete Zusammenhänge) aufzeigen, es sei denn, der Zusammenhang kann nur in eine Richtung verlaufen;[88] auch bei exogenen *random shocks* und bei statistischen Analysen mit Hilfe von Instrumentenvariablen kann die Ökonometrie unter bestimmten Voraussetzungen kausale Zusammenhänge belegen.[89]

3. „Big Data“

Mit dem Schlagwort „**Big Data**“ werden digitale Daten belegt, die aufgrund ihres Umfangs, ihrer Komplexität oder ihrer Struktur weder manuell noch mit herkömmlichen Mitteln der Datenverarbeitung analysiert werden können.[90] Sie werden in der Regel entweder durch die Nutzung digitaler Dienste, in sozialen Netzwerken, durch Transaktionen wie Kreditkartenzahlungen oder im Rahmen automatisierter digitaler Prozesse generiert;[91] es bedarf erheblicher Rechenkapazitäten, um diese Daten nutzbringend auswerten zu können. Big Data ist damit oftmals kommerziellen Ursprungs und daher für wissenschaftliche Zwecke nur eingeschränkt verfügbar; ferner ist die datenschutzkonforme und wissenschaftsethisch einwandfreie Erhebung bisweilen nicht gesichert feststellbar, so dass die Daten für wissenschaftliche Zwecke auch nur

---

87 *Handl/Kuhlenkasper,* Fn. 47, S. 6.

88 s.o. B.I.

89 *Engel,* A random shock is not random assignment, Economics Letters 145 (2016) S. 45.

90 Zur Definition oben A.II.2 und Fn. 37.

91 *Trübner/Mühlichen,* Big Data, in: Baur/Blasius (Hrsg.), Fn. 49, S. 134.

beschränkt nutzbar sind.[92] Gleichwohl ist in diesem Bereich künftig eine ebenso erhebliche wie rasante Entwicklung zu erwarten, weil auch staatliche Stellen (etwa Finanzbehörden) Zugriff auf große eigene Datensätze haben, die sie zur effektiveren und effizienteren Erfüllung ihrer Aufgaben – u.a. datenschutz- und antidiskriminierungsrechtskonform – einsetzen.

Big Data wird in der Regel durch **Algorithmen** ausgewertet. Diese algorithmenbasierte Analyse von Big Data ermöglicht eine schnelle Mustererkennung und erlaubt, die Daten insbesondere für kommerzielle Anwendungen gewinnbringend zu nutzen.[93] Vor allem lassen sich auch kleine Effekte aufgrund der Datenmenge oft mit großer Zuverlässigkeit nachweisen. Big-Data-Datensätze unterscheiden sich von in Experimenten gewonnenen Daten und klassischen Felddaten vor allem darin, dass sie in der Regel nicht für einen bestimmten Forschungsgegenstand oder eine konkrete Fragestellung erhoben wurden und daher nicht theoriegeleitet, sondern explorativ genutzt werden.[94] Häufig sind die für Kontrollvariablen erforderlichen Informationen nicht verfügbar, die bei der Generierung von Felddatensätzen in der Regel explizit mit erhoben werden.[95] Vielfach ist auch die Stichprobe, die sich in den Daten niederschlägt, nicht zufällig gewählt, sondern folgt gewissen Mustern (***selection bias***), so dass die Repräsentativität der Daten für eine bestimmte Population nicht gesichert ist.[96] Auf Big Data fußende Forschungsdesigns sind daher häufig besonders fehleranfällig. Damit erklärt sich, dass der Einsatz von Big Data bei kommerziellen Anwendungen – die keine besondere Methodenstrenge verlangen – sehr viel verbreiteter ist als in der Wissenschaft und bei der Beantwortung sozialer und politischer Fragestellungen. Das Potential von Big Data bei Rechtsstreitigkeiten um Diskriminierung sollte aber dennoch nicht unterschätzt werden: Lässt man statistische Beweismittel in solchen Prozessen zu, so

---

92 *Trübner/Mühlichen,* Fn. 91, S. 134 (153); *Friedrichs,* Forschungsethik, in: Baur/Blasius, Fn. 49, S. 67 ff.

93 *Richter,* Big Data, Statistik und die Datenschutzgrundverordnung, DuD 2016, 581.

94 *Mahrt,* Mit Big Data gegen das "Ende der Theorie"? In: Maireder/Ausserhofer/Schumann/Taddicken (Hrsg.), Digitale Methoden in der Kommunikationswissenschaft S. 23 (24); *Trübner/Mühlichen,* Fn. 91, S. 134 (145); S.o. A.II.2.

95 *Mahrt,* Fn. 94, S. 23 (24).

96 Dazu ausf. unten B.VI.1.

könnte bereits ein mit ihrer Hilfe hinreichend plausibilisierter Verdacht einer Diskriminierung für eine Verurteilung ausreichen.[97]

V. Deskriptive Statistik

Die **deskriptive Statistik** dient der Darstellung von Daten – etwa in Tabellen, Graphiken oder Kennzahlen.[98] Mit ihrer Hilfe werden Daten beschrieben, allerdings keine Schlussfolgerungen über Zusammenhänge gezogen.[99] Es werden vielmehr jene Daten dargestellt, die auch tatsächlich erhoben wurden. Die Darstellung der Daten erlaubt in der Regel keine Verallgemeinerungen, sondern bildet lediglich die Verteilung der getesteten Merkmale auf die Merkmalsträger*innen ab.[100] Damit ermöglicht deskriptive Statistik beispielsweise Aussagen über Anteile und Häufigkeiten von Merkmalen.[101]

Für die Rezeption und das Verständnis deskriptiver Statistik ist es wichtig, sich sorgfältig damit auseinanderzusetzen, was in einer Statistik abgebildet wird und was gerade nicht. Wie werden die Variablen konstruiert, welche Definitionen liegen den Merkmalen zu Grunde? Als Beispiel mag hier das Merkmal „Armut“ und die Definition der Armutsgrenze taugen: Für ein vollständiges Verständnis einer sich auf dieses Merkmal beziehenden Statistik ist es wichtig zu wissen, wie diese definiert wird, welche Parameter ausschlaggebend sind.[102]

Häufig werden aus deskriptiven Statistiken – etwa der **polizeilichen Kriminalstatistik** – vermeintliche Einsichten (oder häufiger noch: „Meinungen“) abgeleitet, obwohl die Daten oftmals entweder gar nicht oder jedenfalls nicht so erhoben wurden, dass sie die jeweilige Aussage zu stützen geeignet wären. Dies gilt beispielsweise für die Frage, ob Menschen ohne deutsche Staatsangehörigkeit in Deutschland häufiger straffällig werden als Menschen mit deutscher Staatsangehörigkeit: Bei Betrachtung einer entsprechenden deskriptiven Statistik ist zum einen darauf zu achten, ob die

---

97 *Pundik,* Rethinking the use of statistical evidence to prove causation in criminal cases: A Tale of (im)probability and free will, Law and Philosophy 2020, https://doi.org/10.1007/s10982-020-09389-0 (zuletzt abgerufen am 4.1.2021).

98 *Goerg/Petersen,* Fn. 5, Rn. 419; *Handl/Kuhlenkasper,* Fn. 47, S. 3; *Schäfer,* Fn. 56, S. 47 ff.

99 *Schäfer*, Fn. 56, S. 47.

100 *Kromrey,* Empirische Sozialforschung, S. 408.

101 *Schäfer*, Fn. 56, S. 48.

102 https://www.statistikportal.de/sites/default/files/2020-01/Definition%20Median%20und%20Armutsgefährdungsschwelle_0.pdf (zuletzt abgerufen am 23. Juni 2020).

Darstellung die Tatsache berücksichtigt, dass beispielsweise Straftaten nach dem Aufenthaltsgesetz von Deutschen überhaupt nicht begangen werden können;[103] zum anderen sollte diese Tatsache ggf. bei der Bildung adäquater Vergleichsgruppen berücksichtigt werden. Auch gibt die polizeiliche Kriminalstatistik – weil sie eben von der Polizei und nicht von den Gerichten erstellt wird – den Erkenntnisstand bei Abschluss der polizeilichen Ermittlungen und im Moment der Abgabe des Falles an die Staatsanwaltschaft wieder. Sie verzeichnet daher lediglich Tatverdächtige, trifft indessen keine Aussagen darüber, ob es später auch zu einer Verurteilung gekommen ist. Sie kann darüber hinaus nur solche Fälle abbilden, die zur Anzeige gebracht wurden: Das Dunkelfeld der nicht zur Anzeige gebrachten Delikte lässt sich nur grob schätzen, ebenso bleiben Unterschiede in der Anzeigebereitschaft und im Anzeigeverhalten unberücksichtigt.[104] Die Daten zu Delikten mit einer typischerweise geringen Anzeigebereitschaft (etwa Internetdelikte wie Phishing, Pharming oder Schadsoftware[105] oder bei persönlicher Beziehung zum Täter[106]) werden in derlei Datensätzen also ebenfalls nur unvollständig verzeichnet.

Bisweilen besteht ein Bedürfnis, auch deskriptive Daten statistisch zu beurteilen. Das ist etwa der Fall, wenn abgeschätzt werden soll, wie stark eine betrachtete Stichprobe hinsichtlich leicht zu beobachtender oder ggf. besonders relevanter Faktoren von der Population abweicht, die sie repräsentieren soll, oder wenn bemessen werden soll, wie ähnlich sich Stichproben sind. Letzteres ist etwa dann von besonderem Interesse, wenn anhand von mit gewissem zeitlichem Abstand wiederholten Erhebungen bestimmt werden soll, ob eine Entwicklung stattgefunden hat. Zu den wichtigsten Kennzahlen, die dies beschreiben, gehören die *Lagemaße* und die *Streuungsmaße*.[107]

1. Lagemaße

Die **Lagemaße** geben die **Mittelwerte** einer Verteilung an und zeigen, auf welchen Wert sich eine Verteilung konzentriert: Dies kann als **Modus**, **Median** oder **arithmetisches Mittel** ausgedrückt werden. Der Modus (auch „Modalwert“) gibt den am

---

103 Die polizeiliche Kriminalstatistik macht inzwischen diesen Unterschied: PKS Jahrbuch 2019, Band 3 Version 2.0, S. 22 f.

104 PKS Jahrbuch 2019, Band 3 Version 2.0, S. 6.

105 Der Deutsche Viktimisierungssurvey 2017, S. 40.

106 *Kölbel*-Mü-Ko-StPO § 158, Rn. 13 m.w.N.

107 *Goerg/Petersen,* Fn. 5, Rn. 428; *Hagen,* Statistik für Juristen, S. 60; *Schäfer,* Fn. 56, S. 52.

häufigsten (typischen) vorkommenden Wert an („Mode“); der Median gibt jenen Wert an, für den gilt, dass sich die übrigen Werte gleichmäßig links und rechts von ihm verteilen (50 % Trennmarke – die Hälfte der Werte liegt links, die andere Hälfte der Werte rechts vom Median) und der arithmetische Mittelwert (auch „Durchschnitt“) beschreibt die Summe aller Einzelwerte geteilt durch die Anzahl der Werte.[108]

Ein Beispiel mag verdeutlichen, wie sich die verschiedenen Lagemaße unterscheiden: In einer Straße stehen zwei Mehrfamilienhäuser, in denen jeweils 30 Parteien leben. In einem Haus leben 30 Parteien mit insgesamt 41 Kindern. Fünf Familien haben je drei Kinder, neun Familien haben je zwei Kinder, acht Familien haben je ein Kind und acht Familien keine Kinder. Der Modalwert dieser Verteilung ist 2, der Median liegt bei 1 und der arithmetische Mittelwert ist 1,13. Im zweiten Haus hat eine Familie 14 Kinder, vier Familien haben drei Kinder, neun Familien haben je zwei Kinder, acht Familien je ein Kind und acht Familien keine Kinder. Der Modalwert liegt auch bei dieser Verteilung bei 2 und der Median ist nach wie vor 1. Allerdings ändert sich das arithmetische Mittel und liegt nun bei 1,73.

An diesem kleinen Beispiel zeigt sich, dass jeder Mittelwert Vor- und Nachteile hat. Das arithmetische Mittel zeigt sich als nicht so robust gegenüber Ausreißern, da alle Werte gleichwertig in die Berechnung eingehen.[109] Der Median hingehen ist robuster gegenüber Ausreißern, der hohe Wert von 14 Kindern verändert ihn nicht.[110] Unterscheiden sich Modalwert und Median, so bedeutet dies, dass die Verteilung nicht symmetrisch ist, in einem solchen Fall ist auch die Aussagekraft des arithmetischen Mittels reduziert.

Die Daten lassen sich in zwei verbreiteten Darstellungsformen der deskriptiven Statistik darstellen – als **Tabelle** und als **Histogramm**:

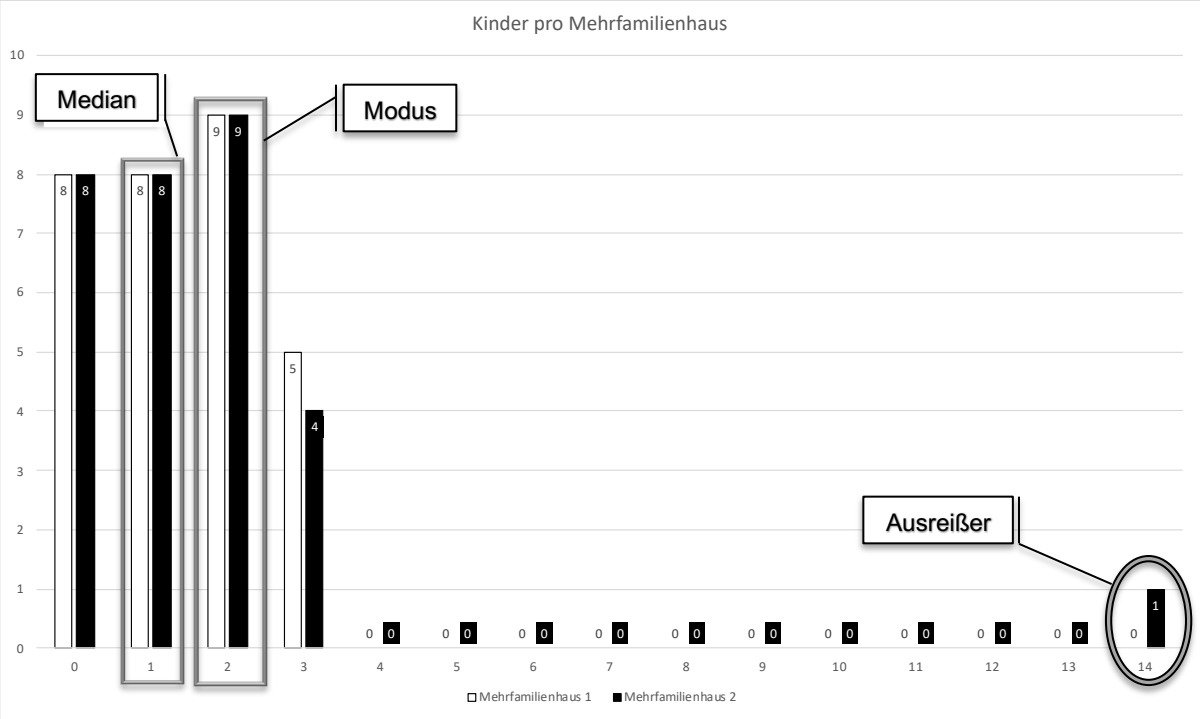
---

108 *Goerg/Petersen,* Fn. 5, Rn. 429 ff.; *Hagen,* Fn. 107, S. 60 ff.; *Schäfer,* Fn. 56, S. 52 ff.; *Kosfeld/Eckert/Türk,* Deskriptive Statistik, S. 68.

109 *Schäfer,* Fn. 56, S. 57.

110 *Goerg/Petersen,* Fn. 5, Rn. 434.

	Mehrfamilienhaus 1	Mehrfamilienhaus 2
kein Kind	8	8
ein Kind	8	8
zwei Kinder	9	9
drei Kinder	5	4
14 Kinder	0	1
Parteien	30	30
Summe Kinder	0 + 8 + 18 + 15 + 0 = 41	0 + 8 + 18 + 12 + 14 = 52
„durchschnittliche“ Zahl Kinder pro Wohnung		
Modus	2 Kinder	2 Kinder
Median	1 Kind	1 Kind
arithmetisches Mittel	1,13 Kinder	1,73 Kinder



2. Streuungsmaße

Die **Streuungsmaße** sind von Bedeutung, da je nach Verteilung der arithmetische Mittelwert zweier graphischer Darstellungen von Häufigkeitsverteilungen (**Histogramme**) sehr ähnlich sein kann, ein Blick auf die Verteilung aber zeigt, dass die Streuungen sehr unterschiedlich sind. Eine Beschränkung auf die Lagemaße ist nicht

sinnvoll, da die Betrachtung des arithmetischen Mittelwerts ohne Berücksichtigung der Streuung wenig aussagekräftig ist.[111]

Zu den wichtigsten Streuungsmaßen gehören die **Stichprobenvarianz** und die **Standardabweichung**. Sie beziehen sich auf den arithmetischen Mittelwert und fragen danach, wie eng bzw. weit die Werte um den Mittelwert verteilt sind, geben also Auskunft darüber, wie zuverlässig ein Mittelwert die Verteilung repräsentiert. Je größer die Streuung der Stichprobe, desto größer ist auch die Stichprobenvarianz.[112]

Die Stichprobenvarianz gibt dabei die durchschnittliche *quadratische* Abweichung der Beobachtungswerte von ihrem Mittelwert an. Zu ihrer Berechnung wird von jedem einzelnen Wert in der Stichprobe der Mittelwert abgezogen und das Ergebnis der Differenz quadriert. Nach Addition dieser Quadrate wird die Summe durch die Anzahl der Beobachtungen (minus 1) dividiert.[113] Das Ergebnis der Berechnung wird in quadrierter Form ausgegeben und ist daher nicht leicht zu interpretieren.[114] Daher wird häufig auch die Standardabweichung angegeben, die mit der Stichprobenvarianz eng zusammenhängt: Sie ist definiert als die Quadratwurzel der Varianz und misst die *durchschnittliche* Abweichung vom arithmetischen Mittel.

Aus dem obigen Beispiel (Kinder in Mehrfamilienhäusern) ergeben sich damit die in der folgenden Tabelle dargestellten Werte, die uns ein konkretes Maß dafür an die Hand geben, was wir intuitiv bereits erfasst haben: dass die Einzelwerte in Mehrfamilienhaus 2 eine deutlich höhere Streuung (Varianz) aufweisen als jene in Mehrfamilienhaus 1.

---

111 *Schäfer,* Fn. 56, S. 60, 66.

112 *Goerg/Petersen,* Fn. 5, Rn. 437.

113 *Goerg/Petersen,* Fn. 5, Rn. 438

114 *Kosfeld/Eckey/Türck,* Fn. 108, S. 120.

	Mehrfamilienhaus 1				Mehrfamilienhaus 2			
arithmetisches Mittel	1,13				1,73			
	Anz.	Diff. (Δ)	Δ^2	Summe	Anz.	Diff. (Δ)	Δ^2	Summe
kein Kind	8	-1,13	1,28	10,24	8	-1,73	2,99	23,92
ein Kind	8	-0,13	0,02	0,16	8	-0,73	0,53	4,24
zwei Kinder	9	0,87	0,76	6,84	9	0,27	0,07	0,63
drei Kinder	5	1,87	3,50	17,5	4	1,27	1,61	6,44
14 Kinder	0	12,87	165,64	0	1	12,27	150,55	150,55
Summe der Quadrate (*SS*)				34,74				185,78
Stichprobe (*N*)	30				30			
Stichprobenvarianz	34,74 / (30-1) = 1,20				185,78 / (30-1) = 6,41			
Standardabweichung (*SD*)	1,10				2,53			

VI. Inferenz- und Bayesianische Statistik

Während mit Hilfe deskriptiver Statistik nur Aussagen über die Verteilung von Merkmalen innerhalb der Stichprobe, aus der die Daten stammen, möglich sind, hat die **Inferenzstatistik** – auch **induktive** oder **frequentistische Statistik** genannt – den Vorteil, aus der Analyse einer Stichprobe Aussagen über die Verteilung von Merkmalen einer Population (Grundgesamtheit) treffen zu können – also ohne dass die Daten zu dieser Population vollständig vorliegen.[115]

Grundlage für diese Schlüsse sind **Wahrscheinlichkeiten**, die in der Inferenzstatistik als **relative Häufigkeit in Zufallsexperimenten** definiert sind.[116] Wird ein fairer Würfel unzählige Male geworfen, so liegt die Wahrscheinlichkeit eine „6“ zu würfeln bei einem Sechstel; Gleiches gilt für die übrigen Zahlen des Würfels.[117] Auf Schwierigkeiten stößt dieser Wahrscheinlichkeitsbegriff in solchen Fällen, in denen das Ereignis, dessen Wahrscheinlichkeit bestimmt werden soll, bereits stattgefunden hat und das Ergebnis feststeht, oder wenn es keine Vielzahl von Ereignissen gibt. In beiden Fällen

115 *Handl/Kuhlenkasper,* Fn. 47, S. 223; *Schäfer,* Fn. 56, S. 109 f.; *Hagen,* Fn. 107, S. 135.

116 *Finkelnstein,* Basic Concepts of Probability in Statistics and the Law, S. 1; *Tschirk,* Statistik: Klassisch oder Bayes, S. 18.

117 Vgl. auch unten die Ausführungen zum Gesetz der Großen Zahl, B.VI.1.

liegt kein Zufallsexperiment vor, so dass die Wahrscheinlichkeit nicht mittels relativer Häufigkeiten erfasst werden kann.

Einen anderen Wahrscheinlichkeitsbegriff, der für sich in Anspruch nimmt, auch solche Wahrscheinlichkeiten erfassen zu können, bietet die **Bayesianische Statistik**, die bislang ihren Anwendungsbereich vor allem in den Technikwissenschaften und in der künstlichen Intelligenz findet, aber auch für Wahrscheinlichkeiten im Recht nutzbar gemacht werden kann, etwa im Rahmen einer Beweiswürdigung im Strafprozess.[118] In der Bayesianischen Statistik wird Wahrscheinlichkeit subjektiv als **Grad persönlicher Überzeugung** definiert. Dieser Wahrscheinlichkeitsbegriff ist subjektiv, weil sog. ***a-priori*-Wahrscheinlichkeiten** (auch „*priors*“) in die Berechnung mit einbezogen werden. A-priori-Wahrscheinlichkeiten fußen auf bereits bekannten Befunden, die etwa aus vorangegangenen empirischen Studien gewonnen wurden, die aber auch aus Naturgesetzen, Erfahrungswissen, Expertenmeinungen oder Annahmen und Vermutungen herrühren können. Es werden also, anders als bei der Inferenzstatistik, zur Bestimmung von Wahrscheinlichkeiten nicht nur solche Informationen berücksichtigt, die sich unmittelbar aus der Stichprobe ergeben – und die eine Aussage darüber erlauben, mit welcher Wahrscheinlichkeit die Stichprobe zur Hypothese passt[119] –, sondern das gesamte Wissen über den zu beurteilenden Fall; damit wird eine Aussage dazu möglich, wie wahrscheinlich es ist, dass die Hypothese stimmt.[120] Der besondere Charme der Offenlegung dieser „*priors*“ liegt darin, dass die Annahmen der eine statistische Analyse durchführenden Wissenschaftler*innen transparent gemacht werden und ihrerseits in Frage gestellt werden können; allerdings sind die anzustellenden Berechnungen ungleich komplexer.

Verdeutlichen lässt sich die unterschiedliche Herangehensweise am Beispiel des Münzwurfs. Mit den Methoden der Inferenzstatistik liegt die Wahrscheinlichkeit, dass beim Münzwurf „Kopf“ obenauf ist, bei 50 %. Allerdings zieht diese Berechnung nicht die Möglichkeit in Betracht, dass die Münze durch Abnutzung oder Beschädigung uneben oder gar manipuliert ist. In der Bayesianischen Statistik kann hingegen der Prior

---

118 *Tschirk,* Fn. 116, S. 3; *Finkelnstein,* Fn. 116, S. 3; *Janßen,* Bayessche Netze in der Rechtsprechung, S. 9 ff; *Schweizer,* Beweiswürdigung und Beweismaß, S. 168 ff..

119 *Tschirk,* Fn. 116, S. 2.

120 *Tschirk,* Bayes-Statistik für Human und Sozialwissenschaftler, S. 8; *Döring/Bortz*, Fn. 62, S. 615 ff.

„unebene Münze“ Berücksichtigung finden. Damit ist sie nach Auffassung ihrer Befürworter besser geeignet, aussagekräftige Ergebnisse mit geringerer Fehleranfälligkeit hervorzubringen.[121] Auch mit Blick auf das Antidiskriminierungsrecht ist die Bayesianische Statistik mit ihrer Berücksichtigung von *a-priori*-Wahrscheinlichkeiten (*priors*) ein interessanter und vielversprechender Ansatz. Da sie aber in der (rechts-) empirischen Forschung und Literatur bislang nur höchst selten eingesetzt wird, bleibt sie im Folgenden außen vor.[122]

1. Stichproben

Eine Stichprobe erlaubt nur dann einen zulässigen Schluss auf die Grundgesamtheit, wenn sie **repräsentativ** ist. Die Personen, die in der Stichprobe herangezogen werden, müssen demnach „durchschnittlich“ jenen Personen entsprechen, die die Grundgesamtheit bilden. Die Ergebnisse einer Umfrage unter Grundschüler*innen lassen sich z.B. nicht auf Berufsschüler*innen verallgemeinern und übertragen. Es entstünde ein verzerrtes Bild der Grundgesamtheit Schüler*innen, da bestimmte Elemente der Grundgesamtheit in der Stichprobe nicht vertreten sind.[123] In diesem Fall liegt ein **Selektions-Bias**[124] vor, da Berufsschüler*innen bei der Auswahl der Stichprobe nicht berücksichtigt wurden. Häufig ist ein Selektions-Bias auch bei Online-Befragungen zu beobachten, weil selbst bei zufälliger Ziehung der Stichprobe nur eine bestimmte Gruppe – nämlich Menschen mit Internet-Zugang und entsprechender Digitalkompetenz – als Studien-Teilnehmer*innen gewonnen werden können. Dies gilt auch für Big Data: Hier ist die Grundgesamtheit in der Regel nicht definiert bzw. es gibt einen Selektions-Bias zugunsten von Nutzer*innen digitaler Dienste, Social-Media-Plattformen etc., die Stichprobe ist damit häufig nicht repräsentativ.[125]

Die Stichprobe kann aber auch dann **verzerrt** sein, wenn nicht alle befragten Personen auch alle Fragen beantworten, oder wenn der Rücklauf verschickter Fragebögen

---

121 *Tschirk,* Fn. 116, S. 2.

122 Zu den Gründen s. *Tschirk*, Fn. 120, S. 9. Wir werden aber im Rahmen der Beschreibung möglicher Fehlerquellen durch die Beurteilung von Wahrscheinlichkeiten durch Entscheider*innen noch einmal auf die Grundsätze der Bayesianischen Statistik zurückkommen; vgl. C.II.

123 *Handl/Kuhlenkasper,* Fn. 47, S. 307 ff.

124 *Finkelnstein,* Fn. 116, S. 98.

125 *Trübner/Mühlichen,* Fn. 91, S. 143 (147); *Mahrt,* Fn. 94, S. 23 (27).

gering ist (***Nonresponse-Bias***).[126] Dies ist häufig der Fall bei Fragen, die von den Befragten als zu persönlich empfunden werden, wie zum Beispiel Fragen nach dem Haushaltseinkommen oder dem Vermögen; bei Online-Befragungen ist ein häufiges Problem, dass die Einstellungen von Soft- oder Hardware (etwa Browser oder Smartphone) die Aufzeichnung bestimmter Daten verhindern.[127] Ebenso kann es sein, dass bei freiwilligen Befragungen nur solche Teilnehmer*innen antworten (bzw. manche Fragen nur von solchen Teilnehmer*innen beantwortet werden), die von der Frage in irgendeiner Weise betroffen sind.[128]

Bisweilen werden Umfragen ausdrücklich als „repräsentativ" bezeichnet. Sofern die für solche Umfragen herangezogene Stichprobe nicht zufällig aus der Grundgesamtheit gezogen wurde (was selten der Fall sein dürfte), sind sie allerdings tatsächlich lediglich *hinsichtlich der beobachteten Merkmale* (z.B. Geschlecht, Alter, Wohnort) repräsentativ.

Die Repräsentativität von Stichproben für eine Gesamtpopulation kann allein und ausschließlich dadurch gewährleistet werden, dass die Stichproben zufällig aus der Population gezogen werden, wie aus dem **Gesetz der großen Zahlen** von *Jacob Bernoulli* (1654-1705) folgt.[129] Danach nähert sich die **relative Häufigkeit** eines Zufallsereignisses der **Wahrscheinlichkeit** dieses Zufallsereignisses an, je häufiger es durchgeführt wird. Anders ausgedrückt: Durch häufige Durchführung einer zufälligen Ziehung können systematische Zusammenhänge oder Muster sichtbar gemacht werden. Das paradigmatische Beispiel ist der Wurf eines Würfels. Die Zahlen 1 – 6 sind gleichmäßig auf diesem Würfel verteilt, jede Zahl sollte also gleich häufig geworfen werden. Wird der Würfel nun sechs Mal geworfen, so ist es nicht wahrscheinlich, dass jede Zahl auch einmal geworfen wird und die Würfe so die gleichmäßige Verteilung der Zahlen auf dem Würfel repräsentieren. Bei tausend oder zehntausend Würfen aber wird die beobachtete relative Häufigkeit jeder Augenzahl in etwa gleich sein,

---

126 *Proner,* Ist keine Antwort auch eine Antwort? S. 33; *Handl/Kuhlenkasper,* Fn. 47, S. 308; *Finkelnstein,* Fn. 116, S. 101.

127 *Trübner/Mühlichen,* Fn. 91, S. 143 (148).

128 *Handl/Kuhlenkasper,* Fn. 47, S. 308; *Cleff,* Angewandte Induktive Statistik und Statistische Testverfahren, S. 6 ff.

129 *J. Bernoulli,* Ars Conjectandi [The Art of Conjecturing] 225 (1713), zit. in *Stigler*, The History of Statistics: The Measurement of Uncertainty Before 1900, S. 65.

sofern es sich um einen fairen Würfel handelt; ist der Würfel manipuliert, lassen sich die tatsächlichen Wahrscheinlichkeiten für die jeweiligen Ausgänge eines Wurfes an der relativen Häufigkeit einer jeden Augenzahl in der gezogenen Stichprobe ablesen.

Dies bedeutet, dass eine häufig und zufällig gezogene Stichprobe im Mittel hinsichtlich aller Merkmale repräsentativ für die Population ist, da sie bei vielfacher zufälliger Ziehung dem Durchschnitt der Population entspricht. Ist die Stichprobe zu klein, so ist sie zwar hinsichtlich eines Merkmals repräsentativ, es besteht aber die Gefahr, dass sie hinsichtlich weiterer Merkmale untypisch ist und damit nicht der tatsächlichen Häufigkeit dieser Merkmale in der Population entspricht. Da Sinn und Zweck der Stichprobe aber gerade ist, mit Hilfe der beobachteten Werte die Werte in der Population (die sich in der Regel nicht vollständig erfassen bzw. erheben lässt) zu schätzen, ist eine möglichst exakte Abbildung der Population in der Stichprobe unabdingbar. Je größer also eine Stichprobe ist, umso zuverlässiger sind die aus ihr abgeleiteten Einsichten.

2. Hypothesen

Die aus der Theorie und dem Modell abgeleiteten Hypothesen, die entweder auf einen Zusammenhang zwischen Merkmalen oder auf einen Unterschied zwischen bestimmten Gruppen gerichtet sind, werden im Rahmen eines **Hypothesentests** daraufhin untersucht, mit welcher Wahrscheinlichkeit der Zusammenhang bzw. Unterschied besteht.[130]

Hierfür bedarf es zunächst einer sog. **Nullhypothese** (H_0) und der **Alternativhypothese** (H_1). Die Nullhypothese geht dabei in der Regel davon aus, dass ein Unterschied oder Zusammenhang nicht besteht, während die Alternativhypothese (Forschungshypothese) vom Bestehen eines Unterschieds oder Zusammenhangs ausgeht.[131] Sie beschreibt den erwarteten Effekt, den die unabhängige Variable (z.B. Geschlecht) auf die abhängige Variable (z.B. Gehalt) haben soll. Aus epistemischen Gründen können Hypothesen nur verworfen werden; deshalb sollten Nullhypothese und Alternativhypothese so formuliert sein, dass sie gemeinsam alle möglichen Erklärungen beinhalten und sich gleichzeitig gegenseitig ausschließen (**MECE**-Regel: ***m**utually **e**xclusive, **c**ollectively **e**xhaustive*), die Alternativhypothese also wirklich die

---

130 S.o. B.II.; *Schäfer,* Fn. 56, S. 139.

131 *Schäfer,* Fn. 56, S. 158.

eine Alternative zur Nullhypothese darstellt: Denn dann liegt in der Widerlegung der Nullhypothese automatisch die Bestätigung der Alternativhypothese.

Ist die Wahrscheinlichkeit gering, dass ein durch eine Hypothese behaupteter und in den Daten beobachteter Unterschied zwischen Merkmalsträger*innen oder Zusammenhang zwischen Merkmalen auf einem Irrtum beruht oder durch den Zufall erzeugt wird (und eben nicht durch einen musterzeugenden Mechanismus[132]), so wird dieses Ergebnis als **signifikant** bezeichnet. Ein Irrtum kann entweder darin bestehen, dass ein Unterschied angenommen wird, obwohl er nicht existiert (**Fehler 1. Art**, *type I error* oder **α-Fehler** => „falsch positiv"); oder darin, dass ein tatsächlich bestehender Unterschied nicht erkannt wird (**Fehler 2. Art**, *type II error* oder **β-Fehler** => „falsch negativ"). Ein strengeres **Signifikanzniveau** birgt eine höhere Wahrscheinlichkeit für einen Fehler 2. Art, also die fälschliche Ablehnung der Alternativhypothese.[133]

	Nullhypothese (H_0) ist	
	wahr	**falsch**
Test weist H_0 zurück	**Fehler 1. Art** (= **falsch positiv**) (Wahrscheinlichkeit = α)	Richtiges Ergebnis (wahr positiv) (Wahrscheinlichkeit = 1 - β)
Test weist H_0 nicht zurück	Richtiges Ergebnis (wahr negativ) (Wahrscheinlichkeit = 1 - α)	**Fehler 2. Art** (= **falsch negativ**) (Wahrscheinlichkeit = β)

Mit welcher Wahrscheinlichkeit ein Unterschied oder Zusammenhang signifikant ist, wird durch das Signifikanzniveau (α) festgelegt. Die Konventionen für die Bestimmung des Signifikanzniveaus sind je nach Disziplin unterschiedlich, die in der Rechtsempirie häufigste Konvention ist wohl 10 % = marginal signifikant, 5 % = signifikant; 1 % = hoch signifikant;[134] das bedeutet, dass ein Ergebnis dann als „statistisch signifikant" (und damit als verlässlich) eingestuft wird, wenn die Wahrscheinlichkeit, dass der Zusammenhang irrig angenommen wird, kleiner als 5 % ist. Empirische Studien berichten häufig den ***p*-Wert** ihrer Tests, der Auskunft über die statistische Signifikanz des

---

132 Ein mustererzeugender Mechanismus kann beispielsweise ein aus dem *rational actor model* abgeleitetes Verhaltensmuster („as if") sein, so dass sich dieser Ansatz besonders für sozialwissenschaftliche empirische Studien eignet.

133 *Hagen,* Fn. 107, S. 137.

134 *Goerg/Petersen,* Fn. 5, Rn. 444.

Ergebnisses gibt: Er bezeichnet die Wahrscheinlichkeit für einen Fehler der 1. Art; nur wenn er kleiner ist als das Signifikanzniveau, ist das Ergebnis statistisch signifikant.[135]

3. Testen von Zusammenhängen

Insbesondere zum Nachweis struktureller Diskriminierungen eignen sich statistische Testverfahren, die Aussagen darüber treffen können, ob zwischen zwei sozialen Faktoren (**Variablen**) ein statistischer Zusammenhang besteht.[136] Dabei bezeichnet man das soziale Phänomen, dessen Auftreten erklärt werden soll, als **abhängige Variable** und das Phänomen, das die Veränderungen in der abhängigen Variablen erklären soll, als **unabhängige Variable**.[137]

a) Korrelationen

Sollen die Zusammenhänge zwischen zwei gemessenen Variablen untersucht werden, so erfolgt dies mit Hilfe von Korrelationen.[138]

Nicht selten werden Zusammenhänge zwischen Variablen in grafischen Darstellungen wie zum Beispiel Streudiagrammen bereits sichtbar: Die Datenpunkte verlaufen ungefähr in einer Linie und auf X- und Y-Achse gleichermaßen ansteigend (**positive Korrelation**) oder absteigend (**negative Korrelation**). Sind die Daten nicht miteinander korreliert, so ist die Punktwolke eher kreisförmig, ein Zusammenhang nicht erkennbar.

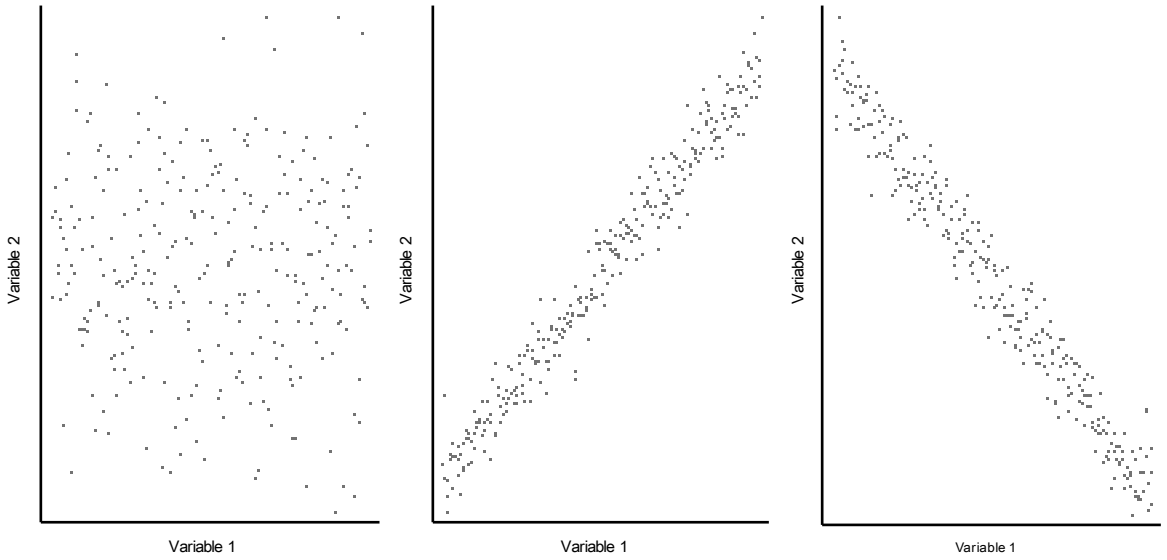
---

135 *Goerg/Petersen,* Fn. 5, Rn. 444.

136 *Hagen,* Fn. 107, S. 89 ff.

137 *Goerg/Petersen,* Fn. 5, Rn. 459.

138 S.o. B.I.



Quelle: *Goerg/Petersen,* Fn 5, Rn. 459.

Um Aussagen darüber treffen zu können, wie stark dieser Zusammenhang ist, kann die Korrelation berechnet werden.[139] Der hierfür verwendete Wert ist der sog. **Korrelationskoeffizient** *r* (nach Pearson), der den (linearen) Zusammenhang zwischen den Variablen widerspiegelt, die sich einer linearen Funktion annähern. Der Wert des Korrelationskoeffizienten bewegt sich dabei zwischen 1 und -1, wobei bei einem Korrelationskoeffizienten von 0 die Variablen nicht korreliert sind, also in keinerlei Zusammenhang stehen. Ob ein Effekt nennenswert – relevant – ist, wird mit Hilfe von Maßen für die **Effektstärke** beurteilt, etwa mit Cohens *d*. In den Sozialwissenschaften wird ein Korrelationskoeffizient von $r > 0{,}1$ als schwach, von $r > 0{,}3$ als mittel und ab $r > 0{,}8$ als stark eingestuft; letztlich ist die Einstufung aber abhängig von den *a priori* Erwartungen des Modells.

Würde sich in einem – fiktiven – Datensatz zeigen, dass die Abiturnote im Fach Deutsch perfekt positiv mit der Staatsexamensnote korreliert ist ($r = 1$), dann würde ein Streudiagramm für jeden zusätzlichen Punkt der Abiturnote Deutsch eine Steigung von 1,2 Punkten im Staatsexamen zeigen. Als reales Beispiel mag hier eine 2011 veröffentlichten Studie dienen, die unter anderem das Ausmaß gruppenbezogener Menschenfeindlichkeit untersuchte und einen Zusammenhang zwischen Fremden- und Islamfeindlichkeit zeigte; der Korrelationskoeffizient betrug $r = 0.59$. Auch Fremdenfeindlichkeit und Antisemitismus waren mit einem Korrelationskoeffizienten von

---

139 *Goerg/Petersen,* Fn. 5, Rn. 459; *Schäfer,* Fn. 56, S. 91.

r = 0.41 korreliert, es handelt sich also um Korrelationen mittlerer Stärke, die einen Zusammenhang zwischen den untersuchten Merkmalen nahelegen.[140]

b) Regressionen

Während die Korrelation Auskunft darüber gibt, *ob* zwischen zwei Variablen ein Zusammenhang besteht, kann mit Hilfe einer Regression genauer ermittelt werden, *wie stark* der Zusammenhang zwischen abhängiger (zu erklärender) und einer oder mehreren unabhängigen (erklärenden) Variablen ist.[141] Die Regression ermöglicht auch die Vorhersage, ob die unabhängige Variable die abhängige beeinflusst (Ursachenanalyse), eine Veränderung der unabhängigen Variable die abhängige ebenfalls verändert (Wirkungsanalyse) oder wie sich die abhängige Variable mit der Zeit verändert (Zeitreihenanalyse).[142]

Hierfür nutzt sie die Korrelation von Variablen, um die Werte der einen Variable aus den Werten der anderen Variablen vorherzusagen, so dass beispielsweise die Frage beantwortet werden kann, wie das Einkommen einer Arbeitnehmer*in von Bildung, Alter und Geschlecht abhängt oder wie sich Geschlecht und/oder Herkunft von Examenskandidat*innen auf ihre Examensergebnisse auswirken.[143]

In einem Regressionsmodell wird die Funktion geschätzt, die die Punktwolke aus der Korrelation am besten beschreibt. Wird beispielsweise ein linearer Zusammenhang angenommen (wie er sich in den oben abgebildeten Punktwolken aufdrängt), so wird die aus der Oberstufenmathematik in ihrer Grundform bekannte Funktion

$$y = f(x) = \alpha + \beta \bullet x + \varepsilon$$

„geschätzt"; bei quadratischen Zusammenhängen wird eine quadratische Funktion (x^2) geschätzt, ferner kann auch jede andere Art bekannter Zusammenhänge (etwa logarithmische) geschätzt werden. Der Anschaulichkeit halber wird die (hinreichend komplexe) Funktionsweise von Regressionen im Folgenden mit der – im Rahmen

---

140 *Zick/Küpper/Hövermann,* Die Abwertung der Anderen: eine europäische Zustandsbeschreibung zu Intoleranz, Vorurteilen und Diskriminierung, S. 79 ff.

141 *Goerg/Petersen,* Fn. 5, Rn. 467.

142 Ausführlich zu Regressionen *Goerg/Petersen,* Fn. 5, Rn. 466 ff.; *Schäfer,* Fn. 56, S. 103; *Kosfeld,* Fn. 108, S. 225 ff.

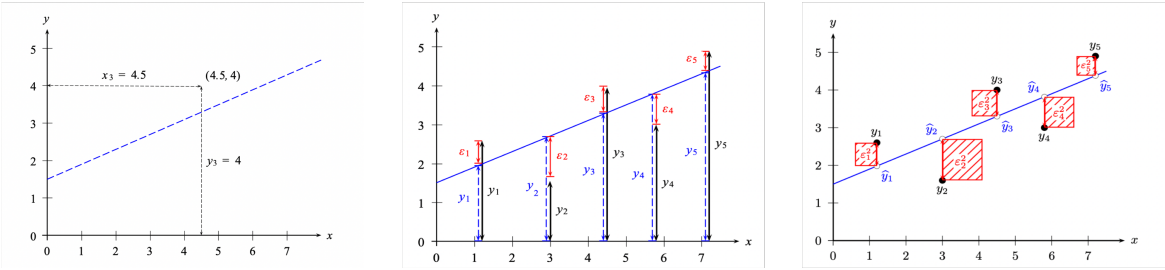
143 *Traxer/Glöckner/Towfigh,* Fn. 66, S. 115 ff.

rechtsempirischer Forschung wohl am häufigsten eingesetzten – linearen Regression (*ordinary least squares* = **OLS-Regression**) dargestellt.

Die **Steigung** dieser Funktion wird mit dem Koeffizienten β bezeichnet. Sie spiegelt den Einfluss der unabhängigen Variablen auf die abhängige Variable wider. α bezeichnet die **Konstante**, also den Wert, an dem die Funktion die y-Achse schneidet (auch y-Achsenabschnitt genannt); und ε bezeichnet den Fehlerterm (***error term***, auch **Störgröße** oder **Residuum**), also den Abstand des jeweiligen konkreten Punkts in der Punktwolke von der Funktion. Dabei handelt es sich gleichsam um die Summe der nicht beobachteten Variablen, die nicht mit x korreliert sein dürfen, folglich im Mittel alle Ziehungen der Stichprobe gleichermaßen betreffen – und damit als „weißes Rauschen" betrachtet werden dürfen.

Die OLS-Regression bestimmt die zu den Daten am besten „passende" Funktion (d.h. insbesondere die angemessenen Koeffizienten), indem sie die Abweichungen jedes einzelnen Punkts der Punktwolke quadriert, dieses Ergebnis addiert und aus den möglichen Funktionen jene auswählt, bei der die Summe der quadrierten Abweichungen am geringsten ist.

1. 2. 3.



Darüber hinaus kann auch der Einfluss mehrerer unabhängiger Variablen (z.B. Geschlecht und Herkunft) auf die abhängige Variable (z.B. y = Note im juristischen Staatsexamen) geschätzt werden, etwa wie folgt:

$$y_x = \alpha + \beta1 \bullet Geschlecht_x + \beta2 \bullet Herkunft_x + \gamma \bullet Kontrollvariablen_x + \varepsilon_x$$

Geschlecht und Herkunft beeinflussen mit ihren Koeffizienten β1 und β2 die Steigung der Funktion y. Je nachdem, ob sich die Variable „Herkunft" auf Examenskandidat*innen mit oder ohne Migrationshintergrund bezieht oder ob die Examenskandidat*innen männlich oder weiblich sind, verändert sich die Funktion.

Neben dem Koeffizienten β sind auch **Gütemaße** der Regression und die Signifikanzen der Regressionskoeffizienten von Bedeutung. Das **Bestimmtheitsmaß** (R^2) und die ***F*-Statistik** stellen dabei gängige Gütemaße dar; sie geben Auskunft darüber, wie gut die geschätzte Regressionsfunktion die empirischen Daten abbildet.

Das Bestimmtheitsmaß R^2 kann Werte zwischen 0 und 1 annehmen; je höher es ist, desto besser wird die in den Daten vorhandene Streuung[144] durch die Regressionsgleichung erklärt. Um mit dem Koeffizienten β Vorhersagen über die Stichprobe hinaus treffen zu können, sollte R^2 möglichst groß sein. Wäre R^2 zu klein, bedeutete das für unser Beispiel, dass Herkunft bzw. Geschlecht keine geeigneten Faktoren wären, um Unterschiede im Examensergebnis zu erklären.

Die *F*-Statistik gibt Auskunft darüber, ob das Ergebnis einer Regressionsrechnung von der Stichprobe auf die Population übertragen werden kann. Ist die *F*-Statistik signifikant, so bedeutet dies, dass generell ein statistisch robuster Zusammenhang zwischen mindestens einer unabhängigen und der abhängigen Variablen besteht. Da der *F*-Test aber keine Aussage darüber treffen kann, ob alle bzw. welche unabhängigen Variablen in einem signifikanten Zusammenhang mit der abhängigen Variablen stehen, wird zusätzlich mittels der ***t*-Statistik** für jede einzelne unabhängige Variable geprüft, ob ein signifikanter Zusammenhang mit der abhängigen Variablen besteht.

Mit der Regressionsanalyse ist es ferner möglich, **Interaktionseffekte** zu erkennen. Als Interaktionseffekt werden solche Einflüsse auf die abhängige Variable bezeichnet, die durch die gleichzeitige Berücksichtigung bzw. durch die Kombination mehrerer unabhängiger Variablen entstehen. Es kann also untersucht werden, welche Auswirkungen das gemeinsame Auftreten unabhängiger Variablen auf die abhängige Variable hat. Für die Antidiskriminierungsforschung ist diese Möglichkeit wertvoll, da mit ihr **intersektionale Diskriminierung** erforscht werden kann.[145] Es ließe sich zum Beispiel zeigen, wie sich die abhängige Variable „Gehalt“ verändert, wenn neben dem Merkmal „Geschlecht“ auch noch das Merkmal „Herkunft“ oder „Kinderzahl“ in der Regressionsanalyse berücksichtigt wird.

---

144 S.o. B.V.2.

145 *Rouhani,* Intersectionality-informed Quantitative Research: A Primer; *Bowleg,* When Black + Lesbian + Woman ≠ Black Lesbian Woman: The Methodological Challenges of Qualitative and Quantitative Intersectionality Research, Sex Roles 59 (2008), 312 ff.

VII. Anforderungen an die Messung

Ganz gleich, welcher Art die zugrunde liegenden Daten und das angewendete Testverfahren sind, in jedem Falle ist wichtige Voraussetzung für seriöse Forschungsergebnisse, dass der Test objektiv (**Objektivität**), zuverlässig (**Reliabilität**) und gültig (**Validität**) ist.[146]

1. Objektivität

Durch die Objektivität eines Tests soll sichergestellt werden, dass die Ergebnisse vergleichbar bleiben, auch wenn unterschiedliche Personen den Test durchführen.[147] Die Messung muss unabhängig von der Leiter*in des Tests sowie von den mit der Auswertung betrauten Personen sein. Durch einen hohen Grad an Standardisierung und durch multiple Beurteiler*innen kann erreicht werden, dass wenig Spielraum verbleibt, der einen Einfluss der durchführenden Personen auf die Art der Durchführung, der Auswertung und der Interpretation der Messungen zulässt.[148]

Objektivität ist besonders bei der Analyse von Big Data eine Herausforderung.[149] Typischerweise sind diese Datensätze sehr umfangreich, die in ihnen enthaltenen Daten stehen oftmals nicht in einem inneren Zusammenhang. Auch ist häufig der zugrunde liegende Erhebungsmechanismus nicht klar, da die Erhebung nicht theoriegeleitet erfolgte,[150] so dass die Gefahr besteht, bei der Operationalisierung etwas womöglich von der Analyst*in Erwünschtes in die Daten hineinzulesen.

Laborexperimente weisen dagegen typischerweise eine hohe Objektivität bei der Durchführung auf, da hier die Testbedingungen konstant sind und kaum eine Beeinflussung der Testpersonen etwa durch das Auftreten oder die Art und Weise der Fragestellung durch die den Test durchführende Person erfolgen kann. Gerade die Auswertung offener Fragen bedarf indessen klarer Vorgaben, wie mit unterschiedlichen Antworten umzugehen ist und wie diese gegebenenfalls zu gewichten sind.[151]

---

146 *Glöckner/Towfigh,* Messgenauigkeit und Fairness in Staatsprüfungen, AnwBl 2016, 706; *Cleff,* Fn. 128, S. 5.; *Kosfeld/Eckey/Türk,* Fn. 108, S. 21; *Krebs/Menjold,* Gütekriterien quantitativer Sozialforschung, in: *Baur/Blasius* Fn. 49, S. 489 ff.

147 *Moosbrugger/Kelava,* (Hrsg.)*,* Testtheorie und Fragebogen-Konstruktion, S. 8.

148 *Glöckner/Towfigh* Fn. 146, S. 706; *Moosbrugger/Kelava,* Fn. 147, S. 8.

149 S.o. B.IV.3.

150 S.o. B.II.1.

151 *Krebs/Menjold,* Fn. 146, S. 491 ff.

Mit Blick auf die Interpretation der Testwerte ist eine eindeutige und klare Standardisierung unerlässlich, damit die Schlussfolgerungen einheitlich und vergleichbar sind.[152] In der Antidiskriminierungsforschung sind es häufig Interessenvertretungen – wie etwa die Antidiskriminierungsstelle des Bundes oder sich der Antidiskriminierungspolitik widmende Stiftungen[153] –, die entsprechende Forschung initiieren und ihren Fokus darauf richten, Diskriminierungen nachzuweisen. Das Objektivitätserfordernis stellt nicht nur in diesen Fällen hohe Anforderungen an die Fähigkeit der Forschenden zur (Selbst-)Reflektion ihrer Subjektivität und Perspektivität.[154]

2. Reliabilität

Reliabilität bezeichnet die Messgenauigkeit eines Tests.[155] Keine Messung ist fehlerfrei. Messfehler, die dazu führen, dass der Messwert nicht mit der tatsächlichen Merkmalsausprägung übereinstimmt, können dabei in der Person, in der Situation oder in deren Zusammentreffen liegen. Reliabilität wird unter anderem dadurch hergestellt, dass die Ergebnisse von unabhängigen Teilen des Tests miteinander korreliert werden oder vergleichbare Tests wiederholt durchgeführt werden. Beispiele hierfür sind etwa das *Test-Retest*-Design (bei dem mit zeitlichem Abstand eine Messung wiederholt wird) oder die in den Sozialwissenschaften eher selten eingesetzten Paralleltests (bei der zwei verschiedene Messinstrumente verwendet werden und eine Übereinstimmung der Werte ein Indiz für ihre Reliabilität ist).[156] Auch die Reliabilität wird als Korrelationskoeffizient ausgedrückt, der für den Nachweis der Messgenauigkeit hoch sein muss.[157] Sie wird etwa mit **Cronbachs** α angegeben.

3. Validität

Auf der Grundlage der in statistischen Testverfahren – wie z.B. der Regressionsanalyse – ausgewerteten Daten werden schließlich die eigentlichen Schlussfolgerungen gezogen. Diese Schlussfolgerungen sind in ihrer Aussagekraft begrenzt und können

---

152 *Goldhammer/Hartig,* Interpretation von Testresultaten und Testeichung, in: Moosbrugger/Kelava (Hrsg.), Fn. 147, S. 174; *Döring/Bortz,* Fn. 62, S. 70.

153 *Lenhart/Roth,* Antidiskriminierung als zivilgesellschaftliches Projekt, in: Scherr/El-Mafaalani/Yüksel (Hrsg.), Fn. 25, S. 626 ff.

154 Zur wissenschaftstheoretischen Dimension dieser Frage vgl. *Döring/Bortz,* Fn. 62, S. 61 ff.

155 *Glöckner/Towfigh,* Fn. 146, S. 707.

156 *Häder,* Empirische Sozialforschung, S. 110 ff. auch zu weiteren Verfahren zur Überprüfung der Reliabilität und ihrer Schwachpunkte.

157 *Moosbrugger/Kelava,* Fn. 147, S.11.

aus den oben erwähnten epistemischen Gründen keinen sicheren Beweis darstellen. Sie sind aber in ihrer Begrenztheit verwertbar und aussagekräftig, solange sie valide sind. Validität bewertet und beurteilt in diesem Zusammenhang die **Qualität einer Schlussfolgerung** (vgl. im Unterschied hierzu die Ausführung zur Validität *des Tests* oben unter III.). Dabei müssen verschiedene Validitätskriterien erfüllt sein: (a) **statistische Validität**, (b) **interne Validität** und (c) **externe Validität**.[158]

Die Validität ist das wichtigste Kriterium zur Beurteilung der Qualität einer Messung, Reliabilität und Objektivität allein vermögen einen nicht validen Test nicht zu retten.[159] Die Messergebnisse einer falsch geeichten Waage hängen nicht von der die Messung durchführenden Person ab (sind also objektiv) und ändern sich auch bei wiederholter Messung nicht (sind folglich auch reliabel), sie sind aber dennoch nicht valide, da eine korrekt geeichte Waage andere – richtige – Messergebnisse hervorbringen wird.

a) Statistische Validität

Statistische Validität einer Schlussfolgerung wird dann bejaht, wenn davon ausgegangen werden kann, dass die **Beobachtung nicht zufällig** war, sondern auf einer Gesetzmäßigkeit beruht. Ein Indiz hierfür ist die statistische **Signifikanz** eines Effekts.[160] Aber auch das Gesetz der großen Zahlen wird zur Beurteilung der statistischen Validität herangezogen: Erst ab einer gewissen Zahl an Beobachtungen (**Teststärke** oder ***statistical power***) kann eine Aussage darüber getroffen werden, ob eine Beobachtung bloß zufällig ist oder ob ihr eine Gesetzmäßigkeit zugrunde liegt; so ist beispielsweise die statistische Validität einer Studie mit nur wenigen Beobachtungen gering.

b) Interne Validität

Die interne Validität beurteilt, ob das Ergebnis kohärent ist. Sie soll verhindern, dass bestimmte **Störfaktoren** nicht ausgeschlossen wurden bzw. berücksichtigt blieben und damit das Ergebnis beeinflusst haben. Interne Validität wird sichergestellt durch ein gutes Forschungsdesign, welches sämtliche Faktoren identifiziert und berücksichtigt, die Einfluss auf die erklärenden oder zu erklärenden Variablen haben können. An

---

158 S*hadish/ Cook/Campbell,* Experimental and quasi-experimental designs for generalized causal inference, S. 33 ff.

159 *Hartig/Frey/Jude,* Validität, in: Moosbrugger/Kelava (Hrsg.)*,* Fn. 147, S.144 ff.; *Döring/Bortz,* Fn. 62, S. 469 ff.

160 S.o. B.VI.1.

interner Validität und Aussagekraft der Studienergebnisse kann es auch dann fehlen, wenn potentielle Störfaktoren nicht beobachtbar oder messbar sind.

Schließlich ergibt sich eine gewisse Nähe zur Operationalisierung der Forschungsfrage und der ihr zugrundeliegenden Konzepte: Misst der Test mehr oder andere Konzepte (also etwa Sprachkenntnis statt Sprachkompetenz oder neben der Einstellung zu Ausländer*innen auch den Nationalismus), so mangelt es der Messung auch an interner Validität.[161]

c) Externe Validität

Eine empirische Studie soll den induktiven Schluss von der Stichprobe auf die Population, mithin eine **Verallgemeinerung**, ermöglichen. Ein Schluss auf allgemeine Gesetzmäßigkeiten geht damit nicht zwingend einher: Die Ergebnisse einer Studie, die mit Student*innen durchgeführt wurde, können zwar auf „alle Student*innen" übertragen werden, aber nicht ohne Weiteres auch auf Rentner*innen. Hierzu müsste das Forschungsdesign entsprechend angepasst werden und auch Rentner*innen mit einbeziehen. Die externe Validität (**Verallgemeinerbarkeit**) ist also beschränkt: Sie genügt für die Annahme, dass die Ergebnisse der Stichprobe auf die Population der Student*innen im Allgemeinen übertragbar sind, reicht aber nicht so weit, dass angenommen werden darf, die Befunde gälten auch für andere Populationen.

Die externe Validität verlangt daher regelmäßig ein umfangreiches Forschungsprogramm und die Berücksichtigung einer großen Zahl von Szenarien und Variablen. Damit kollidiert sie häufig mit den Anforderungen an die interne Validität, die gerade auf eine Begrenzung der zu untersuchenden Faktoren unter Ausschluss möglicher Störfaktoren setzt. Hier einen angemessenen Ausgleich zu finden, ist eine Herausforderung für empirische Studien.

VIII. *Testing*-Verfahren

Den tatsächlichen Schwierigkeiten des Nachweises von Diskriminierung in einem gerichtlichen Verfahren trägt das AGG mit einer Modifizierung der allgemeinen Regeln zur Darlegungs- und Beweislast Rechnung.[162]

---

161 S.o. B.III.

162 S. → *Muthorst*, § 19a zum Beweisrecht; *Krieger/Günther,* Vorsicht Falle! Diskriminierungsnachweis durch Testing-Verfahren?, NZA 2015, 262; *Schleusener,* Diskriminierungsfreie Einstellung

Unter anderem sog. ***Testing*-Verfahren** sollen dabei helfen, die Vermutung einer Diskriminierung mit entsprechenden Indizien zu untermauern.[163] Im Rahmen von *Testing*-Verfahren wird das Verhalten z.B. von Arbeitgeber*innen, Wohnungseigentümer*innen oder Gastronom*innen gegenüber einer Person, welche eines der Merkmale des § 1 AGG trägt, verglichen mit dem Verhalten gegenüber einer Vergleichsperson, die dieses Merkmal nicht trägt. Zum Beispiel werden identische Bewerbungsunterlagen eingereicht, die sich lediglich im Merkmal „ethnische Herkunft" unterscheiden, oder es werden in im Übrigen identischen Bewerbungsunterlagen unterschiedliche Bewerbungsfotos (männlich/weiblich, mit/ohne religiöses Kopftuch usw.) verwendet. Die Motivation für ein *Testing*-Verfahren liegt in der Regel darin, verdecktes, unmittelbar diskriminierendes Verhalten im Rahmen einer bewusst herbeigeführten Situation sichtbar zu machen, um in einem antidiskriminierungsrechtlichen Prozess den Nachweis einer Diskriminierung führen zu können.[164]

Testing-Verfahren ermöglichen also in erster Linie die Aufdeckung von Diskriminierung in einem **Einzelfall** und stellen ein pragmatisches Mittel dar, in der Rechtsdurchsetzung den Nachweis von Diskriminierung zu erleichtern. Sie bieten die Möglichkeit, kontrafaktische Situationen aufzuzeigen: Wird die Bewerbung einer Bewerber*in mit Migrationshintergrund abgelehnt, eine erneute – bis auf das Merkmal des Migrationshintergrunds identische – Bewerbung auf dieselbe Stelle hingegen nicht, so sind für diesen konkreten Fall die *Indizien* stark, dass die Ablehnung der Bewerbung diskriminierend war.

Aus ökonometrischer bzw. statistischer Perspektive sind *Testing-Verfahren* jedoch kein zuverlässiges Mittel, um Diskriminierung nachzuweisen, da sie in der Regel nicht geeignet sind, verallgemeinerbare Erkenntnisse zu Diskriminierung zu schaffen. Während die Inferenzstatistik aus der Betrachtung einer Stichprobe Aussagen für eine

---

zwischen AGG und Frauenförderungsgesetz, NZA-Beilage 2016, 50; *Grünberger/Reinelt,* Konfliktlinien im Nichtdiskriminierungsrecht, S. 27.

163 BT-Drs. 16/1780, S. 47; *Thüsing*-Mü-Ko- § 22 AGG, Rn. 14; *Payandeh,* Rechtlicher Schutz vor rassistischer Diskriminierung, JuS 2015, 695 (700); *Krieger/Günther,* Fn. 162, S. 262; LAG Schleswig-Holstein, Urt. V. 9.4.2014, ArbRAktuell 2014, 364; *Franke/Schlenzka,* Diskriminierung aufgrund der ethnischen Herkunft und rassistische Diskriminierung im Spiegel von Daten und Rechtsprechung, ZAR 2019, 179 (181); *Arndt,* Das Problem der Diskriminierung – Rechtssoziologische Fallstudien zum Allgemeinen Gleichbehandlungsgesetz (AGG) § 5 A. (Diss., nicht veröffentlicht).

164 So etwa im Rahmen von Verfahren strategischer Prozessführung in Leipzig, s. hierzu Kinsky, Mit Recht gegen Rassismus – Grenzen strategischer Prozessführung im Rahmen des Allge-meinen Gleichbehandlungsgesetzes (AGG) am Beispiel diskriminierender Einlasskontrollen vor Diskotheken, S. 70 ff.

Population treffen kann,[165] kann im *Testing*-Verfahren nur eine Aussage über den konkret „getesteten“ Fall getroffen werden.

Für die Durchführung von *Testing*-Verfahren gilt, dass ihre Tauglichkeit und ihr Wert ganz maßgeblich von der Qualität ihrer Ausführung abhängen. Um ausschließen zu können, dass eine Ablehnung nicht auf Unterschieden beruht, die nicht das Diskriminierungsmerkmal betreffen, müssen die Profile der Merkmalsträger*in und der Vergleichsperson tatsächlich völlig identisch sein, sie dürfen sich nur im Diskriminierungsmerkmal unterscheiden. Schon die unterschiedliche Gestaltung einer inhaltlich ansonsten identischen Bewerbung ist geeignet, Zweifel daran zu begründen, dass die Ablehnung der Bewerber*in (allein) auf das Diskriminierungsmerkmal zurückgeführt werden kann.

Ferner sind die *Testing*-Verfahren unter ethischen Gesichtspunkten nicht unproblematisch. Durch den verdeckten Charakter des Verfahrens befinden sich die „Tester*innen“ unter Umständen im Grenzbereich strafbarer Handlungen.[166] Die Verwendung unrichtiger Bewerbungsunterlagen kann etwa eine Urkundenfälschung nach § 267 StGB darstellen (wobei es in der Regel wohl am subjektiven Tatbestand fehlen dürfte).[167] Abseits von strafrechtlichen Implikationen ist es jedoch auch problematisch und steht im Widerspruch zu den Standards guter wissenschaftlicher Praxis, wenn Testpersonen über ihre Rolle als Testperson getäuscht werden, was beim *Testing*-Verfahren regelmäßig der Fall ist, da es für ein erfolgreiches *Testing*-Verfahren unabdingbar ist, dass „überprüfte“ Arbeitgeber*innen keinerlei Kenntnis von der *Testing*-Situation haben und über die Ernsthaftigkeit der Bewerbungen getäuscht werden.[168] Das mag in einem kontradiktorischen Gerichtsverfahren eine akzeptable Beweislasterleichterung darstellen (denn hier gelten andere Standards als jene guter *wissenschaftlicher* Praxis), im Hinblick auf Methodenstrenge (im Sinne intersubjektiver Vermittelbarkeit) stellt die heimliche Datenerhebung indessen ein Problem dar.

Die Gerichte in Deutschland tun sich (noch) schwer mit der Akzeptanz von in *Testing*-Verfahren gewonnen Beweisen. Ihre tatsächliche Bedeutung ist in Gerichtsverfahren

---

165 S.o. B.VI.

166 *Klose/Kühn,* Fn. 169, S. 23.

167 *Krieger/Günther,* Fn.162, S. 264.

168 Vgl. zum forschungsethischen Problem der Täuschung von Versuchspersonen *Döring/Bortz,* Forschungsmethoden und Evaluation, S. 123 ff.; *Friedrichs,* Fn. 92, S. 67 ff.

daher nach wie vor gering.[169] In einem der wenigen Gerichtsverfahren wegen Diskriminierung, in denen *Testing* eine Rolle spielte, wurde daher gerade mit dem Argument, dass die Diskriminierung „provoziert" worden war, die dem Kläger dem Grunde nach zustehende Entschädigung um die Hälfte gekürzt.[170] Durch *Testing*-Verfahren gewonnenen Indizien hängt daher immer der Makel der Intentionalität an – was allerdings mit Blick auf die Regeln des Zivilprozesses und das Problem der „Waffengleichheit" angesichts der besonderen Schwierigkeit des Nachweises versteckter Diskriminierung jedenfalls so lange nicht zu beanstanden ist, wie das Instrument nicht missbraucht wird („AGG-Hopping").

Die genannten Schwierigkeiten und die fehlende Verallgemeinerbarkeit der in *Testing*-Verfahren gewonnenen Aussagen könnten ein Grund für das vergleichsweise geringe Interesse der Wissenschaft an *Testing*-Verfahren sein, deren Bedeutung also eher rechtspraktisch, bei der Lösung konkreter beweisrechtlicher Fragen anzusiedeln ist.[171]

C. Fehlerquellen beim Einsatz empirischer Methoden

Bei der Rezeption empirischer Studien sind Sorgfalt und Aufmerksamkeit erforderlich; Wissenschaftler*innen und Praktiker*innen, die sich für ihre Arbeit mit empirischen Studien zu Diskriminierung befassen (müssen), sollten im besten Fall in der Lage sein, Schwachstellen einer Studie zu erkennen. Es gilt, Fehler – wie z.B. die angesprochenen Scheinkorrelationen[172] – zu vermeiden bzw. sie bei der Rezeption einer Studie zu erkennen. Jurist*innen sind in der Mehrzahl hierfür nur unzureichend ausgebildet und – Achtung: Vorurteil! – scheuen häufig die ihnen unsympathische Befassung mit „Zahlen". Dabei können potentielle Schwachstellen empirischer Studien von aufmerksamen Rezipient*innen häufig auch ohne ein vertieftes Verständnis der mathematischen Zusammenhänge erkannt werden, wenn sie typischerweise auftretende Fehler und Verzerrungen kennen.

---

169 *Klose/Kühn,* Expertise zur Anwendbarkeit von Testingverfahren im Rahmen der Beweislast, § 22 Allgemeines Gleichbehandlungsgesetz. Expertise, 2010, S. 9; zu Testing-Verfahren in anderen europäischen Ländern vgl. *Klose/Kühn* ebd. S. 12 ff.

170 AG Oldenburg, 23.7.2008, E2 C 2126/07: Rn. 23.

171 Hierzu ausführlich → *Muthorst*, § 19a.

172 S.o. B.I.

I. Verzerrung durch ausgelassene Variablen

Nur wenn möglichst alle relevanten Variablen mit einbezogen werden, kann ein umfassendes Bild entstehen und unter Umständen sogar ein der Diskriminierung zu Grunde liegendes Vorurteil sichtbar gemacht werden. Anderenfalls droht eine Verzerrung der Ergebnisse durch ausgelassene Variablen (***omitted variable bias***).[173] Ausgelassene Variablen verzerren die Schätzung des Regressionskoeffizienten.[174] Dieser gibt dann unter Umständen nicht mehr die kausale Wirkung der unabhängigen Variablen auf die abhängige Variable an.

II. Bayes-Theorem und Basisraten-Fehler

Der Nachweis einer Diskriminierung wird häufig im Rahmen eines gerichtlichen Verfahrens zu führen sein, er ist daher untrennbar mit dem Beweisrecht verbunden.[175] Gerichtliche Entscheidungen sind dabei in einem gewissen Maß immer Entscheidungen unter Unsicherheit.[176] Diesem Umstand trägt das Recht mit Beweislastregeln, Rechtsvermutungen und Fiktionen Rechnung, um Entscheidungen zu ermöglichen, die mit einem hohen Grad an Wahrscheinlichkeit zutreffend sind.[177] So müssen selbst Strafrichter*innen für eine Verurteilung von einem Sachverhalt nicht vollkommen überzeugt sein, „zwingende“ Gewissheit ist nicht erforderlich. Der BGH hält vielmehr ein nach der Lebenserfahrung ausreichendes Maß an Sicherheit, das vernünftige Zweifel nicht aufkommen lässt, für ausreichend.[178] Auch für den Zivilprozess hat der BGH ausgeführt, dass das Gesetz eine von allen Zweifeln freie Überzeugung nicht voraussetzt:

> „Auf die eigene Überzeugung des entscheidenden Richters kommt es an, auch wenn andere zweifeln oder eine andere Auffassung erlangt haben würden. Der Richter darf und muss sich aber in tatsächlich zweifelhaften Fällen mit einem für das praktische Leben brauchbaren Grad von

---

173 S.o. B.I.

174 S.o. B.VI.3.b).

175 Vgl. hierzu ausführlich → *Muthorst*, § 19a.

176 *Engel,* Uncertain Judges, Journal of Institutional and Theoretical Economics, 176 (1), S. 44, 46; *Engel/Timme/Glöckner,* Coherence-Based Reasoning and Order Effects in Legal Judgments, Psychology, Public Policy, and Law, 2020, S. 333 ff; *Engel/Güth,* Modeling a satisficing Judge, Rationality and Society 2018, 30(2), S. 220 ff.

177 *Hagen,* Fn. 107, S. 9.

178 BGH NStZ-RR 2004, 238.

> Gewissheit begnügen, der den Zweifeln Schweigen gebietet, ohne sie völlig auszuschließen."[179]

Ob sich die richterliche Überzeugung dabei auf die Wahrheit der Tatsachenbehauptungen oder auf ihre Wahrscheinlichkeit bezieht, ist eine Frage überwiegend theoretischer Natur.[180] In der Rechtspraxis sind die Auswirkungen gering, da der Begriff der Wahrscheinlichkeit in der Regel nicht in seiner mathematischen Bedeutung verstanden, sondern als eine Heuristik gebraucht wird, um unsicheres Wissen zu umschreiben.[181]

Gleichwohl haben wir es mit unterschiedlichen Konzepten und Maßstäben zu tun, die nicht verwechselt oder vorschnell gleichgesetzt bzw. miteinander verbunden werden sollten. Dem normativen Konzept der „überwiegenden Wahrscheinlichkeit" aus dem Recht lässt sich *per se* kein statistisches Äquivalent zuordnen; allerdings kann eine *normative* Entscheidung getroffen werden, die eine bestimmte statistische Wahrscheinlichkeit zum Zwecke der Beweisführung in einem Gerichtsverfahren ausreichen lässt. Ferner ist zu beachten, dass die Empirie in der Regel Aussagen über die Wahrscheinlichkeit *abstrakter* Fälle trifft, während das Gericht in einem *konkreten* Fall zu seiner Überzeugung gelangen muss. Die Heranziehung von Empirie entbindet daher nicht von der Interpretation solcher Einsichten für den konkreten Fall, was wiederum die Fähigkeit der Richter*innen voraussetzt, mit Daten entsprechend umzugehen.[182] Empirische Befunde können etwa zeigen, dass eine Ungleichbehandlung im konkreten Fall *wahrscheinlich* ist: So könnte eine Studie beispielsweise nachweisen, dass die Wahrscheinlichkeit hoch ist, dass männliche Kandidaten mit ausländisch klingenden Namen im Schnitt eine schlechtere Examensbewertung bekommen; unter Berücksichtigung der sozialen Mechanismen und durch den Ausschluss von Alternativerklärungen durch Kontrollvariablen kann dann ggf. gefolgert werden, dass das *Verfahren* insofern *abstrakt* diskriminierend wirkt. Aussagen zu *konkreten* Einzelbewertungen können aber nicht getroffen werden: Es ist nicht auszuschließen, dass die

---

179 BGH NJW 1970, 946.

180 Mü-KoZPO/Prütting ZPO § 286 Rn. 33; *Engel,* Preponderance of the Evidence vs. Intime Conviction: A Behavioral Perspective on a Conflict between American and Continental European Law, Vermont Law Review, 33, S. 435 (441).

181 *Ihden,* Fn. 9, S. 22 m.w.N.

182 *Engel,* Fn. 180, S. 451.

individuelle Examensnote eines konkreten männlichen Kandidaten mit ausländisch klingendem Namen sogar das Ergebnis einer besonders wohlwollenden Benotung ist.

Der Umgang mit und das Denken in Wahrscheinlichkeiten hält nicht nur in Beweissituationen Fehlerquellen bereit, die bekannt sein sollten, um Fehlurteile zu vermeiden.[183] Häufig widerspricht die menschliche Intuition auch dem zutreffenden Verständnis statistischer Wahrscheinlichkeiten, Schlüsse der Wahrscheinlichkeitstheorie werden häufig als kontraintuitiv erfahren.[184]

Ein Grund für das Auseinanderfallen von menschlicher Intuition und statistischer Wahrscheinlichkeit kann in der Verletzung des in der Bayesianischen Statistik[185] gründenden **Bayes-Theorems** liegen, wonach sich durch die Einbeziehung zusätzlicher Informationen ihr Einfluss auf die „tatsächliche" Wahrscheinlichkeit verändert, bei der Schätzung von Wahrscheinlichkeiten also die ***a-priori*-Wahrscheinlichkeiten** zu berücksichtigen sind.[186]

Ein in der Literatur bekanntes Beispiel – das „Taxi-Beispiel" – kann dies verdeutlichen: In einer Stadt gibt es zwei Taxi-Anbieter, die Taxen des einen Anbieters sind grün, die des anderen Anbieters blau. Firma Grün gehören 85 % der Taxen in der Stadt, Firma Blau 15 %. Ein Taxi ist in einen nächtlichen Unfall mit Fahrerflucht verwickelt und ein Zeuge sagt aus, es habe sich um ein blaues Taxi gehandelt. Dieser Zeuge ist in der Lage, in 80 % der Fälle die Fahrzeugfarbe zutreffend zu identifizieren, in 20 % der Fälle gelingt ihm dies nicht. Wie hoch ist nun die Wahrscheinlichkeit, dass das Unfallfahrzeug tatsächlich blau war? Die intuitive Antwort ist, dass die Aussage des Zeugen mit 80 %-iger Wahrscheinlichkeit korrekt ist. Tatsächlich liegt die Wahrscheinlichkeit einer richtigen Aussage aber nur bei etwa 41 %:[187] Die Fehleinschätzung der

---

183 Vgl. z.B. R v Clark, [2003 EWCA Crim 1020, 11 April 2003] – die Angeklagte *Sally Clark* wurde zunächst wegen Mordes an ihren zwei Säuglingen verurteilt, die beide am plötzlichen Kindstod verstarben, da der Gutachter die Wahrscheinlichkeit, zwei Kinder durch den plötzlichen Kindstod zu verlieren, falsch berechnet hatte.

184 *Schweizer,* Kognitive Täuschungen vor Gericht, S. 124; *Cleff,* Fn. 128, S. 34.

185 S.o. B.VI.

186 *Ihden,* Fn. 9, S. 149 (S. 153 ff. ausführlich zur Diskussion um die Anwendbarkeit des Bayes-Theorems in Gerichtsverfahren); *Schweizer,* Fn. 184, S. 124; *Effer-Uhe/Mohnert,* Psychologie für Juristen, S. 42 ff.; *Janßen,* Bayessche Netze in der Rechtsprechung, S. 9; *Bunzel/Marcoul,* Can racially unbiased police perpetuate long-run discrimination? In: Journal of Economic Behavior & Organization 68 (2008) S. 36 (37); s.o. B.VI.

187 Zur genauen mathematischen Herleitung vgl. *Handl/Kuhlenkasper,* Fn. 47, S. 217; *Janßen,* Fn. 186, S. 9 ff.

Wahrscheinlichkeit hat ihre Ursachen darin, dass die Information unberücksichtigt blieb, dass 85 % der Taxen in der Stadt grün und 15 % blau sind (***base rate neglect*** oder **Basisratenfehler**).[188]

Im Entscheidungsfindungsprozess eines Gerichts oder in Ermittlungsverfahren der Staatsanwaltschaften kann der Basisratenfehler etwa Auswirkungen auf die Einschätzung der Wahrscheinlichkeit einer Tatbegehung durch eine*n Beschuldigte*n haben. Deuten Indizien beispielsweise auf eine Tatbegehung durch einen männlichen Täter mit Migrationshintergrund hin, so darf bei der Einschätzung der Wahrscheinlichkeit der Tatbegehung durch den männlichen Beschuldigten mit Migrationshintergrund nicht nur die Verteilung des Merkmals „Migrationshintergrund" in der Gruppe der männlichen Beschuldigten bzw. Angeklagten, mit denen Richter*innen und Staatsanwält*innen in ihrer täglichen Praxis typischerweise zu tun haben, in den Blick genommen werden. Es muss vielmehr auch berücksichtigt werden, dass in der männlichen Bevölkerung das Verhältnis zwischen Männern mit und ohne Migrationshintergrund ein anderes ist. Richter*innen und Staatsanwält*innen mag also bei Vorliegen entsprechender Indizien die Strafbarkeit einer Person mit Migrationshintergrund fälschlicherweise in höherem Maße wahrscheinlich erscheinen, als es der Fall sein würde, wenn richtigerweise auf die gesamte männliche Bevölkerung abgestellt würde.

III. Kognitive Täuschungen bei Entscheider*innen

Neben den soeben ausgeführten Herausforderungen, die sich bei der Rezeption empirischer Studien und dem Umgang mit Wahrscheinlichkeiten in Entscheidungssituationen ergeben und die sich weitgehend durch Methodenkenntnis und ein Verständnis für die Zusammenhänge bewältigen lassen, sind Entscheider*innen – und damit eben auch Jurist*innen – regelmäßig einer Reihe von kognitiven Täuschungen ausgesetzt, die ihre Urteilsbildung und Entscheidungsprozesse unbewusst beeinflussen können. Zwar hat das Wissen um diese Effekte zunächst keine Auswirkungen auf die Entscheidung im Einzelfall, die Verhaltenseffekte schwächen sich (anders als es beispielsweise beim Basisratenfehler der Fall ist) nicht allein dadurch ab, dass sie bekannt sind. Auch Richter*innen als geübte und unabhängige Entscheider*innen unterliegen

---

188 *Glöckner/Dickert,* Base-Rate Respect by Intuition: Approximation Rational Choices in Base-rate Tasks with Multiple Cues, Preprints of the Max Planck Institute for Research on Collective Goods, No. 2008, 49.

ihnen in vergleichbarem Maße wie Nicht-Richter*innen.[189] Dennoch sollten wir sie kennen und uns bewusst machen sowie, soweit möglich, Strategien einsetzen, die die Gefahr dieser kognitiven Fehlleistungen reduzieren.[190]

Die Urteilsbildung kann beeinflusst werden durch den **Rückschaufehler** (***hindsight bias***)[191] und durch überzogenen Optimismus (***overconfidence***).[192] Sie bilden oftmals die Kehrseite von Heuristiken, derer wir uns bedienen, um in komplexen Situationen mit Hilfe einer „kognitiven Daumenregel" handlungsfähig zu sein, die aber das Risiko von Verzerrung und Fehlurteil bergen.[193] Der Rückschaufehler führt dazu, dass die Vorhersehbarkeit eines Ereignisses retrospektiv systematisch (deutlich) überschätzt wird. Menschen neigen dazu, in der Rückschau solche Ereignisse für wahrscheinlicher zu halten, die bereits eingetreten sind, als alternative Szenarien, deren Ausgang zwar im konkreten Fall hypothetisch, aber möglicherweise tatsächlich wahrscheinlicher ist.[194] Die Gefahr des Rückschaufehlers besteht also insbesondere dann, wenn ein Sachverhalt *ex post* betrachtet wird, jedoch *ex ante* zu beurteilen ist, etwa im Polizeirecht (Anscheinsgefahr) oder im Schadensrecht (Fahrlässigkeitsmaßstab).[195] Die nachträgliche Betrachtung einer polizeilichen Maßnahme und die Einschätzung, ob sie rassistisch diskriminierend war, kann also ebenso vom Rückschaufehler beeinflusst sein wie die Frage, ob in Fahrlässigkeitsfällen, in denen der Schaden bereits eingetreten ist, die erforderliche Sorgfalt eingehalten wurde. Empirisch nachgewiesen wurde der Rückschaufehler z.B. bei Jurist*innen, die ihre Fähigkeit überschätzten, beurteilen zu können, ob eine Gerichtsentscheidung berufungs- bzw. revisionsfest sei.[196]

Überzogener Optimismus führt dazu, dass wir uns selbst und unsere Fähigkeiten systematisch überschätzen.[197] Die Überschätzung kann sich z.B. in der falschen

---

189 *Guthrie/Rachlinsky/Wistrich,* Inside the Judicial Mind, 86 Cornell Law Review 777, 782 ff; *Dickert/Herbig/Glöckner/Gansen/Portack,* The More the Better? Effects of Training, Experience and Information Amount in Legal Judgments, Appl. Cogn. Psychol. 26 (2012), S. 223 ff.

190 Ausführlich zu Behavioral Economics im Antidiskriminierungskontext, Magen, § XXX.

191 *Schweizer,* Fn. 184, S. 209; *Beck,* Behavioral Economics, S. 69.

192 *Schweizer,* Fn. 184, S. 261; *Beck,* Fn. 191, S. 58; *Effert-Uhe/Mohnert,* Fn. 186, S. 49 ff.

193 *Englerth/Towfigh,* Fn. 68, § 8 Rn. 507; *Beck,* Fn. 191, S. 25 ff.

194 *Englerth/Towfigh,* Fn. 68, Rn. 512 ff; *Mohnert/Effert-Uhe,* Der Rückschaufehler, RECHTS|EMPIRIE, 28.6.2019, DOI: 10.25527/re.2019.08.

195 *Mohnert/Effert-Uhe,* Der Rückschaufehler, RECHTS|EMPIRIE, 28.6.2019, DOI: 10.25527/re.2019.08.

196 *Guthrie/Rachlinsky/Wistrich,* Fn. 189, S. 802; *Engel,* Fn. 176, S. 46; *Schweizer,* Fn. 184, S. 214 ff. mit weiteren Beispielen zu Studien.

197 *Englerth/Towfigh,* Fn. 68, § 8 Rn. 514.

Einschätzung von Wahrscheinlichkeiten äußern. So geben Heiratswillige trotz einer tatsächlichen Scheidungsrate in Deutschland von rund 30 % typischerweise die Wahrscheinlichkeit, dass die eigene Ehe scheitern werde, mit null Prozent an;[198] auch dürfte sich die weit überwiegende Zahl von Pkw-Fahrer*innen „überdurchschnittliche" Fahrfähigkeiten attestieren.

Ein anderes, mit überzogenem Optimismus verwandtes Phänomen ist der ***self-serving bias***.[199] Der *self-serving bias* führt dazu, dass wir Erfolge unseren eigenen Fähigkeiten zuschreiben, während Misserfolge als zufällig oder diskriminierend eingeordnet werden. Ein*e Bewerber*in wird typischerweise bei einem erfolgreichen Bewerbungsgespräch den Erfolg auf den guten und sicheren Auftritt im Gespräch zurückführen, bei einem erfolglosen hingegen die unfairen Fragen der Auswahlkommission hierfür verantwortlich machen. Gleichzeitig wird Diskriminierung gegenüber Gruppen, denen man sich selbst zugehörig fühlt, typischerweise überschätzt, während Diskriminierung gegenüber Gruppen, denen man sich nicht zugehörig fühlt, als nicht stark ausgeprägt empfunden.[200] So ist etwa eine Mehrheit der Afro-Amerikaner*innen der Auffassung, durch Polizei, vor Gerichten, in öffentlichen Schulen und im Gesundheitswesen schlechter behandelt zu werden als weiße Bürger*innen, die wiederum dieser Auffassung nur in weit geringerer Zahl zustimmen.[201] Dieser *bias* birgt das Risiko, dass Diskriminierungen von Verwaltung und Justiz nicht als solche erkannt und geahndet werden, insbesondere wenn Richter*innen, Staatsanwält*innen und Verwaltungsbeamt*innen nicht Teil der (potentiell) diskriminierten Gruppe sind.

Zu den Effekten, die unsere Entscheidungen beeinflussen, gehören ferner der **Ankereffekt** und das ***Framing***. Der Ankereffekt beschreibt das Phänomen, dass in einem Entscheidungs- oder Einschätzungsprozess Informationen die Entscheidung beeinflussen, die uninformativ (das heißt für die Entscheidung nicht unbedingt relevant) oder auch völlig willkürlich gesetzt sind.[202] Wir richten uns dann in Situationen, in

---

198 *Englerth/Towfigh,* Fn. 68, § 8 Rn. 514.

199 *Beck,* Fn. 191, S. 58.

200 *Heidhues/ Köszegi/Strack,* Overconfidence and prejudice, S. 2, Preprint, arXiv:1909.08497v1.

201 *Horowitz/Brown/Cox,* Race in America 2019, Pew Research Center, April 2019, https://www.pew-socialtrends.org/2019/04/09/race-in-america-2019/#majorities-of-black-and-white-adults-say-blacks-are-treated-less-fairly-than-whites-in-dealing-with-police-and-by-the-criminal-justice-system, zuletzt abgerufen am 28. April 2020.

202 *Kahneman/Tversky,* Judgment under Uncertainty: Heuristics and Biases, Science, Vol. 185 (1974). S. 1124 (1128)*; Nickolaus,* Ankereffekte im Strafprozess, S. 25 ff; *Glöckner/Englich,* When

denen wir uns für einen Zahlenwert entscheiden müssen, unbewusst an einem Zahlenwert aus, der bereits im Raume steht, und zwar unabhängig davon, ob ein Zusammenhang mit der Entscheidungssituation besteht oder nicht. Der Ankereffekt spielt also überall da eine Rolle, wo es um Preisverhandlungen geht, aber auch das Strafmaß oder die Höhe eines Schmerzensgeldanspruchs sind Situationen, in denen der Ankereffekt das Ergebnis beeinflussen wird.[203] Die Höhe der Schmerzensgeldforderung einer Anwält*in in einem Prozess hat für Gerichte also ebenso die Wirkung eines Ankers wie die Höhe eines in vergleichbaren vorangegangenen Prozessen zugesprochenen Schmerzensgeldes – die unabhängig vom Ankereffekt freilich auch aus normativen Gründen wie etwa der Rechtsanwendungsgleichheit und Rechtssicherheit ein Orientierungspunkt ist.[204] Nachgewiesen wurde dieser Effekt zum Beispiel anhand eines Experiments, bei dem zunächst für die Versuchspersonen ein Glücksrad mit Zahlen zwischen 0 und 100 gedreht wurde. Anschließend sollten sie schätzen, ob mehr oder weniger Prozent der afrikanischen Staaten Mitglied der Vereinten Nationen seien als die vom Glücksrad angezeigte Zahl. Sodann sollte die tatsächliche Zahl der Mitgliedsländer geschätzt werden.[205] Es zeigte sich, dass die Schätzungen der Versuchspersonen an der zufällig durch das Glücksrad angezeigten Zahl orientiert waren.[206]

Der als ***Framing*** bekannte Effekt wird auch **Darstellungseffekt** genannt und bezeichnet den Umstand, dass Entscheidungen von Menschen in identischen Situationen anders ausfallen, je nachdem, ob sich ihnen die Situation als Verlust oder Gewinn darstellt.[207] Wird eine Situation derart beschrieben, dass sich die Entscheider*innen zwischen zwei Optionen entscheiden müssen, die jeweils Verluste beschreiben, so wird überzufällig häufiger die risikoreichere Variante gewählt. Haben sie jedoch die Wahl zwischen zwei (mit den im Verlust-*Frame* beschriebenen identischen) Alternativen,

---

Relevance Matters – Anchoring Effects Can be Larger for Relevant Than for Irrelevant Anchors, Social Psychology 2015, Vol. 46, S. 4 ff.; ausführlich zum Ankereffekt, *Beck,* Fn. 191, S. 145 ff.

203 *Nickolaus,* Fn. 202, S. 47 ff; *Ihden,* Fn. 9, S. 57; *Chapman/Bornstein,* Applied Cognitive Psychology, 10 (1996), S. 519 ff.

204 *Arndt,* Fn. 163, S. 430.

205 *Kahneman/Tversky,* Fn. 202, S. 1128.

206 Vgl. ferner ausführlich mit weiteren Beispielen zum Ankereffekt *Jacowitz/Kahneman:* Measures of Anchoring in Estimation Tasks. Personality and Social Psychology Bulletin 21, 1995, S. 1161–1166.

207 *Kahneman/Tversky,* Rational Choice and the Framing of Decisions, The Journal of Business, Vol. 59 No. 4, S. 251.

die als Gewinn formuliert sind, so entscheiden sie sich überzufällig häufiger für die risikoärmere Variante. Nachgewiesen wurde dies im ***Asian-Disease***-Experiment[208], bei dem die in zwei Gruppen aufgeteilten Versuchspersonen jeweils mit zwei Szenarien konfrontiert wurden, die Pläne zur Rettung der Bevölkerung im Falle einer das Leben von 600 Menschen bedrohenden Seuche enthielten. Im Ergebnis waren die Szenarien identisch, sie unterschieden sich aber in der Darstellung. Die eine Gruppe hatte die Wahl zwischen Plan A, wonach mit Sicherheit 200 Menschen gerettet würden, und Plan B, wonach mit einer Wahrscheinlichkeit von 1/3 alle 600 bedrohten Personen, mit der Restwahrscheinlichkeit aber niemand gerettet würde. Die zweite Gruppe sollte zwischen Plan C, dessen Folge der sichere Tod von 400 Menschen war, und Plan D, wonach eine 1/3 Chance bestand, dass niemand stirbt, während mit einer 2/3 Wahrscheinlichkeit alle 600 Menschen sterben würden, wählen.

Der Unterschied bestand also darin, dass sich in der einen Gruppe die Szenarien als Gewinnoptionen darstellten (Plan A und B) und in der anderen Gruppe als Verlustszenarien (Plan C und D). Die Versuchspersonen, die mit dem negativen *Frame* konfrontiert waren, entschieden sich für die riskantere Variante, während die mit dem „Gewinn-*Frame*“ konfrontierten Versuchspersonen die risikoärmere Variante wählten.[209]

Im Kontext eines gerichtlichen Verfahrens führt der *Framing*-Effekt dazu, dass sich Kläger*innen, weil sie sich in einem Gewinn-Frame befinden, eher für den risikoärmeren Vergleich entscheiden, während die Beklagten eine Fortführung des Verfahrens bevorzugen, auch wenn dies für sie riskanter ist.[210]

D. Fazit

Das Antidiskriminierungsrecht wird immer dann besonders herausgefordert, wenn es um den Nachweis einer Diskriminierung geht. Gleichzeitig ist es aber darauf angewiesen, Diskriminierung aus dem Bereich der individuellen Betroffenheit herauszuholen und zum Gegenstand von Gerichtsverfahren, Gesetzgebungsvorhaben und Verwaltungshandeln zu machen. Häufig wird eine erlebte Diskriminierung als besondere Empfindlichkeit der diskriminierten Person oder als Einzelfall abgetan. Vor allem

---

208 *Kahneman/Tversky,* Fn. 207, S. 260.

209 *Englerth/Towfigh*, Fn. 68, § 8 Rn. 535.

210 *Schweizer,* Fn. 184, S. 90; *Rachlinsky,* Southern California Law Review, 1996, 113.

mittelbare Diskriminierungen lassen sich nur schwer erfassen und nachweisen. Von diskriminierenden Strukturen (auch unbewusst) profitierende Akteur*innen sehen häufig keinen Handlungsbedarf, während sich auf Seiten der Diskriminierten oft großer Leidensdruck aufgebaut hat. Die Debatten, die sich um Diskriminierungsfragen drehen sind daher häufig kontrovers und werden zuweilen hitzig geführt.

Mit der empirischen Forschung gewinnt das Antidiskriminierungsrecht ein mächtiges Werkzeug, das richtig genutzt einen essentiellen Beitrag zu diesen Diskursen leisten kann. Mit ihrem Fokus auf statistische Evidenz kann *lege artis* durchgeführte empirische Forschung wesentlich zu einer Objektivierung und Versachlichung der Debatte beitragen. Die Mechanismen und Strukturen struktureller Diskriminierung können identifiziert und Ausgangspunkt für gezielte Maßnahmen werden, diese Zustände zu verändern und Chancengleichheit herzustellen.