Statistik und Statistiken richtig verstehen

Was steht hinter typischen Missverständnissen von Konzepten der Statistik und bei der Interpretation von Statistiken?

Peter von der Lippe

November2013

Es gibt viele Bücher über Schwierigkeiten mit Statistiken (bzw. darüber, wie man von Statistiken getäuscht wird oder sich täuschen lässt). In ihnen werden Verständnisprobleme, mit denen offenbar viele Menschen bei Statistiken zu kämpfen haben zu "Lügen mit Statistik" oder "Lügen mit Zahlen" gemacht und beliebt ist es auch, diese Probleme damit zu erklären, dass die Denkweisen im Alltagsleben und in der Statistik oft sehr verschieden sind. Das wird mit zahlreichen Beispielen von sog. "fallacies" und "biases" illustriert, wobei oft Phänomene, die eng miteinander verwandt sind oder gar auf das Gleiche hinauslaufen unter verschiedenen Namen bekannt sind (z.B. Scheinkorrelation und Simpson's Paradoxon). In diesem Papier wird versucht, zu zeigen, was hinter solchen "biases" und "Paradoxien" steht und wie man entsprechende Fehlinterpretationen systematisieren und analysieren kann. Dabei versuchen wir mit einem gemäßigten Schwierigkeitsgrad[1] in puncto Mathematik und Statistik die Natur der häufig gemachten Fehler zu bestimmen und auch mögliche Zusammenhänge zwischen ihnen und den Denkgewohnheiten vieler Menschen aufzuzeigen.

[1] Einige Vorkenntnisse (z.B. aus einführenden Lehrveranstaltungen zur Statistik) wären sehr zu begrüßen.

1. Was heißt "verstehen" im Zusammenhang mit Statistik?

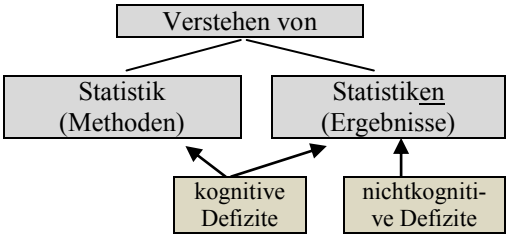
Nachdem Darrell Huff 1954 mit "How to Lie with Statistics" einen Bestseller gelandet hat sind unzählige ähnliche Büchern von Nachahmern erschienen, nicht nur in englischer[2] auch in deutscher Sprache und die Produktion entsprechender populärwissenschaftlicher Bücher geht auch heute noch, 60 Jahre nach Huff (1954) ungebremst weiter. Dies zeigt zumindest, dass viele ein Interesse an diesem Gegenstand haben, vielleicht auch weil sie alle ihre Not mit der Statistik haben.

Ähnlich verhält es sich mit anderen, thematisch eng verwandten Büchern, in denen versucht wird, unsere Schwierigkeiten mit Statistik mit entwicklungsbiologisch bedingten Defiziten unserer Denkstrukturen zu erklären. Danach sind es sozusagen die Lebensbedingungen der Steinzeit gewesen, auf die wir zwar seinerzeit gut eingestellt waren, die es uns aber jetzt, viele tausend Jahre später, so schwer machen, bei Statistiken nicht intuitiv naheliegenden Fehlschlüssen zu erliegen. Offenbar hatte das Verstehen und Interpretieren von Statistiken keine Selektionsvorteile beim survival of the fittest geboten, denn sonst wären wir inzwischen wohl schon besser gerüstet für den Umgang mit den vielen Statistiken, die uns täglich geboten werden.

Es geht bei Statistiken – und das ist im Kern die Botschaft der besagten Bücher – um eine ganz andere Art zu denken als die im Alltagsleben gewohnte, und das ist es auch, was es erklärt, dass so viele, auch intelligente Menschen bei der Interpretation von Statistiken zu falschen Überlegungen neigen und ihnen dann die richtige Interpretation nur so schwer einleuchtet.[3]

> In beiden Fällen (Schriften über sog. "Lügen" mit Statistik und über unsere atavistischen Denkstrukturen) hat man bei den vielen neuen Büchern auffallend oft das Gefühl, wieder Altbekanntes zu treffen. Es scheint, als werde alle paar Jahre ein Bestseller aus den bekannten Thesen und Beispielen früherer Bestsellern komponiert.[4]

Dass viele Verständnisschwierigkeiten haben, ist unbestritten. Aber worin bestehen sie und was heißt überhaupt "verstehen" im Zusammenhang mit Statistik? Es ist zu unterscheiden:



Für uns stehen hier Methoden der Statistik im Vordergrund, also – wie in Lehrbüchern – das Fach Statistik und weniger die Ergebnisse von Statistiken, und deren inhaltliche Interpretation.

In den erwähnten Schriften über sog. "Lügen" mit Statistik wird gerne der Umstand thematisiert, dass bestimmte Messkonzepte (z.B. Messung der Arbeitslosigkeit durch die Arbeitslosenquote oder der Wirtschaftsleistung durch das Inlandsprodukt) in der Statistik (z.B. in der amtlichen "Wirtschaftsstatistik") unbefriedigend sind. Dass es Konzepte (Begriffe) gibt, die "naturgemäß" schwer zu messen sind und man sich dabei mit Konventionen hilft, die kontrovers sind und dies auch immer sein werden, kann man der Statistik nicht anlasten. Es ist ziemlich unsinnig, hier von (indirekten) "Lügen" zu sprechen.

So etwas und die konzeptionellen Probleme der *Messung von Konstrukten*, wie z.B. auch von "Gerechtigkeit" oder "Glück", was aktuell in Mode ist, soll uns hier nicht beschäftigen. Es geht uns hier vielmehr nur um Schwierigkeiten des Verstehens von im engeren Sinne *statistisch-methodischen Inhalten*. Dies und Gründe für das häufige Missverstehen von Statistik ist auch Gegenstand einer sich offenbar allmählich unter dem Namen "statistical cognition" entwickelnden Teildisziplin innerhalb der Statistik.[5]

---

[2] Mit Campbell (1974), Kimble (1978) Jaffe/Spirer (1983), Wang (1993) und vor allem Robert Hooke (1983) habe ich nur einige einschlägige Bücher unten bei den Literaturangaben aufgeführt.

[3] Sicher ist es zumindest ein Denken, das zusammen mit der dazugehörigen Disziplin "Statistik" erst sehr spät (vor etwa 300 Jahren) aufkam.

[4] Plagiatsjäger müssten hier auf ihre Kosten kommen, aber die jagen ja bevorzugt Dissertationen von Politikern und nicht populärwissenschaftliche Schriften.

[5] Cumming definiert es als "the empirical study of how people understand, and misunderstand statistical concepts and presentations."

Man kann kognitive und nicht-kognitive Defizite beim Verstehen von Statistik unterscheiden. Zu den kognitiven gehören Defizite in puncto Mathematik, logisches Denken und "number sense".[6]

Es ist klar, dass Statistiken nicht richtig verstanden werden, wenn Defizite schon bei einfacher Mathematik auftreten, wenn man z.B. glaubt, dass sich nach einer Zunahme um 20% im letzten Jahr (t-1) und einer darauf folgenden Abnahme um 20% in diesem Jahr (t) per Saldo nichts verändert hat, während es in Wahrheit um eine Abnahme um 4% (vom Stand zu Beginn des Jahres t-1) geht.

Andererseits ist auch nicht nur die Mathematik betroffen, wenn es um das "Verstehen" geht. Um Statistik zu "verstehen", muss man nicht auch die entsprechenden Beweise (z.B. von Grenzwertsätzen) begriffen haben. Oft wäre es schon gut, wenn es nicht an der (gerade bei der Mathematik geforderten) Fähigkeit und Bereitschaft mangeln würde, sich ausreichend zu konzentrieren und sich die nötige Zeit zu nehmen, um sich ein Problem sorgfältig gedanklich zurechtzulegen.[7] Man kann auch bei der Interpretation von Statistiken nicht erwarten, dass man halb bei der Sache und "von jetzt auf gleich" alles durchblicken kann.

Zu den nichtkognitiven Fähigkeiten bzw. Defiziten gehört die – sich vielleicht seinerzeit unter Steinzeit-Bedingungen als vorteilhaft entwickelte – Neigung,

- zu vereinfachen (bzw. sich schnell mit einer plausiblen Lösung zufrieden zu geben) und als weniger wichtig Empfundenes ausblenden;
- selektiv Erwartetes wahrzunehmen, Bestätigung für bislang Geglaubtes zu suchen, aber Anderes zu ignorieren und schnell wieder zu vergessen;
- nach Ursachen zu suchen und dabei den Einfluss von Zufall zu unterschätzen;
- persönlichen Eindrücken und in Gruppen herrschenden Meinungen mehr zu glauben als "objektiven" Zahlen[8] und einer skeptischen Überprüfung (wenn diese überhaupt versucht wird) und
- Zahlen als Fakten zu nehmen, und sich nur wenig oder gar nicht in die Bedingungen hineinzudenken unter denen die Statistik diese Zahlen gewonnen hat.

Gerade der zuletzt genannte nichtkognitive Fehler hat sicher nichts mit zu wenig Mathematikkenntnissen zu tun. Wir bringen zwei Beispiele für zu wenig Phantasie und gesunden Menschenverstand, was das Zustandekommen der Zahlen einer Statistik und deren Aussagefähigkeit betrifft:

1. Ein in der Literatur oft zitiertes Beispiel war die Untersuchung der Schäden an britischen Flugzeugen nach Kampfeinsätzen im Zweiten Weltkrieg. Bei der Entscheidung, auf ein "extra armour plating on the places with no or few bullet and flak-holes" zu verzichten und nur die besonders beschädigten Stellen zu verstärken, wurde nicht daran gedacht, dass zwar der Rumpf oft deshalb nicht beschädigt war, weil gerade die hier besonders beschädigten Flugzeuge gar nicht erst vom Kampfeinsatz zurückgekommen sind.[9]

2. Ein anderes Beispiel für mangelnde Phantasie bei der Würdigung der Aussagefähigkeit einer Statistik habe ich selbst Mitte der 90er Jahre an der Universität Essen (damals noch nicht fusioniert mit der Universität Duisburg) erlebt:

Im Rahmen von Befragung von Studenten zur Qualität der Lehre, die seinerzeit in Mode kamen, wurde vom Rektorat ergänzend zu einer Tabelle extra (und besonders vorwurfsvoll) vermerkt, dass ein Student gesagt habe, ein Professor des Fachbereichs würde seit 40 Jahren

[6] Keith Devlin unterscheidet "number sense", ein angeborenes (und offenbar schon bei einigen Tieren vorhandenes) Gefühl für "numerocity" und "numerical ability" sowie die höher entwickelte "algorithmic ability" (die Fähigkeit Sequenzen von Operationen mit Zahlen durchzuführen). Anders als das Verstehen von statistische Analysen hat number sense durchaus Überlebensvorteile: Erkennen, ob angesichts der (möglichst schnell und nur grob geschätzten) Zahl der Gegner Angriff oder Flucht die bessere Option ist.

[7] Anders als bei der Statistik zweifelt man bei der Mathematik nicht daran, dass so etwas nötig ist.

[8] "We prefer stories to statistics" und der Mensch ist ein "story teller" (Thomas Kida). Das entspricht sehr stark dem, was ich (ohne von Kidas Schrift zu wissen) oft "impressionistische" Methode nannte.

[9] Man kann auch sagen, dass die untersuchte Gesamtheit wegen der survivor bias (mehr dazu später) gar nicht "repräsentativ" war.

genau dieselbe Vorlesung halten. Ich habe dann in einem Brief an das Rektorat gefragt,

- wie alt der Student ist, wenn er schon 40 Jahre lang in Essen studiert,[10]
- wie jemand es fertig bringt, sich 40 Jahre lang die gleiche Vorlesung anzuhören,
- und ich habe mich auch dazu bekannt, dass ich seit ca. 20 Jahren in meiner Vorlesung die gleiche Formel für die Dichtefunktion der Normalverteilung benutze, und das auch ganz ohne dabei ein schlechtes Gewissen zu haben.

Eigentlich hatte ich damit gerechnet, dass der Brief als humoristischer Beitrag eines Sonderlings in den Akten verschwinden würde, aber er wurde von der Rektorin, die wir damals an der Uni Essen hatten, wiederholt zitiert und überall sehr ernsthaft diskutiert, was wohl zeigt, dass einem solche naheliegenden Gedanken gar nicht gekommen sind.[11]

Nichtkognitive Fehler dürften nicht selten sein, gleichwohl liegt die Ursache für Fehlurteile bei der Interpretation von Statistiken wohl doch oft eher im kognitiven Bereich, insbesondere auf dem Gebiet der Logik (sind die Schlüsse, die ich aus den Zahlen ziehen korrekt?), weshalb wir deshalb auch im Abschnitt 2 mit solchen Fragen beginnen wollen.Sowohl in Abschn. 2, als auch in Abschn. 3 geht es um nicht ganz einfach zu verstehende Gegenstände aus der Wahrscheinlichkeitsrechnung. Man mag sich fragen: warum gerade dieser vielleicht eher abschreckende Anfang dieses Papiers, zumal doch auch die Lehrveranstaltungen in Statistik meist mit der Deskriptiven Statistik und nicht mit der Wahrscheinlichkeitsrechnung zu beginnen?

Dazu ist zu sagen, dass dieses Papier nicht ein kurz gefasstes Lehrbuch der Statistik sein soll, sondern nur das Verständnis einiger häufig benutzter Denkmuster in der Statistik behandeln soll,[12] bei denen jeweils Wahrscheinlichkeiten eine wichtige Rolle spielen.

Außerdem hatten wir in Sachen Statistik zwei historische Einschnitte, wovon der zweite wohl der wichtigere ist, nämlich

- der Beginn einer vorwiegend numerischen statt bis dato ausschließlich verbalen Beschreibung von Beobachtungen (der Beginn der Statistik als eine beschreibende Staatenkunde),[13] und
- die Nutzbarmachung der Wahrscheinlichkeitsrechnung für die Statistik, zunächst in Gestalt von Stichproben und dann – bis heute anhaltend und unvermindert erfolgreich – mit sog. stochastischen (d.h. Zufallsvariablen enthaltenden) "Modellen".

Der zweite Einschnitt ist für die Statistik nach modernem Verständnis bedeutsamer; er machte die Sache aber wohl auch "schwieriger", und das führt uns zu einem Thema, auf das hier, zu Beginn, schon vor der Behandlung der eigentlich statistischen Gegenstände, hinzuweisen nützlich sein mag:

So, wie man darüber streiten kann, wie uns heutzutage vielleicht noch Relikte aus der Steinzeit ein korrektes Verstehen von Statistiken verbauen, kann man auch darüber spekulieren, was unsere heutige Welt des Surfens im Internet, SMS Schreibens usw. zumindest auf lange Sicht bedeuten könnte für unsere Fähigkeit und Neigung, sich mit anspruchsvollen statistischen Methoden zu beschäftigen. Eine sehr pessimistische Sicht hierauf findet man bei Richard M. Restak, einem US Neurologen, der unter der Überschrift "Technology and the new brain"

[10] Es sollte klar sein, dass er eigentlich schon im Rentneralter sein müsste und somit nicht repräsentativ war (außerdem bestand die Uni Essen damals auch noch keine 40 Jahre). Damit war die zitierte Aussage bestenfalls ein Gerücht und wenig wert.

[11] Wie wenig Zahlen allein, ohne Berücksichtigung ihrer Entstehungsgeschichte wert sind zeigt auch die folgende im Buch von Stephen Campbell erwähnte Geschichte. Danach wollte Ende des 16. Jahrhunderts ein gewisser Herr Weirus, der für den Herzog von Kleve arbeitete festgestellt haben, dass es aktuell auf der Welt 7.405.926 Dämonen gäbe.

[12] Es geht uns hier mehr um "Hintergründe" von Methoden, was einen Einstieg mit der Wahrscheinlichkeitsrechnung und deren Dominieren über die Deskriptive Statistik (wie meist in US Lehrbüchern) rechtfertigen mag, während man nach deutscher Tradition eher mit der für die Praxis wichtigeren und wohl auch am Anfang leichter zu verstehenden Deskriptiven Statistik beginnt.

[13] In dem Papier "Statistik und Manipulation" bin ich auf diese seinerzeit (zur Zeit der Aufklärung im 18. Jahrhundert) durchaus revolutionäre Neuerung der Quantifizierung näher eingegangen.

schreibt: "I think we might lose the ability to analyze things with any depth and nuance" Er führt hierzu als Begründung (oder Illustration) an

- als Folge des alles dominierenden SMS Stils "…anything longer than a sentence or two may well go unread. Since there isn't tolerance under these circumstances for anything beyond raw information … the creativity … is diminished",
- als Folge von "our increasing exposure to and reliance on images" haben wir " a lessening of our brain's capacities for information analysis, critical thinking, imagination and reflection" und schließlich
- "Distraction is everywhere… Short attention spans are becoming the communication norm" dies und das Surfen "from topic to topic" bleibt nicht ohne schädliche Konsequenzen für "our brain's most powerful functions, concentration and focus."

Ohne "concentration and focus" kann niemand Statistik und die dahinterstehende Mathematik verstehen. Es ist mir durchaus bewusst, dass gerade die nächsten beiden Abschnitte in dieser Hinsicht sehr anspruchsvoll sind. Ich habe sie, auch auf die Gefahr des "going unread" hin gleichwohl an den Anfang gestellt, weil ich (aus meiner Lehrerfahrung heraus) der festen Überzeugung bin, dass an der Mühe des längeren, genauen Nachdenkens kein Weg vorbeiführt.[14]

2. Wahrscheinlichkeiten, logisches Denken und das Theorem von Bayes

Fehleinschätzungen bei Zahlen kann man oft nur vermeiden, wenn man sich die Zeit nimmt für einige einfache Zusammenhänge aus der Logik, Mengenlehre und Wahrscheinlichkeitsrechnung.

a) Logische Schlussweisen und bedingte Wahrscheinlichkeiten

Der folgende Schluss heißt *modus ponens*; er ist gültig und charakterisiert die *Deduktion*

1. wenn A dann B (das ist die Bedingung [Prämisse], eine als richtig erkannte konditionale Aussage)
2. da A gegeben ist
3. also gilt B

In 2 haben wir eine mit dem *Vorder*satz konforme Aussage. Der Schluss ist korrekt. Dagegen ist die folgende Schlussweise falsch[15]

1. wenn A dann B (wie bisher)
2. gegeben B (dem *Nach*satz konform)
3. also gilt A

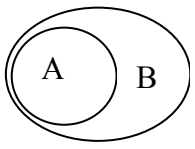
Taleb bringt in seinem Buch für die Prämisse (Satz1) das Beispiel "all people in the Smith family are tall". Man kann jetzt deduktiv schließen:

He belongs to the Smith family → he is tall

aber nicht reduktiv

He is tall → he belongs to the Smith family.

Wir können leicht sehen, dass sich hier richtig und falsch auch aus der Wahrscheinlichkeitsrechnung ergibt. Die Prämisse (1) besagt danach, dass A eine (echte) Teilmenge von B ist (symbolisiert mit $A \subset B$).



A ist eine Teilmenge von B (und damit eine hinreichende [aber nicht eine notwendige] Bedingung für B);
Es ist auch P(B) > P(A)

Daraus folgt für die Wahrscheinlichkeiten P(AB) = P(A) und P(B|A) = P(AB)/P(A) = 1 (wenn A [bedingt durch A] dann auch B also: wenn A gegeben, was symbolisiert ist mit (|A), dann auch B oder "if A then B")

Aber es gilt *nicht*: Wenn B gegeben ist, dann auch A; denn für die entsprechende bedingte Wahrscheinlichkeit P(A|B) erhält man P(A|B)

[14] Betrachtet man die Flut an immer beliebter werdenden deutschen und englischen Schriften unter dem Motto "Even you can understand Statistics" genauer, so sieht man, dass man oft zwar viele Seiten gelesen hat (nur um keine Formeln sehen zu müssen) aber die gleiche Zeit mit Nachdenken über die angeblich so unnötigen Formeln besser genutzt hätte. Ich halte gar nichts von derartigen Büchern.

[15] Bochenski, S. 101 nennt sie "Reduktion" und nach Taleb, S. 270 halten gut 70% der Bevölkerung diese falsche Schlussweise für richtig.

= P(AB)/P(B) = P(A)/P(B) < 1 und sie ist kleiner als 1, weil ja P(B) > P(A) und das deshalb, weil die Wahrscheinlichkeit für das Auftreten von "nicht A und B" (symbolisiert mit $P(\overline{A}B)$) größer ist als Null, denn

$$P(B) = P(AB) + P(\overline{A}B) = P(A) + P(\overline{A}B)$$

und deshalb ist auch

$$P(\overline{A}|B) = P(\overline{A}B)/P(B) = 1 - P(A)/P(B) > 0.$$

Das ist auch leicht zu sehen weil ja stets gilt

$$P(\overline{A}|B) + P(A|B) = 1.$$

Bei gleicher Bedingung B kann nur entweder A oder $\overline{A}$ (also "nicht A") eintreten, ein Drittes gibt es nicht.

In der Abbildung ist – wie gesagt – A eine Teilmenge von B ($A \subset B$) und damit nur *hinreichend* für B; denn B kann auch aus anderen Gründen als A auftreten. Entsprechend ist in dieser Situation B eine *notwenige* Bedingung für A, denn ohne B könnte es auch A nicht geben.

Wie nützlich es ist, eine verbale Aussage in mathematischer Notation noch einmal zu präzisieren zeigt auch das folgende Beispiel. Der britische Psychologe Stuart Sutherland erwähnte in seinem Buch "Irrationality"[16] die folgende Frage (quasi als Test für die Neigung zur Irrationalität)

> "Smoking increases the risk of lung cancer by a factor of ten and a fatal heart disease by a factor of two: do more smokers die of lung cancer than of heart disease?"

Es hat genug Leute gegeben, die auf diese Frage mit "yes" geantwortet haben, wenn sie nicht sogar geantwortet haben, dass doppelt (oder fünf Mal) so viele smoker an cancer als an heart attack sterben.

Dabei besteht kein Zusammenhang zwischen $P(C|S) = 10 \cdot P(C|\overline{S})$ und $P(H|S) = 2 \cdot P(H|\overline{S})$, und wir wissen auch nichts über die Wahrscheinlichkeit P(S) und über die Anzahl der Raucher.[17]

b) Lernen durch Erfahrung: das Bayessche Theorem

Oft wird in der Literatur das folgende, offenbar im "Alltags"verständnis von Wahrscheinlichkeiten als schwer empfundene Problem genannt: Jemand geht zum Arzt und lässt einen Test machen zur Erkennung einer Krankheit D, die man bei ihm vermutet. Der Test verläuft positiv (Ereignis T) und der Arzt sagt, dass der Test sehr zuverlässig ist weil er mit 95% positiv ausfällt.[18] Der Patient geht nach Hause und ist total beunruhigt, weil er meint, er habe mit 95% Wahrscheinlichkeit die Krankheit D. Zur Beruhigung kann man darauf verweisen,

- dass die 95% eine *bedingte* Wahrscheinlichkeit betreffen und P(T|D) = 0,95 und
- die 0,95 nur dann gelten, *wenn man die Krankheit D auch hat*, es ist aber nicht gesagt, dass der Patient sie hat (was vor allem dann nicht zutreffen muss, wenn diese Krankheit generell recht selten ist) und außerdem gilt, dass
- der Test auch fehlerhaft anzeigen kann, also auch $P(T|\overline{D}) > 0$ existiert; und da es keinen Grund gibt, dass P(T|D) und $P(T|\overline{D})$ in der Summe 1 sein müssen, denn die Bedingungen D und $\overline{D}$ sind ja unterschiedlich, können wir für die folgende Berechnung einmal annehmen, es sei $P(T|\overline{D}) = 0{,}1$.

Nicht P(T|D), sondern P(D|T) ist für den Patient relevant, d.h. die Wahrscheinlichkeit, D zu haben, wenn der Test positiv war, was er ja war (die Bedingung T ist also gegeben). Nach der Definition der bedingten Wahrscheinlichkeit gilt

$$(1) \quad P(D|T) = \frac{P(DT)}{P(T)} = \frac{P(T|D) \cdot P(D)}{P(T)}.$$

Diese Gleichung stellt das Theorem von Bayes[19] dar. Für die im Nenner erscheinende "totale" (unbedingte) Wahrscheinlichkeit P(T)

[16] zuerst publiziert 1976, später auch posthum (Sutherland starb 1998) veröffentlicht 2007 und 2013.

[17] Es gibt auch keinen Grund, dass P(C|S) + P(H|S) = 1 sein muss, denn ein Raucher kann auch weder an C noch an H sterben.

[18] Dabei betont er vielleicht nicht so sehr: *wenn* man die entsprechende Krankheit hat (bedingt durch D.

[19] Thomas Bayes (1701 – 1761), presbyterianischer Pfarrer in Turnbridge Wells, Kent.

gilt $P(T) = P(TD) + P(T\overline{D})$ und berücksichtigt man das, so erhält man

$$(2)\quad P(D|T) = \frac{P(T|D)\cdot P(D)}{P(T|D)\cdot P(D) + P(T|\overline{D})\cdot P(\overline{D})}\,.$$

Die Wahrscheinlichkeiten P(D) und $P(\overline{D})$ = 1-P(D) heißen a priori (vor der Erfahrung) Wahrscheinlichkeiten oder "prior probabilities" (oder einfach priors).

In ihnen kommen z.B. allgemeine oder aus früheren Studien gewonnene Kenntnisse über die Verbreitung der fraglichen Krankheit zum Ausdruck. Wenn danach z.B. die Krankheit relativ selten ist, etwa P(D) = 0,05, dann ist die für den Patienten eher interessante Wahrscheinlichkeit P(DT) = P(TD) = P(T|D)P(D), was mit 0,95·0,05 = 0,0475 doch schon viel besser aussieht als die 0,95 für P(T|D).

Wie man sieht, kommen in der Gleichung für das Bayessche Theorem drei Arten von Wahrscheinlichkeiten vor

1) die ***a priori Wahrscheinlichkeiten*** (oder Priors) P(D), auch incidence oder im deutsch "Prävalenz" genannt, und $P(\overline{D})$, wobei natürlich $P(\overline{D})$ = 1 - P(D), und
2) die ***Likelihoods*** P(T|D) (auch sensitivity genannt)und $P(\overline{T}|D)$ = 1 - P(T|D), ferner $P(T|\overline{D})$ und $P(\overline{T}|\overline{D})$;wie man sieht, gibt es *zwei* mögliche Fehldiagnosen, nämlich $P(\overline{T}|D) > 0$ und $P(T|\overline{D}) > 0$, das wären Fehler erster und Fehler zweiter Art wenn die Hypothese "Patient hat Krankheit D" zur Diskussion steht;[20] und
3) $P(D|T), P(\overline{D}|T)$ die ***a posteriori Wahrscheinlichkeiten*** (posteriors oder posterior probabilities).

Die a priori Wahrscheinlichkeiten werden auch "base rates" genannt und der Fehler, sich von 95% für P(T|D) unnötig beunruhigen zu lassen, weil man die geringe Wahrscheinlichkeit von P(D) = 0,05 ignoriert wird "***base rates fallacy***" genannt.[21] In diesem vorliegenden Papier zeigen wir viele Beispiele[22] dafür, dass eine Aussage über die Likelihood P(T|D) [symptoms T, given disease D] fälschlich als eine Aussage über die eigentlich interessierende a posteriori Wahrscheinlichkeit P(D|T) verstanden wird. Weil es beim Unterschied zwischen P(T|D) und P(D|T) vor allem auf P(D) ankommt, kann man das als "base rates fallacy" bezeichnen.

Die Likelihoods kann man auch ganz allgemein definieren als

Likelihood = P(evidence | hypothesis)

und die Hypothesen, die es zu beurteilen gilt, müssen nicht notwendig nur zwei sein, etwa D_1 = D und $D_2 = \overline{D}$ (oder Null- und Alternativhypothese H_0 und H_1), sondern es können $D_1, D_2, D_3, \ldots$ sein.

Man sieht an Gl. 2, dass P(T) ein gewogener (Gewichte P(D) und $P(\overline{D})$ = 1 – P(D)) Mittelwert ist aus $P(T|D)$ und $P(T|\overline{D})$, oder allgemein aus $P(T|D_1)$, $P(T|D_2)$.

Die Wahrscheinlichkeit für D_i wird durch die Erfahrung nach oben korrigiert $P(D_i|T) > P(D_i)$, wenn T unter der Bedingung D_i überdurchschnittlich wahrscheinlich ist, also $P(T|D_i) > P(T)$; entsprechend wird sie nach unten korrigiert wenn $P(T|D_i) < P(T)$.

Der Wert des Theorems liegt vor allem darin, dass es zeigt, wie und unter welchen Voraussetzungen man durch Erfahrung lernt, was so viel heißt wie, dass sich die a posteriori Wahrscheinlichkeiten von den a priori Wahrscheinlichkeiten unterscheiden.[23] Im Zahlenbeispiel ist

$$P(D|T) = 0{,}0475/0{,}1425 = 1/3 > P(D) = 0{,}05$$

und entsprechend $P(\overline{D}|T) = 2/3$ und P(T) = 0,0475 + 0,095 = 0,1425.

[20] Man beachte, dass hier eine Hypothese (z.B. "Patient hat D") eine (subjektive) Wahrscheinlichkeit (in Gestalt der Priors) hat (im Sinne eines Grades der Überzeugtheit von der Richtigkeit einer Hypothese), was vom Standpunkt der "klassischen" (frequentistischen) Testtheorie eine sonderbare Vorstellung ist; denn entweder hat der Patient D oder er hat nicht D. Eine subjektive Wahrscheinlichkeit, wonach z.B. mehr für D als für "Nicht D" spricht, hat hier keinen Platz. Auf die unterschiedliche Herangehensweise der klassischen und der Bayesschen Testtheorie kann hier nicht weiter eingegangen werden.

[21] Man kann jedoch von einem anderen, "nicht-bayesianischen" Standpunkt aus gesehen bezweifeln kann, ob hier überhaupt eine "fallacy" vorliegt.

[22] Siehe Abschn. 2d, aber auch den Anhang (S. 53f).

[23] In der Literatur wird "Lernen" auch als "updaten" von Wahrscheinlichkeiten definiert und nennt dies die diachronic (zeitlich) interpretation of Bayes.

Was man aus dieser Betrachtung mitnehmen sollte ist, dass es den wenigsten Menschen möglich sein dürfte, rein gefühlsmäßig, ganz ohne Mathematik und ihrer Symbolsprache auf eine Wahrscheinlichkeit von 1/3 (also 33%) zu kommen. Wir sind als Menschen, wie dies in unzähligen Beispielen gezeigt wurde, ausgesprochen schlecht darin, die Größe einer Wahrscheinlichkeit "aus dem hohlen Bauch" zu schätzen.[24]

c) Bedingungen für "Lernen" aus der Erfahrung nach dem Bayesschen Theorem

Interessant ist nun, was passiert, wenn man spezielle Annahmen über die beiden in die Rechnung eingehenden Wahrscheinlichkeiten macht:

	alle gleich	extreme Werte
Likelihoods	1	2
priors (a priori)	3	4

zu 1: Das *Diagnoseinstrument* differenziert nicht, ist also *unbrauchbar*: denn $P(T|D) = P(T|\overline{D}) \Rightarrow$ Posteriors = Priors; was nach unserer Definition bedeutet, nichts aus der Erfahrung gelernt zu haben;

zu 2: $P(T|D) = 1$ und auch[25] $P(T|\overline{D}) = 0$ (perfektes, unfehlbares Diagnoseinstrument) $\Rightarrow$ $P(D|T) = 1$, d.h. die Posteriors sind auch extrem, unabhängig davon wie groß die Priors waren,

zu 3: $P(D) = 0{,}5$ und damit auch $P(\overline{D}) = 0{,}5$ (Prinzip des mangelnden Grundes) $\Rightarrow$ die Posteriors werden allein durch die Likelihoods bestimmt, $P(D|T) = \frac{P(T|D)}{P(T|D) + P(T|\overline{D})}$.[26]

Das ist praktisch der Fall der klassischen Test- und Entscheidungstheorie, bei der allein die Erfahrung(en) – d.h. allein die Likelihoods – die Entscheidung bestimmen und keine [subjektiven, persönlichen] a priori Wahrscheinlichkeiten in sie eingearbeitet werden.[27]

Am Rande: Man könnte jetzt meinen, die klassische Betrachtungsweise ohne die a priori Wahrscheinlichkeit $P(D) = 0{,}05$ sei schlechter als die bayesianische und sie beunruhigt den Patienten unnötig mit der Wahrscheinlichkeit von $\pi = 0{,}905$ statt nur 1/3 für $P(D|T)$. Aber das wäre nur richtig, wenn man auch einen unbestritten korrekten Wert von $P(D)$ kennt, was bei entsprechenden Anwendungen des Theorems aber meist nicht der Fall ist.

zu 4: $P(D) = 1$ (*Dogmatismus*; egal was der Test zeigt, jeder hat die Krankheit, oder keiner hat sie $P(D) = 0$) $\Rightarrow$ die Posteriors sind jetzt auch genauso extrem, und zwar unabhängig vom Diagnoseinstrument (die Likelihoods interessieren nicht)

$$P(D|T) = 1 = \frac{P(T|D)\cdot 1}{P(T|D)\cdot 1 + P(T|\overline{D})\cdot 0},$$ wenn wir $P(D) = 1$ und entsprechend wäre $P(D|T) = 0$ bei $P(D) = 0$. Ergebnis (sehr plausibel):

> Es ist kein Lernen aus der Erfahrung möglich, wenn das Diagnoseinstrument nicht differenziert (Fall 1) und wenn (Fall 4) dogmatisch von extremen (0 und 1) a priori Wahrscheinlichkeiten ausgegangen wird.

Interessant ist es, Fall 4 und Fall 3 zu vergleichen:

- man kann *nicht* durch die Erfahrung lernen, wenn man *extreme a priori Wahrscheinlichkeiten* annimmt (Fall 4), und
- man kann auch *nur* durch die Erfahrung lernen, wenn es, wie im Fall 3 *keine a priori Wahrscheinlichkeiten* gibt (oder – was auf das Gleiche hinausläuft – sie alle gleich große sind).[28]

Das Theorem "funktioniert" – wie gesagt – nicht nur bei der Entscheidung zwischen D und $\overline{D}$ (oder D_1 und D_2), sondern auch,

---

[24] Wie sehr man (genauer: wohl die meisten Menschen) "gefühlsmäßig" danebenliegen kann, wird auch unten beim "Ziegenproblem" deutlich (Abschn.2e).

[25] was nicht aus $P(T|D) = 1$ folgt.

[26] Der Patient im obigen Beispiels, dem gesagt wurde $P(T|D) = 0{,}95$ müsste dann tatsächlich mit einer großen Wahrscheinlichkeit rechnen, die Krankheit zu haben, zwar nicht mit 0,95 wohl aber mit $0{,}95/(0{,}95 + 0{,}1) = 0{,}905$. Die Likelihoods addieren sich, wie gesagt, nicht zu 1. Wenn das Instrument völlig fehlerfrei wäre, dann wäre für den Patienten die Wahrscheinlichkeit auch nicht 0,95, sondern 1 (das wäre Fall 2).

[27] In gewisser Weise ist also die bayesianische Betrachtungsweise die allgemeinere.

[28] Fall 3 entspricht der in der Fußnote 20 erwähnten "klassischen" Testtheorie, bei der die mit einer Hypothese gemachte Aussage in der Realität ja entweder *zutrifft oder nicht zutrifft*, also keine subjektiven Einschätzungen der Glaubwürdigkeit einer Hypothese ins Spiel gebracht werden.

wenn über n Hypothesen zu entscheiden ist, etwa D_1, D_2, … D_n (i = 1,…,n). Dann ist

$$(3)\quad P(D_i|T) = \frac{P(T|D_i)\cdot P(D_i)}{\sum_i P(T|D_i)\cdot P(D_i)}.$$

Wenn verschiedene Hypothesen in Erwägung gezogen werden sollten, kann Gl. 3 auch ein Lernen in einem Sinne formalisieren, wie man es aus der Kriminalistik kennt: zu einer positiven Feststellung (D_1 ist der Täter) gelangen, indem man die nicht in Frage kommenden Möglichkeiten D_2, D_3, … ausschließt. Aus Gl. 2 und 3 folgt nämlich: wenn die Likelihood für die konkurrierende Hypothese kleiner wird (etwa $P(T|D_i) \to 0$; $i \neq 1$) dann wird die a posteriori Wahrscheinlichkeit der Hypothese, also $P(D_1|T)$ größer.

Diese Art, einen Täter zu überführen ist aber problematisch. Ein berühmter Fall, bei dem nämlich diese Logik pervertiert wurde war der Mordprozess Sally Clark.[29] Zwei ihrer Kinder starben (Ereignis T) und die Frage war, ob

- beide eines natürlichen Todes starben, und zwar an der sehr seltenen Krankheit (Sudden Infant Death Syndrome SIDS oder einfach S) (Hypothese S_1S_2), oder
- ermordet wurden durch die Mutter (Hypothese MM).

Das Gericht nahm nicht nur fälschlich[30] an, dass die Wahrscheinlichkeit, dass sowohl Kind 1 als auch Kind 2 an SIDS sterben $P(S_1S_2) = (P(S))^2$ sei, was verschwindend gering ist, da (nach allgemeiner Einschätzung) die Wahrscheinlichkeit an SIDS zu sterben nur $P(S) \approx 1/8543$ sei und (ein zweiter, noch gravierender Fehler) weil diese Wahrscheinlichkeit (eigentlich müsste es $P(T|S_1S_2)$ sein) so gering sei, kann die andere Hypothese, nämlich MM angenommen werden, also weil $P(S_1S_2|T)$ praktisch Null ist, muss P(MM|T) praktisch 1 sein. Andere Hypothesen (etwa MA Mord durch eine andere Person) wurden nicht in Betracht gezogen.[31]

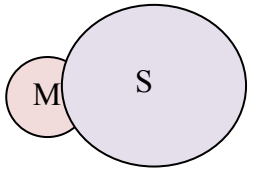
d) Vertauschung der Konditionalität und mehr zum Bayesschen Theorem[32]

Aus Gl. 1 folgt für $P(T|D) \to P(D|T)$

$$P(D|T) = P(T|D)\cdot\frac{P(D)}{P(T)}.$$

Im Zahlenbeispiel mit dem Test auf eine Krankheit war P(T|D) = 0,95 und P(D|T) = 1/3.

Die Vertauschung der Konditionalität soll beim Mordprozess des seinerzeit prominenten Sportlers O. J. Simpson eine Rolle gespielt haben.[33] Simpsons Anwälte argumentierten, dass es selten vorkommt, dass Männer, die ihre Frau schlagen (was im Falle von S als gegeben vorauszusetzen war) selten ihre Frau ermorden, so dass also $P(M|S)$ klein ist.



Männer, die ihre Frau schlagen (S) … ermorden (M):

P(S) > P(M) und P(M|S) ist klein weil P(MS) klein ist

Abgesehen davon, dass eine Wahrscheinlichkeitsaussage kein Beweis für die Erfüllung eines Tatbestands ist, kann[34] P(M|S) = P(S|M)P(M)/P(S) auch wegen einer (im Vergleich zu P(S)) geringen Wahrscheinlichkeit P(M) sehr klein sein.

Ebenso wenig wie

- man aus einem großen Wert für P(M|S) nicht darauf schließen kann, dass ein Mann, der seine Frau geschlagen hat auch ihr Mörder ist
- folgt aus einem kleinen Wert für P(M|S), dass er *nicht* ihr Mörder ist.

Wir haben auch keine Vorstellung, von der Aussagefähigkeit des Schlagens, weil wir das Verhältnis von Likelihoods $\frac{P(S|M)}{P(S|\overline{M})}$, also die "likelihood ratio" nicht kennen.

[29] Darüber mehr im Buch von Uri Bram, auf das ich mich hier stütze.
[30] weil S_1 und S_2 nicht notwendig unabhängig sind.
[31] Hinzu kommt, dass das Gericht hier generell mit sehr geringen Wahrscheinlichkeiten operierte. Auch P(T), das was allein beobachtete wurde, also das eigentliche "empirische Fundament" ist sehr klein. Analog zu Gl. 1 und 2 steht auch hier P(T) im Nenner bei der Bestimmung einer posteriori Wahrscheinlichkeit, die hier P(MM|T) ist, statt P(D|T) in Gl. 1 und 2.

[32] Wir bringen im Anhang, S. 54f weitere Beispiele.
[33] Dies ist wieder einen Fall, in dem Juristen, die sich sonst immer rühmen, quasi die Logik gepachtet zu haben, fehlerhaft mit Wahrscheinlichkeiten operieren.
[34] Wie noch zu zeigen ist, kann man auch grundsätzlich aus Wahrscheinlichkeiten keine Aussagen für konkrete Einzelfälle herleiten.

e) Das Ziegenproblem

Unter diesem Namen ist ein Spiel berühmt geworden,[35] das offenbar im US Fernsehen gespielt wurde. Ein Spieler S steht vor drei Türen und hinter einer der Türen ist ein Gewinn (z.B. ein Auto) verborgen und hinter den anderen zwei Türen jeweils eine Niete in Gestalt einer Ziege. S soll eine Tür i nennen, hinter der er den Gewinn vermutet. Dann öffnet der Moderator M eine Tür, wobei er zwei Restriktionen zu beachten hat:

R_1 M öffnet nicht die von S genannte Tür (dann wäre das Spiel auch schon gleich zu Ende) und

R_2 M darf auch nicht die Tür öffnen, hinter der sich der Gewinn verbirgt.

Angenommen S nennt Tür 2 und M öffnet darauf hin Tür 1 (was wir mit T_1 bezeichnen), hinter der eine Ziege steht. Die Frage ist nun, soll S seinen bisherigen Tipp beibehalten, also weiter auf Tür 2 tippen, oder soll er wechseln und jetzt auf Tür 3 tippen.

Vor der Antwort zunächst ein paar Symbole: G_i sei das Ereignis, dass der Gewinn hinter Tür i steht (bei i = 1, 2, 3). Es gilt (für die a priori Wahrscheinlichkeiten) $P(G_1) = P(G_2) = P(G_3) = 1/3$. Und mit $P(T_j|G_i)$ bezeichnen wir die Wahrscheinlichkeit, dass M die Tür j öffnet, wenn der Gewinn hinter Tür i ist.

Die häufigste Antwort auf die Frage "wechseln oder nicht" ist: da feststeht, dass der Gewinn nicht hinter Tür 1 steht, also $P(G_1) = 0$ ist, kann S bei seinem Tipp (Tür 2) bleiben, denn jetzt ist $P(G_2) = P(G_3) = ½$,[36] so dass es egal ist, ob S bei Tür 2 bleibt oder wechselt und jetzt Tür 3 nennt.

Diese offenbar intuitiv sehr naheliegende Antwort ist aber falsch; denn S stellt sich besser, wenn er wechselt und jetzt auf Tür 3 tippt. Sie ist falsch, weil sie die im Verhalten des M steckende Information nur unvollständig nutzt. S kann aus ihr mehr herausholen, um seine Gewinnchancen zu verbessern. Es geht auch nicht um $P(G_2)$ und $P(G_3)$, sondern um $P(G_2|T_1)$ und $P(G_3|T_1)$, weil ja T_1 eingetreten ist. Es wäre egal, ob S wechselt oder bei seinem Tipp bleibt, wenn $P(G_2|T_1)$ und $P(G_3|T_1)$ gleich groß wären. Tatsächlich ist aber $P(G_3|T_1) > P(G_2|T_1)$, und das deshalb, weil es wahrscheinlicher ist, dass M T_1 öffnet wenn der Gewinn hinter Tür 3 steht als wenn er hinter Tür 2 steht. Das liegt an den Restriktionen R_1 und R_2 und ist an den Likelihoods[37] zu sehen, für die gilt:

- wenn der Gewinn hinter Tür 2 steht:[38] $P(T_1|G_2) = P(T_3|G_2) = ½$; denn $P(T_2|G_2) = 0$ wegen der Restriktionen R_1 *und* R_2
- wenn aber der Gewinn hinter Tür 3 steht ist $P(T_2|G_3) = 0$ wegen R_1 und $P(T_3|G_3) = 0$ wegen R_2, so dass für M nur Tür 1 möglich ist, also $P(T_1|G_3) = 1$ ist.

Da $P(T_1|G_3) > P(T_1|G_2)$ ist auch $P(G_3|T_1) > P(G_2|T_1)$[39]. Wegen $P(G_i) = 1/3$ für i = 1, 2, 3

$$P(T_1) = \tfrac{1}{3}\left[P(T_1|G_1) + P(T_1|G_2) + P(T_1|G_3)\right] = ½$$

so dass nun

$$P(G_2|T_1) = \frac{P(T_1|G_2)P(G_2)}{\frac{1}{2}} = \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3} \text{ und}$$

$$P(G_3|T_1) = \frac{P(T_1|G_3)P(G_3)}{\frac{1}{2}} = \frac{1\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

[35] Es sind ganze Bücher darüber geschrieben worden.

[36] Diese Wahrscheinlichkeiten sind eigentlich nicht mehr wirklich a priori (also *vor* der Erfahrung) - und man bräuchte, streng genommen, hierfür andere Symbole - weil in ihnen ja die Information (Erfahrung) verarbeitet wurde, dass der Gewinn nicht hinter Tür steht. Die hinter dem Verhalten von M steckende Information wurde jedoch unvollständig verarbeitet, d.h. in ihr steckte mehr, als genutzt wurde. Die a priori Wahrscheinlichkeiten (vor Öffnen der Tür) waren und bleiben $P(G_1) = P(G_2) = P(G_3) = 1/3$.

[37] Die Bedingung (hinter dem senkrechten Strich) bei den Likelihoods ist G_1, G_2, bzw. G_3 was aber keine Zufallsvariable ist, sondern ein (uns allerdings nicht bekanntes Faktum. Eine Wahrscheinlichkeit kann man aber nur für eine Zufallsvariable definieren. Das ist der Grund, weshalb man hier mit Recht nicht den Begriff probability, sondern den (eigentlich synonymen) Begriff likelihood gewählt hat.

[38] Da der Gewinn definitiv nicht hinter Tür 1 steht, und S deshalb nicht auf Tür 1 tippt sind die $P(T_i|G_1)$ für alle i Null (die Bedingung trifft nicht zu).

[39] Das gilt hier wegen der Gleichheit der a priori Wahrscheinlichkeiten $P(G_1) = P(G_2) = P(G_3) = 1/3$ (das war oben der Fall 3).

Der Spieler S steht also besser da, wenn er wechselt. Das heißt nicht, dass er gewinnt, wenn er jetzt auf Tür 3 tippt, denn $P(G_3|T_1)$ ist ja nicht 1, aber seine Gewinnchance ist doppelt so groß, wenn er wechselt, als wenn er bei seinem Tipp (Tür 2) bleibt.

Man kann leicht nachrechnen, dass man wenn S auf Tür 2 tippt und M Tür 3 statt Tür 1 öffnet gilt $P(G_2|T_3) = 1/3$ und $P(G_1|T_3) = 2/3$, so dass es auch hier für S unsinnig wäre, auf seinem Tipp (T_2) zu beharren.

Die interessante Frage ist natürlich: woher kommt es, dass so viele Menschen hier die naheliegende, aber falsche Antwort "gehupft wie gesprungen" geben? Die Gründe sind bekannt und in der Literatur oft genug genannt worden:

1. Oft wird "gleichermaßen unbekannt" mit "gleichwahrscheinlich" verwechselt und daher auf $P(G_2) = P(G_3)$ getippt.[40] Manche vermuten auch eine natürliche Bevorzugung zu symmetrischen Lösungen, was vielleicht auch der Grund ist für
2. die Neigung, vorschnell Analogien zu bilden: G_2 und G_3 sieht so aus wie Zahl oder Wappen, die beiden Seiten einer Münze, und dort liegt ja auch Gleichwahrscheinlichkeit vor.
3. Die Unfähigkeit, die im Verhalten von M steckende Information zu nutzen liegt sicher auch daran, dass Betrachtungen nach Art des Bayesschen Theorems deutlich abstrakter und komplizierter sind als die naheliegende "einfache" (besser wohl: intuitive) "intuitive" Lösung "gehupft wie gesprungen", wonach S auch bei seinem ursprünglichen Tipp bleiben kann.[41]

Diese falsche Betrachtung, setzt implizit voraus, dass die likelihood ratio $\lambda_{32} = P(T_1|G_3)/P(T_1|G_2) = 1$ ist, tatsächlich ist sie aber 1/0,5 = 2, so wie auch $P(G_3|\ T_1)/P(G_2|\ T_1) = 2$ ist. Man sieht also, dass die falsche "intuitive" Lösung von gleichen Likelihoods ($\lambda_{32}=1$, statt $\lambda_{32} = 2$) ausgeht.

f) Möglichkeit und Wahrscheinlichkeit

Man kann i.d.R. nicht von der Anzahl der Möglichkeiten auf die Wahrscheinlichkeit schließen. Es ist bekannt, dass man nicht einfach wie folgt argumentieren kann: bei dem Spiel kann man gewinnen (G) oder nicht gewinnen ($\overline{G}$), also ist dann die Wahrscheinlichkeit, dass man gewinnt $P(G) = ½$. Das wäre nämlich nur dann richtig, wenn G und $\overline{G}$ gleichwahrscheinlich sind, was aber nicht der Fall sein muss.[42]

Man kann in konkreten Fällen durchaus darüber streiten, ob etwas gleichwahrscheinlich ist, wenn es nicht gerade um so einfache Dinge wie Münz- oder Würfelwurf geht. Aber selbst da waren früher den Menschen die Dinge noch nicht so klar wie uns heute.

Nach Leonard J. Savage hat der berühmte französische Mathematiker und Physiker J. B. d'Alembert (1717 - 1783) noch gedacht, die Wahrscheinlichkeit für mindestens einmal Kopf (K) beim Werfen mit zwei Münzen sei 2/3 und nicht ¾. Vom den vier Möglichkeiten (Z = Zahl) KK, KZ, ZK und ZZ hatte er KZ und ZK als eine Möglichkeit gezählt, so dass er insgesamt von drei, statt vier Möglichkeiten ausging. Es macht vielen auch Schwierigkeiten, zu sehen, dass beim Werfen von zwei Würfeln, einem blauen und einem roten, die gleiche Augenzahl (etwa 3-3 oder 5-5) jeweils *eine* der 36 Möglichkeiten ist, aber verschiedene Augenzahlen, (etwa 3-2 und 2-3) *zwei* Möglichkeiten sind.

Neben dem Rechnen mit bedingten Wahrscheinlichkeiten haben viele auch Schwierigkeiten mit der Frage: Was sagt mir eine Wahrscheinlichkeit darüber,

---

[40] Auf diese Gleichsetzung beruht ja auch die Annahme gleicher a priori Wahrscheinlichkeiten wenn nichts weiteres bekannt ist, man also im Sinne des Prinzips des mangelnden Grundes eine Annahme treffen muss; sie ist also nicht völlig abwegig.

[41] In der Literatur wird hier gerne Ockhams Rasiermesser (nach Wilhelm von Ockham 1288 – 1347) ins Spiel gebracht. Es bedeutet aber nur, dass die einfachere *Hypothese*, bzw. Erklärung (die mit weniger fragwürdigen [weiteren] *Annahmen* verbunden ist) die bessere ist, nicht, dass der einfachere Rechengang der bessere ist. Mit der mit *Annahmen* (nicht Rechenaufwand) am sparsamsten umgehenden Erklärung kann (und sollte) man quasi alle alternativen Erklärungen wegrasieren.

[42] Es mag paradox erscheinen, dass man wissen muss, ob etwas gleichwahrscheinlich ist, um zu wissen, wie wahrscheinlich es ist. Wir können hier auf die damit angesprochene Problematik des Wahrscheinlichkeitsbegriffs (objektiv vs. subjektiv [personalistic]) nicht weiter eingehen.

- ob etwas zutrifft oder nicht (also "richtig" oder "falsch" ist),[43] und
- ob ein Ereignis, z.B. eine Erkrankung bei mir eintritt, oder ob ich beim nächsten Wurf eine 6 würfeln werde oder nicht?

3. Wahrscheinlichkeit, Fakten und Prognosen

Es geht im Folgenden darum, das Besondere an Wahrscheinlichkeitsaussagen zu verstehen.[44] Sie sind keine Feststellungen über Fakten, mit ihnen kann nicht etwas verifiziert oder falsifiziert werden, und man kann auch nicht darauf bauen, dass einer Art von (gerechten) Ausgleich eintreten wird.

a) Wahrscheinlichkeit und Eintritt eines Ereignisses: die "gamblers' fallacy"

Sehr bekannt ist der als "gamblers' fallacy" (Fehlschluss der Glücksspieler) bekannte Fehler, anzunehmen, dass eine Roulettekugel mit einer größeren Wahrscheinlichkeit auf rot (R) fällt, wenn sie in einer entsprechend langen Reihe (etwa 10-mal) hintereinander immer auf schwarz (S) gefallen ist.

Angenommen, die Wahrscheinlichkeit, beim i-ten Wurf der Kugel auf S zu fallen sei $P(S_i)$.

Es ist richtig, dass die Wahrscheinlichkeit $P(S_1) \cdot P(S_2) \cdot \ldots \cdot P(S_{10})$ in jedem Fall gering ist, weil jedes $P(S_i) < 1$ ist. Aber was sagt das darüber aus, ob S oder R (rot) im elften Wurf eintritt?

Man muss sich klarmachen, dass eine Kugel nicht beseelt ist: "schwarz" und "rot" sagt ihr nichts, sie hat auch kein Gedächtnis, es wird ihr auch nach zehnmal schwarz nicht langweilig und sie hätte auch nach noch so viel "schwarz" kein Bedürfnis nach Veränderung. Es ist also nicht $P(S_{11}|S_1S_2\ldots S_{10}) < P(R_{11}|S_1S_2\ldots S_{10})$, wie der gambler annimmt, weil die Kugel eine Abwechslung bräuchte. Was die fallacy ausmacht, ist dass der gambler nicht die Unabhängigkeit der Würfe erkennt, d.h. dass er nicht sieht, dass die bedingte (durch den vorherigen Ablauf) und unbedingte Wahrscheinlichkeit für S (und so auch für R) immer gleich ist, z.B. $P(S_3|S_1S_2) = P(S_3|R_1S_2) = P(S)$ ist.

Diese Unabhängigkeit bei einem Zufallsvorgang dürfte beim Roulette gegeben sein. Es ist aber nicht unbedingt ein Fehlschluss nach Art der "gamblers' fallacy", anzunehmen, dass nach 10 Tagen Regen ein Tag mit Sonnenschein folgen muss (das Wettergeschehen von heute ist nicht unabhängig von dem, wie es gestern war) oder dass der Kurs einer Aktie fallen muss, wenn er eine längere Zeit gestiegen ist, weil die Kursbewegungen ja massenhaft Kauf- und Verkaufsaktionen auslösen können.

Es geht auch nicht um die falsche Denkweise des gamblers, wenn kein Zufall (oder nicht nur Zufall) im Spiel ist. Wenn ein Bewerber B_1 schon x mal bei einer Auftragsvergabe den Zuschlag bekommen hat, kann das auch einfach daran liegen, dass er (systematisch) besser ist als die Mitbewerber und dass die Auftragsvergabe keine Verlosung (wo der Zufall entscheidet) darstellt. Der Irrtum der Glücksspieler hat auch nichts damit zu tun,

- dass es in vielen Fällen Gründe gibt, weshalb "die Bäume nicht in den Himmel wachsen", eine Erscheinung, die auch unter dem Stichwort "regression to the mean" diskutiert wird; oder
- dass "Ausgleichstendenzen nach Art des Gesetzes der großen Zahl(en) auftreten.

Kennzeichnend für die gambler's fallacy ist die Verkennung des Charakters einer Wahrscheinlichkeitsaussage. So etwas liegt auch vor bei dem verbreiteten Missverständnis, dass die Ablehnung einer Hypothese bei einem Signifikanztest so etwas sei, wie das v.a. von Karl Popper geforderte Falsifizieren einer Hypothese, oder bei der Vorstellung, man könne mit der Statistik zwar keine Hypothesen verifizieren, wohl aber falsifizieren.[45] Das ist zum einen falsch

- weil wir mit einem solchen Test – wie noch in Abschn. 6c gezeigt wird – nicht zeigen, ob eine Hypothese *falsch* ist, sondern nur dass es (angesichts des Stichprobenbefunds) *unwahrscheinlich* ist, dass sie zutrifft und zum anderen weil

[43] Dazu mehr in Abschn. 6c.

[44] Auf ein Verkennen des Charakters einer "Wahrscheinlichkeitsaussage" beruht auch das übliche Verständnis von "Repräsentativität" (vgl. Abschn. 6a).

[45] Das ist auch bei Schuyler W. Huck erwähnt als ein Beispiel für seiner 52 statistischen "misconceptions".

- Wahrscheinlichkeitsaussagen von anderer Art sind als z.B. die Beobachtung, dass einmal ein Stein nicht gefallen ist (was dann eine Falsifikation des Gravitationsgesetzes wäre).[46]

Die Wahrscheinlichkeit betrifft– um in dem Bild zu bleiben – *nicht* das Fallen/Nichtfallen *eines Steins*, sondern *aller Steine* unter bestimmten, gleichbleibenden Bedingungen.

b) Die Bäume wachsen nicht in den Himmel: regression to the mean

Es gibt Prozesse, denen eine Art ausgleichende Gerechtigkeit oder Tendenz zum Mittelmaß innewohnt. Aus welchen Gründen auch immer (z.B. Übermut, wenn es sehr gut läuft, oder auch genetische Gründe, wie im folgenden "klassischen" Beispiel) kann $x_{t+1} < x_t$ sein, wenn x_t über dem Durchschnitt lag (also $x_t > \bar{x}$) und $x_{t+1} > x_t$ wenn $x_t < \bar{x}$). Nehmen wir an x_{t+1} hängt im Sinne einer einfachen linearen Regression von x_t ab (etwa der IQ des Sohns x_{t+1} vom IQ des Vaters x_t) gem.

(4) $x_{t+1} = \hat{\alpha} + \hat{\beta} x_t + u_t$ (mit u_t als "Störgröße"),

wobei $\hat{x}_{t+1} = \hat{\alpha} + \hat{\beta} x_t$ die geschätzte Regressionsgerade ist. Weil diese durch den Schwerpunkt mit den Koordinaten $\bar{x}_{t+1}$ und $\bar{x}_t$ geht, ist $\hat{\alpha} = \bar{x}_{t+1} - b\bar{x}_t$. Wir dürfen weiter davon ausgehen, dass der IQ in beiden Generationen (Vater und Sohn) im Mittel gleich ist (also $\bar{x}_{t+1} = \bar{x}_t = \bar{x}$). Man erhält dann

(4a) $x_{t+1} - x_t = (1-\hat{\beta})(\bar{x} - x_t) + u_t$.

Die "regression to the mean" entsteht, wenn $\hat{\beta} < 1$ ist (wäre $\hat{\beta} > 1$, wäre die Annahme $\bar{x}_{t+1} = \bar{x}_t$ nicht mehr haltbar) und wenn man (zur Vereinfachung) u_t vernachlässigt. Denn dann dürfte der Sohn einen höheren IQ haben als der Vater ($x_{t+1} - x_t > 0$) wenn der Vater einen unterdurchschnittlichen IQ hat ($\bar{x} - x_t > 0$). Und ein überdurchschnittlich intelligenter Vater ($\bar{x} - x_t < 0$) hätte einen Sohn mit einem geringeren IQ ($x_{t+1} < x_t$ also $x_{t+1} - x_t < 0$). Um das anschaulicher zu machen sei angenommen: wenn $\hat{\beta} = ½$ muss wegen $\bar{x}_{t+1} = \bar{x}_t = \bar{x} = 100$ gelten $\hat{\alpha} = 50$, und die Regressionsfunktion ist dann $\hat{x}_{t+1} = 50 + 0{,}5 \cdot x_t$

Vater $x_t < 100$ (unterdurchschnittlich)		Vater $x_t > 100$ (überdurchschnittlich)	
Vater x_t	Sohn x_{t+1}	Vater x_t	Sohn x_{t+1}
80	90	110	105
90	95	120	110

Die Regression zum Mittelwert hat nicht die Qualität einer (psychologischen) Regel- oder Gesetzmäßigkeit,[47] als welche sie gerne hingestellt wird oder eines mysteriösen Ausgleichs auf lange Sicht, der "irgendwie" mit den Gesetz der großen Zahlen (law of large numbers) zu begründen wäre. Bei diesem Gesetz geht es nicht um einzelne x-Werte, sondern um zusammenfassende Kennzahlen, (wie z.B. einen Mittelwert $\bar{x}$) einer Stichprobe und deren Verteilung bei immer größer werdenden Stichprobenumfang n. Schon im Falle des IQ gibt es ja auch so einen zwangsläufigen Ausgleich nicht, weil sonst ja die Streuung des IQ um den Mittelwert 100 immer geringer werden müsste. Wir beobachten auch weder eine Polarisierung in immer klüger werdende Familien einerseits und immer dümmer werdende Familien andererseits, noch eine generelle Anhebung oder Senkung des mittleren IQs. [48]

c) Wahrscheinlichkeit und Nichtvorhersagbarkeit eines Ereignisses

Das hinter der "gamblers' fallacy" stehende Problem ist die *unzulässige Vorhersage des Eintretens eines (einzelnen) zufälligen* Ereignisses aufgrund der Wahrscheinlichkeit eines vorangegangen Ablaufs. Denn es ist ja geradezu das Kennzeichen eines Zufallsvorgangs, dass man sein Ergebnis im Einzelfall nicht voraussagen kann, auch nicht im Lichte einer gerade gemachten Erfahrung.

Das ist ein Phänomen, das früher einige Philosophen als ein Paradoxon faszinierte: Zufall bedeutet, dass man gerade *nicht* etwas

[46] Es ist nicht nur so. dass hinter einer Testentscheidung eine Wahrscheinlichkeitsaussage steht, sondern für diese ist auch der Stichprobenumfang von Bedeutung. Wir kommen darauf in Abschn. 6 zurück.

[47] Wie etwa die im folgenden Abschnitt angesprochene Abneigung, etwas mit "Zufall" zu "erklären", und stattdessen immer zu versuchen, es auf konkrete Ursachen zurückzuführen.

[48] Es gibt genug Gründe dafür, allen voran der Umstand, dass der IQ des Sohnes nicht nur von dem des Vaters, sondern auch dem der Mutter oder der Großelterngeneration abhängt. Ein Rückgriff auf die Großelterngeneration war für Francis Galton (1822 – 1911) auch der Grund für die Namensgebung "Regression".

vorhersagen und berechnen kann, und trotzdem gibt es hier aufgrund der Wahrscheinlichkeits*rechnung* etwas zu *berechnen*.

Das ist aber dann nicht mehr paradox, wenn man sieht, dass es nicht das Gleiche ist, was hier berechnet wird, bzw. nicht berechnet werden kann. Einmal ist es *ein einzelne Ereignis* (die Roulettekugel fällt auf rot oder schwarz), wo es nichts zu rechnen gibt, und zum anderen ist es die *Gesamtheit aller möglicher Beobachtungen unter den gleichen Bedingungen* (was passiert, wenn man die Roulettekugel unendlich oft wirft und die Unabhängigkeit der Würfe garantiert ist?).

Es ist sogar so, dass man *nur deshalb* mit Wahrscheinlichkeiten rechnen kann, *weil* der Vorgang (beim Roulette) *allein* vom Zufall bestimmt wird. Es ist also der gleiche Grund, der dahinter steht, wenn man

- das Eintreten eines Ereignis E nicht vorhersagen kann, aber
- andererseits die Wahrscheinlichkeit P(E) von E sehr wohl berechnen kann.

Die Unzulässigkeit von einer Wahrscheinlichkeit auf das Eintreten eines Ereignisses zu schließen steht auch hinter dem bekannten Witz über Statistik: Die Wahrscheinlichkeit, dass an Bord eines Flugzeugs eine Bombe ist, sei P(B) = 1/1000. Einem Politiker ist das zu riskant und er sagt sich, dass es besser ist, selbst eine Bombe mitzunehmen, weil die Wahrscheinlichkeit für zwei Bomben nur $(1/1000)^2$, also 1 zu eine Million sei.[49]

d) Wie groß ist die mich betreffende Wahrscheinlichkeit?

Bei dieser Frage tut sich schon dadurch eine Schwierigkeit auf, dass die für den Einzelnen relevante Wahrscheinlichkeit immer nur eine bedingte Wahrscheinlichkeit sein kann, und es bei der Wahl der Bedingung viele Möglichkeiten gibt, wie das folgende, leider etwas längere Zitat von R. v. Mises zeigt[50]

> "In a sample of American women between the age of 35 and 50, 4 out of 100 develop breast cancer within a year. Does Mrs. Smith, a 49-year-old American woman, therefore have a 4% chance of getting breast cancer in the next year? There is no answer. Suppose that in a sample of women between the ages of 45 and 90 – a class to which Mrs. Smith also belongs – 11 out of 100 develop breast cancer in a year. Are Mrs. Smith's chances 4%, or are they 11%? Suppose that her mother had breast cancer, and 22 out of 100 women between 45 and 90 whose mother had the disease will develop it. Are her chances 4%, 11%, or 22%? … What group should we compare her with to figure out the "true" odds? You might think, the more specific the class, the better – but the more specific the class the smaller its size and the less reliable the frequency … In the limit, the only class that is truly comparable with Mrs. Smith is the class containing Mrs. Smith herself. But in a class of one, "relative frequency" makes no sense."

Mit diesem Dilemma muss man natürlich irgendwie umgehen, wenn man Mrs. Smith z.B. als Krankenversicherer versichern will. Man wird sie in eine "passende" *Gruppe* einordnen und muss damit leben, dass speziell in ihrem Fall der Tarif ein gutes oder ziemlich schlechtes Geschäft sein wird.

> Man beachte, dass wir bei unserer Überlegung davon ausgingen, dass das, was vorhergesagt werden soll, ganz oder zumindest im hohen Maße vom Zufall abhängt und dass wir uns nur auf Wahrscheinlichkeitsaussagen stützen können und Kausalbeziehungen nur unvollkommen kennen. Aber es gibt auch genug Dinge, die von bekannten Einflussfaktoren abhängen und bei denen damit durchaus fundierte Einschätzungen möglich sind, was mit einer konkreten Person passieren wird. Im extremen Fall kann man sogar ziemlich sicher sein, was passieren wird.
>
> Wenn z.B. Mrs. Smith ohne ausreichend Sprit mit einem Sportflugzeug unterwegs ist, kann man eine Aussage wagen, die sehr viel konkreter ist, denn es ist dann ziemlich sicher, dass Mrs. Smith über kurz oder lang in Schwierigkeiten kommen wird.

[49] Hierbei ist übrigens implizit gedacht, dass die Bombe B_1 an Bord ist wenn der Politiker mit seiner Bombe B_2 kommt, und dass beide Ereignisse unabhängig sind, also $P(B_2|B_1) = P(B)$ denn nur dann ist $P(B_2B_1) = (P(B))^2$. Mehr zu Witzen über Statistik im Anhang dieses Papiers.

[50] Richard von Mises (1883 – 1953) war ein bekannter österreichischer Mathematiker und Wahrscheinlichkeitstheoretiker; hier zitiert nach Pinker, S. 349

e) Extrem seltene Ereignisse, Aufmerksamkeit und Aberglaube

Es ist oft festgestellt worden, dass die Menschen dazu neigen, beim Eintritt von Ereignissen, die ihnen extrem unwahrscheinlich erscheinen (man bekommt einen Anruf von jemand, an den man gerade gedacht hatte, oder man trifft "zufällig" einen Bekannten, den man lange nicht mehr gesehen hat, an einem fernen Ort im Ausland usw.) an Gedankenübertragungen, Fügungen usw. zu glauben und den Gedanken, dass so etwas auch allein durch Zufall eintreten könnte weit von sich weisen.

Was dem entgegenzuhalten ist, ist die große Masse entsprechender Vorgänge (z.B. von Telefonanrufen) bei denen nichts Spektakuläres eingetreten ist und denen deshalb keine Aufmerksamkeit geschenkt wird. Berücksichtigt man dies, erscheint das Seltene gar nicht mehr so selten und mysteriös.

Es gilt also zu zeigen, dass auch Ereignisse mit einer geringen Wahrscheinlichkeit, also seltene Vorgänge durchaus häufig auftreten können (sehr wahrscheinlich werden), wenn nur die Zahl entsprechender Zufallsvorgänge nur groß genug ist.

Mit zwei Würfeln jeweils eine 6 zu würfeln hat nur eine Wahrscheinlichkeit von $(1/6)^2 = 1/36 = 0{,}0278$ also weniger als 3%. Wirft man aber n = 100-mal mit zwei Würfeln, so ist die Wahrscheinlichkeit, so etwas einmal zu erleben 17%: $\binom{100}{1}(1/36)^1(35/36)^{99} = 0{,}1708$. Es ist sogar wahrscheinlicher, dieses zweimal zu erleben, denn $\binom{100}{2}(1/36)^2(35/36)^{98} = 0{,}242$.

Im Mittel ist das in 100/36 = 2,78 Fällen zu erwarten.[51] Im gleichen Sinn ist die Anzahl der Beobachtungen relevant beim bekannten Geburtstagsproblem. Betrachtet man n = 2 Personen, so ist es recht selten, dass beide am gleichen Tag Geburtstag haben, aber unter n = 100 Personen zwei zu finden, die am gleichen Tag Geburtstag haben ist gar nicht so "unwahrscheinlich".

[51] Selbst dreimal hat noch eine Wahrscheinlichkeit von 22,55%.

Dass quasi auch ein blindes Huhn ein Korn trifft, wenn der Haufen der Körner nur groß genug ist, kann man sich auch in der Weise zunutze machen, dass man sich unter empirischen Studien (wenn es nur genügend viele sind) diejenige aussucht, die einem das für seine Interessen günstigste Ergebnis liefert. So verfahren u.a. die "miracle pill vendors". Bei sehr vielen Untersuchungen kann es schon einmal eine geben, bei der das Wundermittel tatsächlich zufällig gewirkt hat. Man erfährt nichts darüber, bei wie vielen Studien aber das objektiv ganz wertlose Mittel keine der angepriesenen Wirkungen hatte.

Unter "data mining" versteht man die Suche nach Zusammenhängen in einer großen (gemessen am Umfang N der Gesamtheit, bzw. n > N der Stichprobe) Datenmenge. Das geschieht i.d.R. auch ganz ohne von vornherein (a priori) bestimmte Hypothesen über Kausalzusammenhänge im Auge zu haben. Aus dem Gesagten folgt, dass es gerade wegen der Größe von N bzw. n nicht unwahrscheinlich ist, dass zwei Variablen X und Y nur zufällig miteinander korrelieren. Jetzt gezielt diese Fälle herauszusuchen und so den Eindruck zu erwecken, man habe neue bedeutsame Zusammenhänge entdeckt heißt "data degrading" und ist wieder eine der nicht seltenen Arten des Missbrauchs von Statistik, und wegen "big data" (dazu mehr in Abschn. 5b) kann man damit rechnen, in Zukunft noch mehr solche reinen Zufallsprodukte als neu gefundene "Gesetzmäßigkeiten" präsentiert zu bekommen.

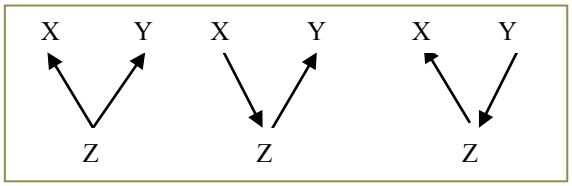
4. Kausalität, Korrelation und Zufall

In diesem Abschnitt wollen wir zeigen wie stark statistische Methoden von Überlegungen nach der Art von "Was (welche Variable X) beeinflusst eine bestimmte Variable Y?" dominiert werden und wie oft auch hinter Fehlinterpretationen von Statistiken falsche Vorstellungen gerade auf diesem Gebiet stehen.

a) Korrelation und Kausalität: warum und wie Kontrollgruppenexperimente?

Findet man eine betragsmäßig nicht unerhebliche Korrelation zwischen zwei Variablen X und Y (also $|r_{xy}| > 0$) bei der Betrachtung der Regressionsfunktion $y_i = \alpha + \beta x_i + u_{yi}$, (mit einer auf y einwirkenden Störgröße u_y) so kann das in puncto Kausalität bedeuten

- X und Y korrelieren nur zufällig miteinander (obgleich in der Stichprobe $|r_{xy}| > 0$ ist, gilt in der Grundgesamtheit $\rho_{xy} = 0$)
- X ist kausal für Y ($X \rightarrow Y$) oder umgekehrt $Y \rightarrow X$ (das kann man bei nur zwei Variablen nicht unterscheiden
- X und Y sind kausal nicht miteinander verbunden und sie korrelieren nur über eine gemeinsame Abhängigkeit von einer dritten Variable Z miteinander, wie es die folgende grafische Darstellung veranschaulichen soll. Es ist die typische Situation einer Scheinkorrelation (spurious correlation).



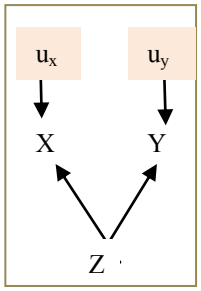
Die erste Möglichkeit kann man mit dem t-Test und der Prüfgröße $t = \hat{\beta} / \hat{\sigma}_{\hat{\beta}}$ oder (äquivalent) mit einem F-Test überprüfen.

Was die Richtung der Kausalität betrifft, so hat man kaum Möglichkeiten, wenn es nur um zwei Variablen X und Y geht. In allen drei Fällen der Scheinkorrelation ist damit zu rechnen, dass für die Korrelationskoeffizienten (in etwa) gilt $r_{xy} = r_{xz}r_{zy}$, so dass die partielle Korrelation verschwindet, also $r_{xy.z} = 0$ ist. Um nicht zu sehr ins Detail zu gehen genügt es, sich vorzustellen, dass bei einer partiellen Korrelation $r_{xy.z}$ der Einfluss von Z auf X und auf Y "ausgeschaltet" ist.[52] Bei kategorialen (≈ qualitativen) Merkmalen entspricht dem eine Betrachtung speziell für eine Teilgesamtheit bzgl. Z, wovon wir in den folgenden Abschnitten (insbes. in 7b) noch wiederholt Gebrauch machen werden.

Wenn etwa Z das Geschlecht ist heißt das, dass der Zusammenhang zwischen X und Y verschwindet, wenn man nur Daten für Männer oder nur Daten für Frauen betrachtet. Der Einfluss des Geschlechts ist nicht nachweisbar, wenn alle das gleiche Geschlecht haben. In der Statistik muss ein Einflussfaktor variieren. Eine Konstante, die bei allen Einheiten gleich groß ist, kann keine Ursache sein.

Man sieht, es gibt genügend Gründe für das Auftreten einer Korrelation zwischen zwei Variablen, auch dann, wenn keine (direkte) kausale Abhängigkeit zwischen ihnen besteht.

Wir müssen auch bedenken, dass wir es – anders als im Beispiel mit dem Motorstillstand bei Spritmangel – mit Variablen (als Ursachen und auch als Wirkungen), die auch einem Zufallseinfluss unterliegen zu tun haben.



In der letzten Abbildung wurden der Übersichtlichkeit halber die Störgrößen u_x und u_y nicht mit eingezeichnet. Berücksichtigt man diese, so ist leicht zu sehen, dass r_{xy} sogar 1 wäre, wenn die Varianzen der Störgrößen und die Kovarianz zwischen ihnen Null wären (also $cov(u_x u_y) = 0$).

Verschwindende Varianzen von u_x und u_y wäre auch genau der unrealistische Fall einer *deterministischen* Abhängigkeit von Z.

Im Folgenden gehen wir wieder von der Vorstellung einer Ursache X aus, also von

Das für Empiriker unüberwindbare Problem mit der "Kausalität" $X \rightarrow Y$ (im Unterschied zur Korrelation r_{XY}), ist die von den Philosophen verlangte *Notwendigkeit* der Verknüpfung (ohne X kein Y) statt nur eines (beobachtbaren) häufigen zusammen oder nacheinander Auftretens von X und Y. Abgesehen von streng determinierten technischen Prozessen (ohne Sprit läuft der Motor nicht und Mrs. Smith stürzt mit ihrem Flugzeug ab) ist "Notwendigkeit" i.d.R. nicht zu "beweisen" (zumindest nicht durch Erfahrung).

Der Nachweis einer ***Kausalität*** $X \rightarrow Y$ kann dann ***nur durch Ausschließen anderer möglicher systematischer Einflüsse*** erfolgen.

Das ist die Art, auf Kausalität zu schließen, wie sie bei einem Experiment praktiziert wird und wie sie allgemein als logisch korrekt und überzeugend empfunden wird. Kennzeichnend für ein Experiment ist, dass ein Einfluss[53] X systematisch variiert werden

[52] In der älteren Literatur sprach man auch von einer "bedingten" Korrelation und schrieb $r_{xy|z}$ statt $r_{xy.z}$ (so etwa J. Pfanzagl, Allgemeine Methodenlehre der Statistik, Bd. II, Berlin 1962, S. 260ff.).

[53] Genauer ein bekannter, "systematischer" Einfluss, der explizit berücksichtigt (manipuliert) werden kann.

kann und andere (störende) Einflüsse u_Y auf Y konstant gehalten (oder "kontrolliert") werden können, so dass eine Wirkung (bezüglich Y) allein auf X zurückgeführt werden kann. Die Kontrolle sonstiger (neben X existierender) Einflüsse ist es, was Experimentdaten von den "bloßen" Beobachtungsdaten unterscheidet. Wenn es nur X ist, was variiert wurde (ceteris paribus, d.h. bei Konstanz *aller* anderen möglichen Einflüsse), dann muss es auch X sein, was die Veränderung bei Y bewirkt hat.

Im einfachsten Fall eines Experiments, also im Fall von "one variable at a time" (OVAT) kann die isolierte Variation eines Faktors sichergestellt werden durch die Bildung von zwei Gruppen, einer Experimentgruppe E, die eine Behandlung (treatment) erfährt, z.B. Einnehmen eines Medikaments (was hier X sei), um dessen Wirksamkeit (Y) es bei dem Experiment geht, und einer Kontrollgruppe K, die diese Behandlung nicht erfährt. Die Kontrollgruppe soll sicherstellen, dass Y (die Wirkung) deshalb eintritt weil X (Medikament) "wirksam" war und nicht etwa nur durch Zufall, oder wegen des "Placeboeffekts". Es ist wichtig, dass K ansonsten (abgesehen vom "treatment") gleich strukturiert ist wie E, weil sonst auch Unterschiede zwischen K und E für eine scheinbare Wirksamkeit von X verantwortlich sein könnten (wenn z.B. in E "im Schnitt" jüngere, gesündere und sportlichere Personen sind als in K). Es darf also keinen systematischen, für Y relevanten Unterschied zwischen der Gruppe E und K geben. Man versucht das dadurch sicherzustellen, dass man den Zufall entscheiden lässt, wer in E und wer in K gelangt (was aber aus ethischen Gründen oft nicht möglich ist).[54]

Mit Randomisierung in Gestalt vom "random assignment" von Versuchspersonen (zu E und K) bzw. von unterschiedlichen Versuchen, die nacheinander durchgeführt werden (wenn die Reihenfolge einen Einfluss hat) ist bereits ein fundamentales Prinzip des "designs of Experiments" (DOE) – eine Teildisziplin der Statistik – angesprochen worden. Andere Prinzipien sind Blockbildung[55] und Wiederholung (replication), worauf hier jedoch nicht eingegangen werden kann. Wenn Erwartungen von Versuchspersonen (Vpn) eine Rolle spielen, sollte es den Vpn nicht bewusst werden ob sie in E oder in K sind. Die Vpn in K sollten also ein dem Medikament ähnliches Placebo erhalten (Blindversuch).[56]

Das soweit beschriebene *Kontrollgruppenexperiment* mit nur einem "Faktor", der in nur zwei Stufen (levels) vorliegt, Einnehmen oder Nichteinnehmen eines Medikaments, betrifft nur einen sehr speziellen Experimenttyp. Mehrere Faktoren F_i (i = 1,…, m) auch mit unterschiedlich vielen Faktorstufen n_i (oft auch mehr als nur zwei) so zu untersuchen wäre nicht rationell. Hierfür sind "*faktorielle*" (eigentlich gemeint: mehrfaktorielle) *Versuchspläne* nötig, auf die hier nicht eingegangen werden kann.

Was aus diesem Abschnitt vor allem folgt ist der Unterschied zwischen wissenschaftlicher und pseudowissenschaftlicher Vorgehensweise (z.B. durch "anecdotical evidence"):

> "Erklären" verlangt das *Ausschließen* aller Möglichkeiten *einer alternativen Erklärung*. Das Experiment zeigt, wie das geschieht und auch wie groß der dafür notwendige Aufwand ist.

Wir werden im Abschnitt 5a noch einmal auf die im engeren Sinne *statistische* Vorstellung von "erklären" zu sprechen kommen.

b) Stochastische Unabhängigkeit

Für Überlegungen, wie sich Ursache und Wirkung in beobachtbaren Größen widerspiegeln ist auch die regelmäßig missverstandene Vorstellung der stochastischen Unabhängigkeit wichtig.

Beim erwähnten Problem des Auftretens von Brustkrebs (C) bei Mrs. Smith ging es um die "richtige" bedingte Wahrscheinlichkeit für C, was deshalb ein Problem ist, weil die (bedingte) Wahrscheinlichkeit für C je nach Bedingung B unterschiedlich ist, also $P(C|B_1)$

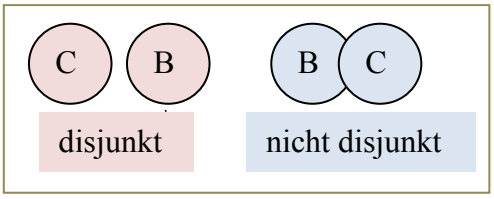
[54] Man denke an ein evtl. wirksames Mittel gegen Krebs. Wem will man es verweigern, wen also in K schicken? Das im Folgenden beschriebene Kontrollgruppenexperiment ist nur ein besonders einfaches Beispiel für ein Experiment. Abgesehen vom OVAT - Fall ist es eher unwirtschaftlich und auch bei gleichzeitiger Betrachtung von mehreren Einflüssen mit Wechselwirkungen (interactions) ineffizient.

[55] Einheiten in (möglichst homogene) Blöcke zusammenzufassen hat bei Experimenten eine ähnliche Bedeutung, wie die "Schichtung" bei der Ziehung einer Stichprobe.

[56] Bei einem Doppelblindversuch ist es bekanntlich auch dem Experimentator nicht bekannt, ob er es mit einer E- oder eine K-Vpn zu tun hat.

$\neq P(C|B_2)$ ist, und wir nicht wissen, ob B_1 oder B_2 die "maßgebende Bedingung speziell im Fall von Mrs. Smith ist. Wären dagegen das Brustkrebsrisiko und die Bedingung B (stochastisch) "unabhängig", würde gelten $P(C|B) = P(C|\overline{B}) = P(C)$, was nichts anderes als die Definition der "Unabhängigkeit" (von C und B, bzw. B und C)[57] darstellt.

Dieses Konzept wird von Studenten regelmäßig im Sinne von "unverträglich" (disjunkt) missverstanden, was aber P(CB) = 0 bedeuten würde. Hier, bei Unabhängigkeit ist aber P(CB) gerade nicht null. Denn wenn es null wäre, müsste ja auch $P(C|B) = P(CB)/P(B)$ null sein, was aber nicht zutrifft, denn bei Unabhängigkeit ist P(C|B) = P (C) > 0.



B und C sind nicht "unabhängig", weil sie nicht gemeinsam auftreten, sondern weil sie nicht häufiger P(BC) > P(B)P(C) und auch nicht seltener P(BC) < P(B)P(C) gemeinsam auftreten als es nach dem Zufall zu erwarten ist.[58]

Der Zusammenhang mit der Assoziation (bzw. Korrelation) wird deutlich mit der sog. Vierfeldertafel

	C	$\overline{C}$	Σ
B	a = P(BC)	b = $P(B\overline{C})$	P(B)
$\overline{B}$	c = $P(\overline{B}C)$	d = $P(\overline{B}\overline{C})$	$P(\overline{B})$
Σ	P(C)	$P(\overline{C})$	1

Wie man leicht sieht, ist (paarweise) Unabhängigkeit $P(B|C) = P(B|\overline{C}) = P(B)$ und damit auch $P(C|B) = P(C|\overline{B}) = P(C)$ oder auch $P(BC)P(\overline{B}\overline{C}) - P(B\overline{C})P(\overline{B}C) = ad - bc = 0$,[59] weil P(BC) = P(B)P(C), $P(\overline{B}\overline{C}) = P(\overline{B})P(\overline{C})$ usw. Die Vierfelderkorrelation ϕ

$$\phi = \frac{P(BC)P(\overline{B}\overline{C}) - P(B\overline{C})P(\overline{B}C)}{\sqrt{P(B)P(\overline{B})P(C)P(\overline{C})}}$$

verschwindet jetzt (ϕ = 0), weil Unabhängigkeit auch Unkorreliertheit impliziert,[60]

Die Vierfelderkorrelation ϕ (auch Phi-Koeffizient genannt) entspricht der "normalen" Produkt-Moment-Korrelation r_{xy} wenn die Variablen X und Y sog. 0-1 Variablen sind.

	Y = 1	Y = 0	Σ
X = 1	a	b	a+b
X = 0	c	d	c+d
Σ	a+c	b+d	n

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Das extreme Gegenteil von Unabhängigkeit wäre die vollkommene Abhängigkeit des Ereignisses C von B. Wenn (analog zu unserer Betrachtung in Abschn. 2a) B hinreichend wäre für C weil B eine Teilmenge von C ist ($B \subset C$), dann ist P(C|B) = 1, b = 0 und

$\phi = \sqrt{\frac{a}{1-a} \cdot \frac{d}{1-d}} \neq 1$. Auch wenn nur $C \subset B$ ist, wäre $\phi \neq 1$. Nur wenn B nicht nur hinreichend sondern auch notwendig für C ist, (also $B \subset C$ und $C \subset B$) ist nicht nur P(C|B) = 1 sondern auch P(B|C) = 1 weil dann ja P(BC) = P(B) = P(C) ist. Man hätte dann auch ϕ = +1 und die folgende Situation: [61]

	C	nicht C	Summe
B	P(BC)	0	P(B)
nicht B	0	$P(\overline{B}\overline{C})$	$P(\overline{B})$
Summe	P(C)	$P(\overline{C})$	1

Was auch Verständnisschwierigkeiten mit sich bringt ist die Vorstellung von "unabhängig" bei unabhängigen Stichproben, "Zügen" oder Wiederholungen eines Zufallsversuchs, die vor allem bei den für die Statistik so wichtigen Grenzwertsätzen entscheidend ist.

Am einfachsten ist es noch, sich "unabhängige Züge" aus einer Urne vorzustellen, wenn "mit

[57] Unabhängigkeit ist eine symmetrische Relation.
[58] Vielmehr gilt P(BC) = P(B)P(C).
[59] also ad – bc = 0.
[60] Unabhängigkeit ist eine strengere Forderung als Unkorreliertheit (das hängt damit zusammen, dass wir auch nichtlineare Zusammenhänge haben): wenn unabhängig, dann auch unkorreliert, aber die Umkehrung gilt nicht.
[61] Wäre B nur hinreichend für C also nur $B \subset C$, dann wäre $0 < \phi < +1$, so dass C nicht vollkommen abhängig von B ist (C kann ja auch eintreten bei "nicht B").

Zurücklegen" gezogen wird (weil sich dann das Mischungsverhältnis von schwarzen zu weißen Kugeln nicht ändert). Zieht man "ohne Zurücklegen" so ist die Wahrscheinlichkeit $P(S_2|S_1) < P(S_1)$, weil es weniger wahrscheinlich ist, wieder eine schwarze Kugel zu ziehen nachdem eine schwarze Kugel gezogen wurde (S_1), denn die erste ist ja nicht zurückgelegt worden.

Unabhängige Stichproben liegen vor, wenn die Zuordnung einer Einheit zu Experiment- oder zur Kontrollgruppe durch Los (Zufall) bestimmt wird. Versucht man den Effekt einer Behandlung (treatment) durch Messung an den gleichen Personen vor und nach der Behandlung festzustellen (repeated measures- oder within design), so liegen abhängige Beobachtungen (abhängige Stichproben) vor.[62]

Bei den sog. Grenzwertsätzen werden *unabhängig* identisch verteilte Zufallsvariablen vorausgesetzt. Die Zufallsvariable X (Augenzahl beim Würfeln) besitzt eine (diskrete) Gleichverteilung

$$f(x) = \begin{cases} 1/6 \text{ wenn } x = 1,2,...,6 \\ 0 \text{ sonst} \end{cases}$$

und die Zufallsvariablen X_1, X_2, … die Augenzahl beim ersten, zweiten … Wurf mit dem gleichen unveränderlichen Würfel sind unabhängig identisch (gleich)verteilt (independently identically distributed i.i.d.). Wir führen diesen Gedanken in Abschn. 6b weiter aus (wo es um Stichprobenverteilungen und um Grenzwertsätze geht).

In der ersten Euphorie über die Erkenntnis des Gesetzes der großen Zahl(en)[63] glaubte man ein Gericht würde umso eher die Wahrheit erkennen, je mehr Richter auf der Richterbank säßen. Dabei wurde vergessen, dass dieses Gesetz Unabhängigkeit voraussetzt, die gerade bei sehr vielen Richtern nicht gegeben sein dürfte (man schließt sich der zu erwartenden Mehrheit an, bildet "Koalitionen" etc.).

Abschließend noch ein Beispiel für ein krasses Missverständnis des Konzepts "stochastische Unabhängigkeit". Im Buch von Nate Silver wird berichtet, dass die US Rating Agenturen bei der bekannten "Subprime" Krise, die Auslöser der Finanzkrise 2008 war, das Kreditausfallrisiko (chance of default) massiv unterschätzt hatten, indem sie Unabhängigkeit statt vollständiger Abhängigkeit der Risiken unterstellten. In einem Poolmodell hat man fünf Kredite zusammengepackt und rechnete bei einem Ausfallrisiko von $P(B) = P(B_1) = 0{,}05$ mit nur $P(B_1B_2…B_5) = (P(B))^5 = 3{,}125 \cdot 10^{-7}$, also wie bei Unabhängigkeit, statt mit der 160.000 mal so großen richtigen Wahrscheinlichkeit von 5%

$$P(B_5|B_4B_3B_2B_1) \cdot P(B_4|B_3B_2B_1) \cdot P(B_3|B_2B_1) \cdot P(B_2|B_1) \cdot P(B_1) = 1 \cdot 1 \cdot 1 \cdot 1 \cdot 0{,}05 = 0{,}05,$$

wie es richtig gewesen wäre weil die Kredite vollkommen abhängig waren, sie "behave exactly alike. That is, either all five mortgages will default or none will".

So entstand die gefährliche Illusion von "almost no chance of defaulting when pooled together", während in Wahrheit das Risiko gerade durch das Zusammenpacken sehr groß war. Die Annahme der Unabhängigkeit macht nur Sinn, wenn der Häusermarkt im Prinzip gesund ist und die Zahlungsunfähigkeit eines Schuldners keinen Einfluss auf einen möglichen Kreditausfall eines anderen Schuldners hat, aber nicht bei einer sich zu einer Blase entwickelnden Überinvestition wenn "there is one common factor that ties the fate of these home owners together" (N. Silver). Der Fehler, zu übersehen, dass evtl. etwas quasi schicksalshaft gemeinsam variiert, ist auch das Thema im folgenden Abschnitt.

c) Die Störgröße bei einer Regressionsfunktion (Endogenitäts-Fehler)

Die folgende, leider etwas sehr "statistisch-technisch" ("formal") anmutende Überlegung betrifft einen nicht selten vorkommenden Mangel in empirischen Untersuchungen. Alle etwas anspruchsvolleren Untersuchungen dieser Art arbeiten mit "Modellen", mit denen bestimmte Größen geschätzt werden. Damit die geschätzten Modellparameter bestimmten Gütekriterien[64] genügen – also wissenschaftlich "brauchbar" sind –

[62] Man kommt mit weniger Versuchspersonen (Vpn) aus, aber in Messungen an der gleichen Vp können sich auch "Interferenzen" auswirken.

[63] Man kann dieses law of large numbers als einen speziellen Grenzwertsatz (diese Sätze müssten eigentlich eher Grenzverteilungssätze heißen) auffassen.

[64] Gütekriterien, wie z.B. Erwartungstreue und Konsistenz werden in Abschn. 6b kurz erwähnt. Aus Platzgründen kann aber hierauf in diesem Papier nicht weiter eingegangen werden. Für Nichtstatistiker ist es oft schwer zu verstehen, warum die Nichterfüllung von solchen Eigenschaften einer Schätzung eine so diskussionswürdige Angelegenheit ist.

müssen bestimmte Modellvoraussetzungen erfüllt sein und wir wollen hier auf eine dieser Voraussetzungen näher eingehen, weil ihre (nicht seltene) Verletzung in der hier diskutierten Literatur gerne thematisiert wird.[65]

Wenn wir es mit einem Experiment zu tun haben, in dem die vermutete Ursachengröße x willentlich nach Belieben isoliert verändert werden kann (d.h. als "exogen" betrachtet werden kann und nicht als Zufallsvariable), sollte die Störgröße (der "error term") u_i in der Regressionsfunktion $y_i = \alpha + \beta x_i + u_i$ im Mittel null sein. Sie ist dann auch quasi automatisch mit dem Regressor x nicht korreliert. Im Experiment ist zwar x keine Zufallsvariable, wohl aber U, und mit der erwähnten "Randomisierung" bei einem Experiment verfolgt man das Ziel, die Geltung der Annahme $E(U_i) = 0$ sicherzustellen. Dies ist eine *Annahme* über die *Grundgesamtheit*, die nicht notwendig zutreffen muss und davon zu unterscheiden ist, dass *in der Stichprobe* mit ihren, mit der Methode der kleinsten Quadrate geschätzten Parametern $\hat{\alpha}$ und $\hat{\beta}$ (für α und β), die *geschätzte*n Störgrößen $\hat{u}_i$ notwendig stets im Mittel null sind.[66]

Bei Beobachtungsdaten, die in den Sozialwissenschaften die Regel sind, und bei denen weder x aktiv manipuliert werden kann, noch sonstige Einflüsse auf y "kontrolliert" (konstant gehalten) werden können, entspricht dem die Annahme, dass die Störgröße u nicht mit der (jetzt auch zufällig schwankenden) Variable x korreliert ist (sein soll).[67]

Im Falle einer Korreliertheit sagt man, dass x nicht (mehr) exogen ist und man einen Endogenitäts-Fehler (endogeneity bias) hat, z.B. wegen

- *Falscher Kausalannahmen*, dass man $X \rightarrow Y$ für gegeben hält, während tatsächlich $Y \rightarrow X$ gilt (nicht erkannte Interdependenz)

  Interdependenz ist etwas, was es in Experimentsituationen nicht gibt (aber bei Beobachtungsdaten nie völlig auszuschließen ist). Wenn man z.B. im Experiment die Düngemittelmenge X vergrößert und einen höheren Ernteertrag Y feststellt steht natürlich nie umgekehrt $X \leftarrow Y$ zur Diskussion, weil schon rein physisch der Ernteertrag nicht für mehr oder weniger Düngemittel sorgen kann.

  Eine "simultaneous causality", also Interdependenz $X \leftrightarrow Y$ (Beispiel Herzkatheter X und Lebensverlängerung Y: es gilt nicht nur $X \rightarrow Y$, sondern auch $X \leftarrow Y$ [nur wenn lebensverlängernde Maßnahmen nötig sind, bekommt man X].[68]

- *Systematischer Einflüsse auf die Störgröße* verbergen (etwa von nicht explizit berücksichtigten Regressoren X_2; X_3, … [omitted variables], so dass z.B. neben $X_1 \rightarrow Y$ auch $X_2 \rightarrow Y$ gilt) und man damit α mit $\hat{\alpha}$ systematisch zu hoch oder zu niedrig schätzt,[69]

  Ein Fall einer omitted variable liegt vor, wenn das richtige Modell lautet $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, man aber $y_i = \alpha + \beta x_{1i} + v_i$ geschätzt hat (es ist klar, dass dann $E(v_i) = E(\beta_2 x_{2i} + u_i) \neq 0$ ist).

- Auch die später (Abschn. 6a) behandelte "sample selection bias" oder Messfehler in den Variablen können hier genannt werden.

Endogenitätsfehler beeinträchtigen nicht unerheblich die Schätzqualität,[70] was jedoch mehr eine Sache der statistischen Theorie ist.

---

[65] Uri Bram behandelt in seinem Buch "Thinking Statistically" im Prinzip nur drei Themen: Sampling Bias (bei uns unten im Abschn. 6a), das Bayesche Theorem und die in diesem Abschnitt behandelte, ihm offenbar sehr wichtige Endogenitäts Bias.

[66] Das gilt wegen der Methode der kleinsten Quadrate. $\wedge$ bezeichnet einen Schätzwert. Es ist anfänglich für viele Studenten schwer, *Annahmen* (*Hypothesen*) über die unbekannte Grundgesamtheit und *Fakten*, die die bekannte (weil "gezogene") Stichprobe betreffen, zu unterscheiden

[67] Nicht nur die U_i (und damit auch die Y_i) sind jetzt Zufallsvariablen, sondern auch die X_i.

[68] Man könnte hier auch von einer omitted variable "Gesundheitszustand" sprechen, die in der Störgröße steckt und eine Korrelation mit X erzeugt. Ein anderes Beispiel (von U. Bram): Wechsel der Versicherung (X) und Gewinn dabei (Y) in Gestalt niedrigerer Prämien. Es mag $X \rightarrow Y$ gelten, aber es gilt auch $Y \rightarrow X$, denn nur solche Leute wechseln, die sich davon einen Gewinn versprechen. X ist also endogen.

[69] und so auch y bei Prognosen mit dem Modell konsequent über- oder unterschätzt.

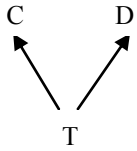
[70] Die Parameterschätzung ist nicht erwartungstreu und nicht konsistent. Auch bei Vergrößerung des

Für alle praktischen Betrachtungen dürfte es interessanter sein, die offensichtlich nicht immer klar zu trennenden Fälle eines Endogenitätsfehlers klar zu unterscheiden.

Als Beispiel für eine nicht berücksichtigte Variable X_2 (Prüfungsfach) nennt Uri Bram den Zusammenhang zwischen Fleiß (effort) X_1 und Studienerfolg Y.[71] Er kann u.U. deshalb nicht – wie erwartet – positiv sein, oder mit den Daten nicht nachgewiesen werden, weil die Studenten das Fach (X_2) frei wählen können und viele ein einfaches Fach wählen, bei dem sie auch bei geringem Fleiß einen großen Erfolg haben. Die Koeffizienten in der Regression von Y auf $X_1 = X$ wären also verzerrt (nicht erwartungstreu) weil X_2 über U aufY einwirkt.

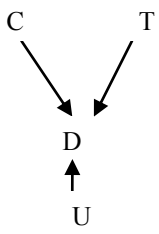
Man kann auch ein Feedback der zu erklärenden Variable Y auf den Regressor X_2 vermuten: *weil* man eine gute Note anstrebt (Y) wählt man ein leichtes Fach. Es ist das sich am Erfolg Y orientierte (Wahl-) Verhalten, das ein feedbak zwischen Y und den kausal für Y verantwortlichen Aktivitäten X erzeugt.

Es ist auch grundsätzlich nicht einfach, das Endogenitätsproblem von dem der Scheinkorrelation zu unterscheiden. Uri Bram bringt das Beispiel der Korrelation zwischen ice cream sales (C) und drowning (D) und deutet das als omitted variable Problem (Nichtberücksichtigung der Variable Temperatur T). Wir würden hierin eher einen Fall von Scheinkorrelation sehen, ganz im Sinne der folgenden Graphik:



Im Sommer (T hoch) wird gerne Speiseeis gekauft (C) und wegen der Hitze auch in Badeseen gebadet, so dass es häufiger zum Ertrinken (D) kommen kann.

im Unterschied zu einem omitted variable Problem in der allerdings ziemlich unsinnigen Vorstellung der Regression $d_i = \alpha + \beta c_i + u_i$,



Man "erklärt" das Ertrinken (D) durch großen Eiskonsum C (abwegig!!) und hat dabei die Variable T "vergessen" (omitted), bzw. diese implizit in der Störgröße "versteckt", so dass U mit C korreliert.

Das Interessante ist hier auch weniger, dass man β (und auch α) in der Regression (die ohnehin unsinnig ist) von D auf C nicht gut schätzen kann, als die gar nicht kausal zu erklärende Korrelation zwischen C und D (also $|r_{CD}| > 0$).

Warum Endogenität ein Problem ist, mag noch deutlicher werden, wenn man ein simultanes (ökonometrisches Mehrgleichungs)modell betrachtet, wo man dann auch von "simultaneous equation bias" spricht. Wie so oft erscheinen gleiche oder zumindest ähnliche Probleme unter verschiedenen Namen. Vgl. hierzu mehr Hinweise im Anhang (S. 52 unten).

d) Modelle und Modellbausteine (kontextabhängige Schätzwerte)

Bleiben wir noch etwas bei der multiplen Regression, etwa bei einem Modell mit zwei Regressoren (Einflussgrößen)[72] X_1 und X_2

$$y_i = \beta_{y.12} + \beta_{y1.2}x_{1i} + \beta_{y2.1}x_{2i} + u_i \quad (i = 1,\ldots,n).$$

Die Konstanten β (weil konstant ohne Subskript i) heißen partielle (oder auch multiple) Regressionskoeffizienten [73] (was bisher α war ist jetzt $\beta_{y.12}$ mit den beiden Regressoren hinter dem Punkt). Viele sind erstaunt, dass sich der Schätzwert $\hat{\beta}_{y1.2}$ (im Folgenden einfach $b_{y1.2}$) verändert wenn man eine andere als die bisherige Variable als Variable X_2 betrachtet oder von $b_{y1.23}$ verschieden ist in einem Modell mit zusätzlich X_3). Das liegt daran, dass die β-Koeffizienten aufgrund eines Gleichungssystems geschätzt werden und rechnerisch alle Koeffizienten miteinander zusammenhängen. Aber was das konkret bedeutet ist vielen nicht ganz klar.

Ich habe dies oft bei Diskussionen mit Nichtstatistikern im Zusammenhang mit Mietspiegeln erlebt: hier ist Y die zu schätzende Quad-

Stichprobenumfangs nähert man sich nicht dem wahren Wert von α und β.

[71] Anders als bei der Scheinkorrelation dürfte hier X_1 und Y auch wegen einer Kausalität $X_1 \rightarrow Y$ korrelieren und nicht nur indirekt über eine dritte Variabel X_2.

[72] oder "unabhängige" (erklärende) Variablen, die man meist mit x bezeichnet. "Multipel" heißt $k \geq 2$ Regressoren. Das Modell ist jedoch nicht "multivariat". Multivariat heißt, dass wir auch mehrere y-Variablen (abhängige Variablen) haben.

[73] Partieller und multipler Regressionskoeffizient ist synonym, aber partielle und multiple Korrelationskoeffizienten sind zu unterscheiden. Erkenntnisse aus einer einfachen Regression ($k = 1$) können evtl. nicht weit über den (rein deskriptiven) Vergleich bedingter Mittelwerte hinausgehen, so dass eine befriedigende Analyse meist erst mit $k \geq 2$ Regressoren möglich ist.

ratmetermiete und X_1, X_2, …, X_K sind Regressoren, wie Wohnungsgröße, Stockwerk, usw. oder auch 0-1-Variablen, wie z.B. Vorhandensein eines Balkons, eines Aufzugs usw. Es fällt manchen schwer einzusehen, warum z.B. der Koeffizient $b_{y1.2...}$ für den Balkon (X_1) plötzlich negativ (oder nichtsignifikant) werden kann, wenn ein Regressor (etwa X_7) durch einen anderen ersetzt wird, wo doch der Einfluss des Balkons vorher positiv und signifikant war. Vermieter meinen dann oft (scherzhaft), dass sie nach Abschlagen des Balkons eine höhere Miete verlangen könnten. Es geht also darum, einzusehen, warum die Schätzung der Modellkoeffizienten sozusagen "kontextabhängig" ist.

Die Erfahrung zeigt, dass viele mit dem dahinterstehenden Gleichungssystem

$$\begin{bmatrix} \sum y \\ \sum x_1 y \\ \sum x_2 y \end{bmatrix} = \begin{bmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 \end{bmatrix} \cdot \begin{bmatrix} b_{y.12} \\ b_{y1.2} \\ b_{y2.1} \end{bmatrix}$$

(oder kompakt in Matrixschreibweise geschrieben $\mathbf{X'y} = \mathbf{X'X}\cdot\mathbf{b}$) wenig anfangen können und dass sie sich nicht anschaulich vorstellen können, was es bedeutet, dass die Schätzung des Vektors $\mathbf{b} = \hat{\boldsymbol{\beta}}$ mit $(\mathbf{X'X})^{-1}\mathbf{X'y}$ eine inverse Matrix $(\mathbf{X'X})^{-1}$ verlangt. In der bei nur zwei Regressoren noch sehr übersichtlichen Darstellung in Gestalt von Rekursionsformeln, wie

$$(5) \qquad b_{y1.2} = \frac{b_{y1} - b_{y2}b_{21}}{1 - b_{12}b_{21}} = \frac{b_{y1} - b_{y2}b_{21}}{1 - r_{12}^2} \text{ und}$$

$$(6) \qquad b_{y2.1} = \frac{b_{y2} - b_{y1}b_{12}}{1 - b_{12}b_{21}} = \frac{b_{y2} - b_{y1}b_{12}}{1 - r_{12}^2}$$

ist die Interdependenz jedoch einfacher zu sehen. Hierin ist b_{y1} die (geschätzte) Steigung in $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_{1i} = b_{y.1} + b_{y1}x_{1i}$ und b_{12} (was i.d.R. ungleich b_{21} ist) ist die Steigung in der (einfachen) Regression von X_1 auf X_2, also $\hat{x}_{1i} = b_{1.2} + b_{12}x_{2i}$ (im Unterschied zur "umgekehrten" Regression $\hat{x}_{2i} = b_{2.1} + b_{21}x_{1i}$).

Damit werden gleich zwei interessante Spezialfälle verständlich

- korreliert der hinzugekommene Regressor X_2 mit X_1 vollständig ($|r_{12}| = 1$) kann $b_{y1.2}$ und $b_{y2.1}$ nicht berechnet werden (Null im Nenner von Gl. 5 und 6 und wir haben es jetzt bei X_1 und X_2 faktisch nicht mit zwei Variablen zu tun, sondern eigentlich nur mit einer);
- sind X_1 und X_2 unkorreliert ($r_{12} = 0$ und damit auch $b_{12} = b_{21} = 0$) dann gilt hier (und nur hier) $b_{y1.2} = b_{y1}$ (und auch $b_{y2.1} = b_{y2}$), so dass hier eine hinzugenommene Variable X_2 nichts an Stärke und Richtung (Vorzeichen) des Einflusses von X_1 auf Y ändert (hier und nur hier gilt, was fälschlich generell erwartet wird, dass nämlich bei einer Erweiterung eines Modells die bestehenden Modellbausteine unverändert bleiben). [74]

Um zu sehen, wie sich unterschiedliche Annahmen über die Variable X_2 auf den mit $b_{y1.2}$ gemessenen Einfluss von X_1 auf Y in Relation zu b_{y1} auswirken stellen wir fest, dass alles von sechs Größen (Varianzen s^2 und Kovarianzen s) abhängt. Wir nehmen fünf Größen als gegeben (s_{y1}, s_y^2, s_1^2, s_{y2} und s_2^2) an und variieren eine (s_{12} und damit r_{12}):

$s_{y1} = 5$, $s_y^2 = 9$ und $s_1^2 = 4$ (Varianz von X_1)
$\Rightarrow r_{y1} = 5/6 = 0{,}833$ und $b_{y1} = 5/4 = 1{,}25$
die nächsten zwei Größen sind $s_{y2} = 8$, und $s_2^2 = 16$ damit erhält man
$\Rightarrow r_{y2} = 8/12 = 0{,}666$ und $b_{y2} = 8/16 = ½$

Unterschiedliche Annahmen über s_{12} (Kovarianz zwischen X_1 und X_2) wirken sich auch auf b_{12}, b_{21} und r_{12} aus (und damit auch auf die partiellen Regressionskoeffizienten)

	s_{12}	r_{12}	$b_{y1.2}$	$b_{y2.1}$
1	6	0,750	1,14286	0,07143
2	3	0,375	1,01818	0,30909
3	0	0	1,25	0,5
4	- 3	- 0,375	1,89091	0,85455
5	- 6	- 0,750	4,57143	2,21429

und wenn r_{12} ungefähr +1 oder -1 ist:

	s_{12}	r_{12}	$b_{y1.2}$	$b_{y2.1}$
6	7,5	+ 0,9375	2,58065	- 0,70968*
7	- 7,5	- 0,9375	18,0645	8,96774

* Wie Fall 6 zeigt, kann sich sogar das Vorzeichen ändern von $b_{y2} = 0{,}5$ zu $b_{y2.1} = -0{,}70968$.

[74] Die Fälle sind der negative und positive Extremfall von (vollständiger) Kollinearität (negativ) und unabhängigen (unkorrelierten) Regressoren (positiv). Nur bei unabhängigen Regressoren ist auch die multiple Bestimmtheit (von y durch x_1 und x_2 und …) die Summe der einfachen Bestimmtheiten (y durch x_1, y durch x_2 usw.).

Man kann leicht nachrechnen, dass $b_{y2.1}$ mit $r_{12} = 0{,}8$ null wird und bei $r_{12} > 0{,}8$ dann auch negativ wird, während $b_{y1.2}$ nur positiv sein kann.

Wir weisen nur darauf hin, ohne dies zu demonstrieren, dass mit zunehmendem Wert von $(r_{12})^2$ also mit gegen Null strebendem Nenner $1 - (r_{12})^2$ in den Gleichungen 5 und 6 die Streuung (der Stichprobenfehler) der geschätzten β-Koeffizienten b größer (und das Konfidenzintervall breiter) wird.

Betrachtet man die Ergebnisse der Tabelle, so neigen viele dazu, die jeweils positiven Koeffizienten[75]

	$b_{y1.2}$	$b_{y2.1}$
1	1,14286	0,07143
2	1,01818	0,30909
3	1,25	0,5

dahingehend zu interpretieren, dass X_1 die "wichtigere" Einflussgröße für Y sei, weil $b_{y1.2}$ durchwegs größer ist als $b_{y2.1}$. Solche Betrachtungen sind jedoch aus mindestens zwei Gründen falsch

- wenn $r_{12} \neq 0$ ist, dann misst $b_{y1.2}$ nicht nur den Einfluss von X_1, sondern z.T. auch den von X_2 und entsprechend $b_{y2.1}$ nicht nur den von X_2, sondern z.T. auch den von X_1 und weil
- auch wenn $r_{12} = 0$ ist, der erste Einwand also nicht Platz greift die Koeffizienten nicht in ihrer Größe vergleichbar sind, weil sich die Variablen X_1 und X_2 in ganz unterschiedlichen Größenordnungen bewegen können.

Um das zu berücksichtigen müsste man die "standardisierten" partielle Regressionskoeffizienten b* berechnen[76]

$$b^*_{y1.2} = b_{y1.2} \cdot \frac{s_1}{s_y} \text{ und } b^*_{y2.1} = b_{y1.2} \cdot \frac{s_2}{s_y}.$$

Da die Standardabweichung von X_2 doppelt so groß ist wie die von X_1 ($s_2 = 4$, $s_1 = 2$ und $s_y = 3$) erhält man $b^*_{y1.2} = 0{,}833$ und $b^*_{y2.1} = 0{,}667$. Der Unterschied ist zwar nicht mehr ganz so groß wie bei den nichtstandardisierten Koeffizienten, aber es ist in jedem Fall falsch zu meinen, X_1 sei 2½ mal so "wichtig" wie X_2 (weil 1,25/0,5 = 2,5).[77]

Zwar spricht man beim Fehlschluss, wonach die Regressionskoeffizienten unberührt sein sollten vom Hinzu- oder Wegtreten von anderen Regressoren oder ihre Beträge gäben Aufschluss über die relative Stärke eines Einflusses nicht von einer "fallacy" oder einer "Paradoxie", aber täte man es, so wäre auch dies wieder mal so ein typisches Verständnisproblem einer statistischen Methode, das manche zum beliebten, aber unsinnigen Vorwurf von "Lügen mit Statistik" veranlasst.

Wie man hier und auch im vorangegangenen Abschnitt über Endogenität sieht, sind auch die in der Literatur immer wieder behandelten "fallacies" "biases" nicht leicht vollständig, überdeckung- und widerspruchsfrei zu klassifizieren und auseinanderzuhalten. Es lohnt sich gleichwohl, zu versuchen, hier etwas System hineinzubringen und daraus auch – im positiven Sinne – allgemeine Lehren für eine gute Statistik zu gewinnen.[78]

5. Einige grundlegende Denkmuster in der Statistik

Statistik "verstehen", wie es im Titel dieses Papiers heißt, bedeutet vor allem, mehr und mehr Zusammenhänge zwischen Methoden zu erkennen und das hinter ihnen bestehende *System* zu sehen. So sieht man z.B. dass Schätzen und Testen verschiedene Darstellungen des gleichen empirischen Befundes sind, oder dass die Varianzanalyse als Spezialfall der Regressionsanalyse aufgefasst werden kann (was zu vermitteln ein Hauptanliegen des Lehrbuchs von Roger Bakeman und Byron Robinson ist).

[75] Die im Folgernden beschrieben falsche Betrachtungsweise kommt schon dann in Schwierigkeiten, wenn die (partiellen) Regressionskoeffizienten unterschiedliche Vorzeichen haben.

[76] Sie heißen auch β-Koeffizienten, sollen aber hier mit b^* bezeichnet werden.

[77] Ich erinnere mich noch gut daran, wie in den 70er Jahren in Marburg ein Habilitand in der VWL-Theorie eine solche Regression mit zwei Regressoren von den Statistikern rechnen ließ und nicht nur X_1 für die wichtigere Variable als X_2 hielt weil für die nichtstandardisierten Koeffizienten $b_{y1.2} > b_{y2.1}$ galt, sondern dies auch noch theoretisch begründete. Rechnete man mit den standardisierten Koeffizienten war das Gegenteil der Fall (X_2 erschien danach wichtiger als X_1).

[78] Das ist im Prinzip das, was wir uns hier mit diesem Papier vorgenommen haben.

Zu den Grundvorstellungen (oder typischen "Denkfiguren") der Statistik gehört auch der Gedanke, dass eine Vergrößerung der Datenbasis i.d.R. von Vorteil ist, was aber jedoch auch – wie unter 5b gezeigt – zu relativieren ist

In diesem Abschnitt 5 können natürlich nicht alle "grundlegenden Denkmuster in der Statistik" dargestellt werden. Zu solchen Mustern gehört z.B. sicher auch die in Abschn. 6b und 6c noch einmal behandelte "Logik" von Signifikanztests und das Konzept der Stichprobenverteilung einer Schätzfunktion (= Stichprobenfunktion) wie z.B. die Schätzfunktion von $\bar{x}$ für μ (oder p für π).

a) Systematisch und zufällig: was heißt "Erklären" in der Statistik?

Bei der Darstellung der hinter dem "Experiment" stehenden "Logik" haben wir bereits einen in der Statistik immer wieder auftretenden Gedanken kennen gelernt: bei den nicht zu kontrollierenden (nicht willentlich zu variierenden) Einflüssen sollte man den Zufall "walten" lassen, weil es oft gerechtfertigt ist, anzunehmen, dass sich diese weniger bedeutsamen und in beide Richtungen (positiv und negativ) wirkenden Einflüsse gegenseitig aufheben. Der gleichen Erwartung entspringt auch die Vorstellung, dass man mit wiederholten Messungen (replications) z.B. bei einem Experiment eher zum richtigen Ergebnis kommt, weil sich (als zufällig anzunehmende) Messfehler gegenseitig aufheben. In diesem Sinne unterscheidet man bei "Fehlern"

Ursache		"heben sich auf" (Ausgleich)
systematisch	bekannt	nein; sie können auch alle nur positiv oder nur negativ sein
zufällig	unbekannt	ja; sie können + *und* - sein

Bei randomization im Experiment macht man sich die die *Ausgleichstendenz beim Zufall* zu Nutze. Ein weiteres Argument dafür, den Zufall walten zu lassen wird bei Stichproben genutzt: hier geht es um die *Vermeidung von Einseitigkeit und Willkür,* um die Eigenschaft des Zufalls, "blind" zu sein, keine Präferenzen zu haben und ohne Ansehung der Person zu agieren (mehr dazu auch in Abschn. 6a).

Die entsprechende Terminologie beim "Erklären" mit "Einflussfaktoren", auf die wir gleich zu sprechen kommen ist:

Variationsquelle (VQ)		F-Test und H_0 dabei
systematisch	erklärt; zwischen (between)	$F = V_E/V_R$ mit V_E = erklärte Varianz V_R = Residualvarianz H_0: nur Zufall (VR)
zufällig	residual, innerhalb (within)	

"Erklären" einer Variablen y bedeutet in der Statistik, dass die Varianz von y in eine "systematische" und eine "zufällige" Varianzkomponente zerlegt werden kann. Die Vorstellung ist, dass y schwankt, weil y von systematischen Einflüssen bestimmt wird, die aber auch noch durch zufällige (auch "noise" genannt) überdeckt werden und es gilt, ein durch "noise" verdecktes Muster aufzudecken. [79] In der einfachen (einfaktoriellen) Varianzanalyse kann z.B. der Ernteertrag Y durch den *qualitativen* Faktor Art des Düngemittels mit den "Stufen" (levels) A_1, A_2, …, A_I erklärt werden Wir unternehmen n_1, n_2, … n_I Messungen mit dem jeweiligen level und erhalten die mittleren Ernteerträge

$$\bar{y}_1 = \frac{\sum_{j=1}^{n_1} x_{1j}}{n_1}, \bar{y}_2 = \frac{\sum_{j=1}^{n_2} x_{2j}}{n_2}, ..., \bar{y}_I = \frac{\sum_{j=1}^{n_I} x_{Ij}}{n_I}$$

und über alle Faktorstufen gerechnet den Gesamtmittelwert $\bar{y} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} \Big/ (n-1)$ mit n $= n_1 + n_2 + \ldots + n_I$. Für die Unterschiedlichkeit der Mittelwerte $\bar{y}_1, \bar{y}_2, ..., \bar{y}_I$ (unterschiedliche mittlere Ernteerträge) im Verhältnis zu $\bar{y}$ ist der "Faktor" (≈ die Ursache) A = Art des Düngemittels verantwortlich, so dass die Summe der Abweichungsquadrate $SSE = \sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2$ als (mit der Düngemittelart) "erklärte" Variationsquelle VQ zwischen den Stufen [between] gelten kann.

Dass trotz gleicher Faktorstufe (gleicher Düngemittelart) y schwanken kann, liegt an zufällig wirkenden anderen Faktoren, die eine nicht erklärte (residuale) Variationsquelle (innerhalb [within] der Stufen) mit

[79] Eine "Konstante", etwa y = 14 kann man nicht mit Mitteln der Statistik "erklären" und außerhalb der Statistik hat "Erklären" eine viel weiter gehende Bedeutung, nämlich etwas auf seine Ursachen zurückführen und, mehr noch (im Idealfall), den Mechanismus aufzeigen, mit dem die fragliche Ursache die beobachtete Wirkung quasi "erzeugt".

$SSR = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{I}(n_i - 1)s_i^2$ darstellen, wobei $s_i^2 = \sum_j (y_{ij} - \bar{y}_i)^2 / (n_i - 1)$ ist.

Wie man sieht, gilt die Zerlegungsformel $SST = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = SSE + SSR$, worin T für "total" steht und man erhält dann die bekannte Tabelle

VQ		df*	Varianz (mean square)
erklärt	SSE	I - 1	MSE = SSE/(I-1)
residual	SSR	n - I	MSR = SSR/(n-1)
Summe	SST	n - 1	Summe nicht sinnvoll

* degrees of freedom (Anzahl der Freiheitsgrade)

und die Prüfgröße F = MSE/MSR, die F-verteilt ist mit I - 1 und n - 1 Freiheitsgraden und mit der die Nullhypothese H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$ (kein Einfluss des Faktors A; Ernteschwankungen nur zufällig) getestet wird.

Man kann zeigen, dass der bekannte t-Test für zwei unabhängige Stichproben der Spezialfall hiervon mit I = 2 und der H_0: $\mu_1 = \mu_2$ oder, was ja identisch ist $\mu_1 - \mu_2 = 0$ (daher auch der Ausdruck *Null*-hypothese) ist.

Betrachtet man nun zur "Erklärung" des Ernteertrags eine Regressionsfunktion mit K Regressoren als *quantitative* Variablen X_1, $X_2,\ldots,X_K$ (etwa Düngemittel*menge*) so ist

(7) $\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_K x_{Ki}$

(i = 1, …, n; wir haben in diesem Papier auch mit den Symbolen a und b statt $\hat{\alpha}$ und $\hat{\beta}$ operiert) als "erklärter" Teil von y_i (bei $y_i = \hat{y}_i + \hat{u}_i$) aufzufassen und $\hat{u}_i$ als nicht erklärter (zufälliger) Teil von y_i.

Da die "Regresswerte" $\hat{y}_i$ wie auch die beobachteten Werte (unsere Daten) y_i im Mittel gleich $\bar{y}$ sind (weil die $\hat{u}_i$ ja bei der LS-Methode im Mittel null sind) gilt

$SSE = S_{\hat{y}\hat{y}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ und

$SSR = S_{\hat{u}\hat{u}} = \sum_{i=1}^{n}\hat{u}_i^2$ sowie

$SST = S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ und $S_{yy} = S_{\hat{y}\hat{y}} + S_{\hat{u}\hat{u}}$.

Man erkennt sofort anhand der Zerlegung

VQ		df	Varianz
erklärt	$S_{\hat{y}\hat{y}}$	K = k-1	$S_{\hat{y}\hat{y}}/K$
residual	$S_{\hat{u}\hat{u}}$	n-K-1	$S_{\hat{u}\hat{u}}/(n-K-1)$
Summe	S_{yy}	n-1	

die Ähnlichkeit mit den vorangegangenen Ausführungen zur Varianzanalyse. Jetzt ist auch die Prüfgröße F ganz analog definiert als $F = \frac{S_{\hat{y}\hat{y}}/K}{S_{\hat{u}\hat{u}}/(T-K-1)}$. Zum Begriff "Freiheitsgrad": In Gl. 7 sind k = K+1 Parameter zu schätzen, nämlich α, $\beta_1,\ldots$, β_K in Gestalt von $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_K$. Hat man $\hat{\beta}_1, \ldots, \hat{\beta}_K$ so steht auch $\hat{\alpha}$ fest weil dann bei der Methode der kleinsten Quadrate (LS-Methode) auch $\bar{y} = \hat{\alpha} + \hat{\beta}_1\bar{x}_1 + \ldots\hat{\beta}_K\bar{x}_K$ gelten muss.

Die Betrachtungen mit der Tabelle und mit F gelten ganz allgemein, für *jedes* K, etwa für K = 1 oder K = 2. Im Abschn. 3b (regression to the mean) hatten wir den Fall einer einfachen linearen Regression (K = 1)[80] und in Abschn. 4d einen Fall mit K = 2 Regressoren (X_1 und X_2).

Wir sahen nicht nur die fundamentale Bedeutung der Unterscheidung systematisch/erklärt vs. zufällig/residual, sondern auch ein Beispiel dafür wie verschiedenen Methoden in der Statistik gleiche Grundgedanken zugrundeliegen.

Verwandt sind auch "schätzen" und "testen". Ein Konfidenzintervall für μ zu berechnen dessen Grenzen beispielsweise 380 und 420 sind und die H_0: $\mu = 400$ anzunehmen (oder wenn die Grenzen 350 und 390 wären H_0 abzulehnen) läuft auf das Gleiche hinaus.[81]

Was die so beliebten "fallacies" betrifft, so läuft beides (Schätzen und Testen) auch darauf hinaus, sehr viel mehr als wir das im Alltagsleben

[80] Der t-Test der Hypothese $\beta = 0$ ist identisch mit dem F-Test, denn bei 1 und n – 2 Freiheitsgraden gilt für die Prüfgrößen $F = t^2$.

[81] Dieses Beispiel von diametral entgegengesetzten Testentscheidungen, aber gleichwohl überlappender Konfidenzintervalle [380; 420] und [350; 390] zeigt, dass die Testentscheidung allein ein falsches Bild zeichnen kann. Das ist die Botschaft des Buches von Geoff Cumming, in dem er sich großspurig als Schöpfer einer "New Statistics" präsentierte (was aber eigentlich alles schon im Buch von Paul D. Ellis enthalten war). Wir gehen darauf in Abschn. 6c weiter ein.

tun auch die Möglichkeit ins Auge zu fassen, dass etwas nicht explizit mit einer bekannten Ursache zu "erklären" ist, sondern (was ja die "Annahme" von H_0 bedeutet) "nur Zufall" sein könnte. Die Nullhypothese zu testen, heißt ja, zu fragen, wie wahrscheinlich der Stichprobenbefund ist, wenn es keine "systematische", sondern nur die zufällige Variationsquelle gibt.

b) "Big data" und Statistik: die große Masse macht's?

Unter dem Stichwort "Big data" versteht man die statistische Auswertung von Daten, die permanent, und in großen Massen, quasi als kostenloses Nebenprodukt generiert werden durch Internet-, Smartphone-, Kreditkartennutzung usw., aber auch durch Verkehrssteuerungssysteme, Kameras usw.

"Amazon überwacht unsere Produktvorlieben und Google unser Surfverhalten… Twitter und Facebook sammeln und speichern Daten über die sozialen Beziehungen der Menschen."[82] Nicht nur die Flut an digitalen Informationen, auch die ebenso gigantisch gestiegenen Rechen- und Analysemöglichkeiten[83] nähren die Hoffnung, so mit statistischen Auswertungen zu empirischen Erkenntnissen zu gelangen (vor allem auch wenn personenbezogene Daten zu Prognosen über das künftige Verhalten dieser Personen und zur *Vorbeugung* von Verbrechen genutzt werden).[84] Die Beschreibung ist hier, wie schon beim "Data Mining" mit automatischen Suchen nach "etwas Wertvollem" (Muster, Assoziationen, Korrelationen, bewusst ohne Anspruch auf Kausalität) etwas vage: man spricht z.B. von[85] "extract new insight, create new forms of value".

Vom Standpunkt der Statistik sind jetzt primär zwei Fragen von Interesse

- Haben wir hier eine quantitative Steigerung oder einen qualitativen Sprung gegenüber dem, was Statistik (und ihren Wert) seit je her ausmacht; systematisch "Massen" (Gesamtheiten mit vielen Einheiten) zu betrachten, weil erst bei Beobachtung vieler (gleichartiger) Fälle Regelmäßigkeiten sichtbar werden?[86]
- Ist Big Data und der Trend zu Data Warehouse Systems (Vorhalten von Daten und Analysen) eine Bedrohung der Sonderstellung der *amtliche*n Statistik als Produzent statistischer Daten?[87]

Sieht man, was sich die Enthusiasten von "Big Data" hiervon versprechen und wie sie "Data Mining" definieren, nämlich als "the art and science of intelligent data analysis", was im Ergebnis "will continue to help change our world for the better",[88] so mag man sich fragen was hier für die "traditionelle" Statistik noch zu tun übrig bleibt.

Eine ziemlich allgemeine Überzeugung scheint zu sein, dass trotz eines gemeinsamen Fokus auf "Massen" bei Statistik und Big Data bei letzterer doch noch eine Art "qualitativer Sprung" vorliegt, wobei die Begründung (im Folgenden nach Mayer Schonberger und Cukier) vage und nicht unproblematisch zu sein scheinen:

- Als prinzipiell neu gilt der Verzicht auf Kausalität (Korrelation genügt) "society will need to shed some of its obsessions for causality in exchange for simple correlation not knowing *why* but only *what*. This challenges our most basic understanding of how to make decisions and comprehend reality."[89]
- mehr Vollerhebungen, Stichproben sind ein Relikt aus der "small data environment" (es fragt sich, ob damit nicht die übliche Unterscheidung zwischen deskriptiver und induktiver Statistik hinfäl-

[82] Wirtschaftswoche 41/2013 (aus einem Artikel zu einem Vorabdruck der deutschen Übersetzung des Buchs von V. Mayer-Schönberger und K. Cukier).
[83] Nate Silver zitiert eine Schätzung von IBM, wonach täglich 2,5 Quintillionen (10^{18}) Bytes Daten hinzu kommen.
[84] Die Befürchtung ist dann, dass eine Person für eine Tat bestraft wird, die sie nicht begangen hat, die sie aber – angesichts ihres aus den Daten herausgefilterten "Profils" – mit einer hohen Wahrscheinlichkeit begehen *könnte*.
[85] Hier und im Folgenden v.a. Zitate aus dem Buch von Mayer-Schönberger und Cukier.
[86] Beispiel: in der "Masse" sind Knaben- und Mädchengeburten etwa gleich häufig, was aber der einzelne auf Grund seines Bekanntenkreises oft nicht erkennt, weil er Familien kennt mit drei Töchtern und keinen Sohn oder mit zwei Söhnen und einer Tochter.
[87] Auf diese Frage bin ich an anderer Stelle (mein Papier "Statistik und Manipulation") eingegangen.
[88] G. Williams, Data Mining with Rattle and R, 2011.
[89] Nate Silver zitiert dagegen (missbilligend) einen US Journalisten, der meinte "that the sheer volume of data would obviate the need for theory, and even for the scientific method."

lig wird, was wohl auch das folgende Argument erklärt) [90],

- "loosen up our desire for exactitude: with less error from sampling we can accept more measurement error" (als ob man eine Art Fehler durch eine andere Art von Fehler substituieren könnte) und
- Expertenwissen (im jeweiligen Sachgebiet) wird weniger wichtig, weil der Computer die korrekten Wahrscheinlichkeiten liefert (gedacht ist an die oben behandelte Vorstellung der Vorhersagbarkeit eines einzelnen Ereignisses aufgrund von Wahrscheinlichkeiten).[91]

Ein Großteil der hier aufgelisteten methodischen Besonderheiten von Big Data gegenüber der Statistik scheint darauf hinauszulaufen, dass man es sich dank mehr Rechen- und Speicherkapazität erlauben kann, methodische Skrupel und Qualitätsansprüche der Statistik über Bord werfen zu können. [92]

Trotz der Gemeinsamkeit von Statistik und "Big Data" beim Motto "Die große Masse macht's" (mit dem Anspruch, gerade mit der Masse der Daten auch substanziell neue Erkenntnisse zu liefern) kann das Ergebnis doch sehr verschieden sein. Die "große Masse" ist also nicht *hinreichend* als Kennzeichen oder Fundament der Statistik. In mancher Hinsicht ist Big Data sogar eher das glatte Gegenteil von Statistik, nämlich

- methodisch, z.B. wenn man bewusst darauf verzichtet Ursachen festzustellen und sich mit Korrelationen begnügt[93]
- und hinsichtlich des Zwecks von Datenanalysen, was im Falle von Big Data nahe an Überwachung, Spionage u. ä. liegt und keinesfalls Aufgabe der Statistik ist.

Neben Enthusiasten gibt es deshalb auch Skeptiker. Man findet Vorbehalte bei Nate Silver und noch deutlich mehr bei Kaiser Fung: "here is a real danger that Big Data moves us backward, not forward. It threatens to take science back to the Dark Ages, as bad theories gain ground by gathering bad evidence and drowning out good theories."[94] Er begründet das mit der zu erwartenden Überflutung (eines gerade deswegen unkritischen Publikums) mit konkurrierenden, zweifelhaften "empirischen Erkenntnissen".

Wir beschränken uns hier bewusst auf mögliche Entwicklungen in puncto Methoden der empirischen Forschung, weil es uns auf die Methoden ankommt, bei Big Data einerseits und Statistik andererseits. Viele sehen das Problem mit Big Data vor allem in der "Anwendung" auf die einzelne Person, also etwas, was in der Statistik gerade nicht vorgesehen ist: individuell zugeschnittene Versicherungsbeiträge, Verhalten des Arztes bei "big data driven diagnosis", oder von Kommissionen bei der Auswahl aus Bewerbern, und der frühe Verdacht (oder gar die Bestrafung) bei Straftaten, die man noch gar nicht begangen hat.

c) Modelle und Prüfung der Modellvoraussetzungen

Man erlebt oft, dass ein Vater von seinem Sohn (der weithin sichtbar der totale Versager ist) sagt, er sei hyperintelligent aber leider sehr faul. Es kommt so gut wie nie vor, dass es heißt, er habe sich enorm bemüht, aber all sein Schuften hat nichts geholfen, weil er nie etwas kapiert hat. Vieles was mit der Statistik gemessen wer-

[90] Dass das blanker Unsinn ist wird schnell klar, wenn man an die statistische Qualitätskontrolle oder an Versuchsreihen in der Pharmaindustrie denkt. Auch bei noch so vielen Daten wird ein Hersteller von Glühbirnen die Lebensdauer seiner Produkte nur an einer kleinen Stichprobe überprüfen, weil er doch den größeren Teil seiner Produktion verkaufen will. Ein anderes Beispiel ist die Capture-Recapture Methode: wie stellt man die Anzahl der Fische in einem Teich fest ohne das Wasser abzulassen und dann nur noch tote Fische zählen zu können? Es ist auch kein Relikt aus einer überwundenen trüben Vorzeit, wenn man neue Medikamente nicht gleich im Großversuch testet.

[91] Die Autoren zitieren Buch und Film "Moneyball", in dem ein Informatiker per Computer die optimale Basketball Mannschaft zusammenstellt, die nach einer Anlaufzeit jedes Spiel gewinnt. Wie man sieht, beflügelt das "Big" zu Utopien.

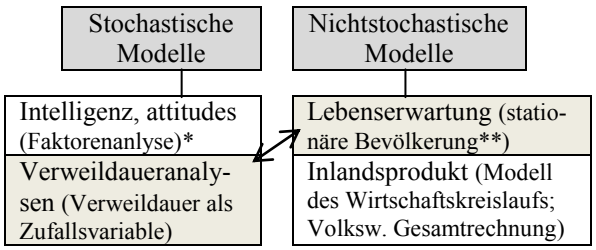
[92] Mayer Schonberger und Cukier erwähnten, dass bei Google mal eben 450 Millionen (!!) mathematische (vermutlich ökonometrische) Modelle mit verschiedener Auswahl von "search terms" als Regressoren auf ihre Prognosefähigkeit ausprobiert wurden.

[93] Das und das blinde Vertrauen in einen nicht durchschauten "Algorithmus" ist nicht nur das Gegenteil von Statistik, sondern auch das Gegenteil von Wissenschaft, vor allem wenn man trotzdem glaubt Täter von noch nicht begangenen Taten aufgrund einer berechneten Wahrscheinlichkeit persönlich verantwortlich machen zu können.

[94] K. Fung, Numbersense 2013.

den soll, tangiert Werte (gut/schlecht) wo man keine "objektiven" Auskünfte von Befragten erwarten kann. Es macht keinen Sinn, eine Umfrage durchzuführen, und die Leute zu fragen: "Sind Sie intelligent?"[95] Es wäre auch sinnlos, einen gerade geborenen Säugling nach ihrer/seiner Lebenserwartung[96] zu fragen. Niemand kennt sein voraussichtliches Sterbealter.

Es gibt also Dinge, die man nur indirekt, nur mit einem Modell messen kann, und auch hier kann die Statistik viel leisten. Es ist zu unterscheiden:



* Man liest in deutschen Texten auch öfter "Faktoranalyse", offenbar wegen "factor analysis", oder weil man nicht weiß, dass i.d.R. mehr als nur ein Faktor "extrahiert" wird.

** = Sterbetafelbevölkerung (eine Berechnung nach Art einer Sterbetafel, allerdings nicht nur mit konstanten, sondern sogar für jedes Alter x *gleichen* Sterbewahrscheinlichkeiten unternehmen wir im Abschn. 6a im Zusammenhang mit der survivor bias).

Die Unterscheidung hängt nicht zwingend mit dem Gegenstand der Untersuchung zusammen. und wir haben deshalb auch ein Beispiel für zwei verwandte Gegenstände farblich hervorgehoben Die Lebenserwartung ist eine Art von "Verweildauer" und man kann so etwas auch mit einem stochastischen Modell untersuchen, wenn man z.B. mit der Weibull Verteilung als Wahrscheinlichkeitsverteilung arbeitet und daraus die hazard rate (entspricht Sterbewahrscheinlichkeit) und Überlebensfunktion (Absterbeordnung bei der Sterbetafel) abgeleitet.

Der Vorteil einer Analyse auf Basis eines *stochastischen* Modells ist, dass nicht nur Hypothesen über Parameter des Modells geprüft werden können, sondern dass ein solches Modell (i.d.R. in Gestalt von einer Gleichung oder einem Gleichungssystem) gerade *wegen der Zufallsvariable(n)* in der (den) Gleichung(en) im Hinblick darauf beurteilt werden kann, wie gut das Modell insgesamt den Stichprobendaten angepasst ist.[97]

Man hat also hier (aber nicht bei den *nichtstochastischen* Modellen) mit Maßen der Güte der Anpassung (goodness of fit) die Möglichkeit, von schlechteren zu besseren Modellen zu schreiten. Man kann und sollte hier mit statistischen Tests prüfen (was bei *nichtstochastischen* Modellen nicht geht):

- Hypothesen über Parameter (meist H_0: $\beta_k = 0$, also kein Einfluss von X_k oder $\beta_1 = \beta_2 = \ldots = \beta_K = 0$, so dass y nur durch eine Konstante und die Störgröße u bestimmt wird),
- über die Güte der Anpassung des Modells als Ganzes, und
- ob die Modellannahmen erfüllt sind.

Es ist durchaus möglich, dass viele oder alle Bestimmungsfaktoren (im Sinne von Regressoren $X_1, X_2, \ldots, X_K$ in einer geschätzten Regressionsgleichung) "relevant" sein können, in dem Sinne, dass die entsprechenden Regressionskoeffizienten signifikant sind, die Anpassung des Modells insgesamt aber gleichwohl unbefriedigend ist, so dass man das "richtige" Modell noch nicht gefunden hat und es mit einer anderen Menge (Auswahl) von Regressoren versuchen muss. Bei unbefriedigenden Schätzergebnissen kann

- es an der *Spezifikation* (z.B. Auswahl der Regressoren) liegen, aber auch
- an den *Daten*, wenn die Variablen z.B. zu wenig streuen oder untereinander zu sehr korrelieren.

Ein Modell kann auch ganz unabhängig von den Daten überhaupt gar nicht schätzbar sein, weil Koeffizienten in den Gleichungen nicht "identifiziert" sind. In einem "unteridentifizierten" Modell müssen Gleichungen (und damit das Modell) geändert werden, um es überhaupt erst schätzbar zu machen.

[95] Auch Einstellungen (attitudes) gehören dazu: man kann keine Umfrage machen und fragen: sind Sie ein Rassist oder ein Antisemit usw.

[96] Mit *der* Lebenserwartung ist meist die einer/eines Nulljährigen gemeint, also e_0. Die Lebenserwartung e_x ist aber eine Funktion des Alters x, so dass sich e_0, e_1, e_2,… unterscheiden. Eine häufige Fehlvorstellung ist auch, dass e_{x+1} stets genau ein Jahr weniger sein müsste als e_x (so etwas würde nur bei einer rechteckigen Überlebensfunktion gelten: alle werden $x = x_{max} = \omega$ Jahre alt und sterben dann alle im gleichen Alter ω).

[97] Das nichtstochastische Modell der stationären Bevölkerung ist z.B. bekanntermaßen unrealistisch (ein heute 20 Jähriger hat in 10 Jahren die gleiche Sterbewahrscheinlichkeit q_{30} wie ein heute 30 jähriger). Man beachte: trotz Sterbe*wahrscheinlichkeit* ist das Modell nichtstochastisch, weil e_x streng funktional (ohne eine Störgröße) von den Größen q_x abhängt.

Das "Prüfen" mit statistischen Tests betrifft nicht nur Schätzergebnisse, sondern auch die für die Schätzung eines Modell zu treffenden Annahmen über die Variablen,[98] die mit einer Gleichung bestimmte Funktionsform und vor allem über die Verteilungen der Störgrößen.[99] Andererseits ist zu bedenken, dass mit dem Vorteil, verschiedene Modelle beurteilen zu können auch verbunden ist, dass die jeweils getroffenen Modellannahmen überprüft werden sollten.

Solche Annahmen sind notwendig damit die Schätzungen bestimmten Gütekriterien[100] genügen). Es wird oft nicht beachtet, dass die Nullhypothese H_0 bei solchen Tests besagt, dass die betreffende Modellannahme erfüllt ist, so dass man – anders als bei den eher gewohnten Tests der Koeffizienten – an der *Annahme* von H_0, und nicht an deren Ablehnung interessiert ist ("nicht signifikant" ist hier also gut).

Maßgebend für die Wahl einer statistischen Methode (und damit meist auch eines stochastischen Modells) sind

1. die Fragestellung und
2. die Qualität der Daten, insbesondere das Skalenniveau der Variablen; aber
3. es gibt auch Fälle, wo es bei der Methodenwahl auf die inhaltliche Aussagefähigkeit der Daten ankommt.

Zu 1 (*Zwei oder mehr Methoden können für die gleiche Fragestellung geeignet sein):* So kann z.B. die Frage, welcher von zwei vorgegebenen Gruppen (wie Käufer/Nichtkäufer eines Produkts oder Diagnose krank/nicht krank) eine Einheit aufgrund ihrer Merkmale (wie Einkommen, Beruf usw. bzw. im Fall gesund/krank Blutdruck, Fieber usw.) zugeordnet werden sollte mit der Diskriminanzanalyse (DA) oder mit der "logistischen Regression" (Logit Analyse, LA) behandelt werden. Bei der DA werden Gewichte β_0, β_1, …, β_K der Linearkombination der x-Werte $\tilde{x}$ so bestimmt, dass eine Einheit i aufgrund von $\tilde{x}_i = \beta_0 + \beta_1 x_{1i} + \ldots + + \beta_K x_{Ki}$ möglichst treffsicher einer der beiden Gruppen zugeordnet werden kann. Bei der LA wird die Wahrscheinlichkeit πi, dass die Einheit i zur Gruppe Käufer bzw. gesund gehört, bzw. genauer die transformierte Wahrscheinlichkeit $\lambda_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ mit $\hat{\lambda}_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_K x_{Ki}$ geschätzt und i nach Maßgabe von $\hat{\pi}_i$ der Gruppe 1 zuordnet (wenn $\hat{\pi}_i > ½$) oder der Gruppe 2 also Nichtkäufer, bzw. krank, wenn $\hat{\pi}_i \leq ½$ (man beachte, dass $\hat{\pi}_i$ - anders als $\tilde{x}_i$ - auf den Wertebereich $0 \leq \hat{\pi}_i \leq 1$ normiert ist).[101]

Die Fragestellung ist also verwandt, nicht aber die Methode und ihre Modellvoraussetzungen. Die DA gilt als Verfahren der "multivariaten Analyse", die LA nicht. Weil es in $\lambda_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_K x_{Ki+} u_i$ K > 1 Regressoren gibt, liegt eine *multiple* Regression vor, "*multivariat*" ist ein Modell, wenn es auch mehrere (oft "latente") y-Variablen gibt (z.B. die latenten "Faktoren" bei der Faktorenanalyse, oder die Linearkombinationen aus K x-Variablen und aus m y-Variablen bei der kanonische Korrelation).

Zu 2: Das *Skalenniveau* spielt eine Rolle bei der Entscheidung für einen parametrischen oder einen nichtparametrischen Test oder die Messung der Korrelation zwischen zwei Variablen X und Y.

Sind X und Y metrisch skaliert kann man mit der üblichen (Produkt-Moment) Korrelation r rechnen, bei Ordinalskalen mit Rangkorrelationen (es gibt hier verschiedene Formeln), ist eine Variable dichotom und die andere Variable metrisch skaliert, kann man point biserial correlation coefficients berechnen, sind beide dichotom kann man mit einen Assoziationskoeffizient (wie ϕ) rechnen, wie es hier in diesem Papier in verschiedenen Abschnitten geschieht.

[98] Auf eine solche Annahme sind wir z.B. in Abschn. 4c eingegangen: keine endogene Regressoren.

[99] Genau genommen liegt jeder der n (i = 1, 2, …, n) Beobachtungen der y_i (das ist eine Zufallsvariable weil u_i eine Zufallsvariable ist) eine Zufallsvariable zugrunde, weshalb man auch von Zufallsvariablen u_1, u_2, …, u_n sprechen kann. Aber weil man unabhängige "Züge" aus *identischen* Verteilungen annimmt kann man auch von *der* Zufallsvariable U sprechen mit u_i als einer ihrer "Ausprägungen".

[100] In Abschn. 4c und 6b werden mit "Erwartungstreue" und "Konsistenz" Beispiele hierfür erwähnt.

[101] Genaugenommen nähert sich die mit der LA geschätzte Wahrscheinlichkeit nur asymptotisch dem Wert 0 bzw. 1.

Zu 3 (inhaltliche Aspekte): Ein Beispiel hierfür ist die Entscheidung für ein Streuungs- oder ein Konzentrationsmaß als beschreibende Statistik. Es gibt Leute, die meinen, beide Maße messen mehr oder weniger das Gleiche, nämlich eine Art "Ungleichheit". Wir gehen darauf im Anhang kurz ein.

Wir können hier keine weiteren Hinweise zur Methodenvielfalt geben, zumal dieser Text ja auch kein Lehrbuch sein soll, wollen aber noch eine Bemerkung zur veränderten Rolle der Statistiker machen.

> Man muss davon ausgehen, dass es den wenigsten Statistikanwendern (und auch nicht vielen Statistikern) vergönnt ist, auf vielen Gebieten auch die fortgeschritteneren Methoden zu kennen und zu verstehen. Andererseits können heutzutage immer mehr Menschen durch leicht zugängliche Statistik-Software auch bei geringer Vorbildung in Sachen Statistik sehr ausgefeilte Methoden der Statistik anwenden, in die man sich früher vielleicht gerade mal als Doktorand hinein vertiefte, und sie lernen solche komplizierten Methoden in erster Linie aus den Handbüchern zu ihrer Statistik Software kennen. Aus alle dem folgt, dass sich die Statistiker vielleicht verstärkt der Aufgabe widmen sollten, kurze und gut verständliche Einführungen in ihre Methoden zu verfassen.

6. Größe und Struktur einer Gesamtheit, und Schlüsse auf Basis von Stichproben

Wie wiederholt betont wurde – zuletzt im Zusammenhang mit "big data" – geht es in der Statistik nicht um einzelne Personen, Betriebe, Gemeinden usw. (allgemein Einheiten), sondern immer nur um Gesamtheiten von solchen Einheiten und dabei ist Umfang und Struktur der betreffenden Gesamtheit von entscheidender Bedeutung.

Das gilt auch für Stichproben: ein gleich großer Unterschied zwischen einer Stichprobengröße, wie $\bar{x}$ und dem entsprechenden Parameter μ der Grundgesamtheit $\bar{x} - \mu$ (oder auch $\hat{\beta}_k$ weil ja meist $\beta_k = 0$ geprüft wird) kann bei einem geringen Stichprobenumfang n nichtsignifikant, aber bei einem größeren Stichprobenumfang durchaus signifikant sein

Viele Verständnisprobleme, die wir mit Statistik haben hängen mit dem Fokus der Statistik auf *Gesamtheiten* und den aus ihnen gebildeten Teilgesamtheiten zusammen, wie

1. Auswahl einer Stichprobe aus einer Gesamtheit (die wir aktiv vornehmen) oder ein Ausscheiden von Einheiten aus einer solchen Gesamtheit (was wir nur passiv hinzunehmen haben),
2. die "Inferenz" (allgemein), bzw. speziell die "Induktion" als Schließen von einer Teilgesamtheit auf die Grundgesamtheit, aus der sie entnommen ist, wobei gerade der Stichprobenumfang n eine sehr wichtige "Stellschraube" ist, und
3. Zusammenhänge zwischen Statistiken (wie z.B. Mittelwerte, Korrelationskoeffizienten) die sich auf *Teilgesamtheiten* beziehen einerseits und solchen, die sich auf eine aus ihnen *aggregierte* größere *Gesamtheit* beziehen andererseits.

Die ersten beiden Punkte werden in diesem Abschn. 6 behandelt, und das dritte Problem ist das Thema von Abschn. 7, einem Abschnitt, der mehr einen Gegenstand der Deskriptiven Statistik behandelt.

a) "Repräsentativität", Zufallsauswahl, selection bias und survivor bias

Es gibt Teilgesamtheiten einer Grundgesamtheit, die nicht von uns gezogene Stichproben[102] darstellen, sondern irgendwie anders reduzierte Grundgesamtheiten darstellen. Für eine solche "Reduktion", die nicht wir absichtlich durch eine Auswahl vorgenommen haben, sondern die anders entstanden ist, gibt es verschiedene Gründe, wie

- *Nichtbeantwortung* (non-response), *oder Ausscheiden* (attrition) bei einer Wiederholungsbefragung (Panel), was eine Art Selbstselektion der Befragten ist, und
- *natürliches Ausscheiden* von Einheiten der Grundgesamtheit im Zeitablauf (Liquidation von Unternehmen etc.)[103] was

[102] Der Begriff ist hier allgemein gemeint. Wir werden später nur dann von Stichproben sprechen wenn sie nach dem Zufallsprinzip gezogen worden sind.

[103] Im Unterschied zur (weiteren) Teilnahmeverweigerung einer Einheit, die an sich zur Zielgesamtheit gehört wie im Fall von Panelmortalität (panel attrition) ist dies ein "unechtes" Ausscheiden, weil das liqui-

wir hier unter dem Stichwort "survivor bias" behandeln wollen.

Auch hier ist wieder das Begriffspaar "systematisch-zufällig" entscheidend. Das Ausscheiden oder die Nichtbeteiligung[104] von Einheiten, oder ganz allgemein das Fehlen von Daten ("missingness") kann

- systematisch, d.h. korreliert mit dem (den) Untersuchungsmerkmal(en), oder
- zufällig sein.

Bei erheblichen echten nonresponse Fällen und z.B. systematisch mehr Ausfällen bei älteren Personen (wenn das Alter ein Untersuchungsmerkmal ist) sind "Korrekturen" erforderlich[105]; denn andernfalls kann man argumentieren, dass wir eine *Verzerrung* (bias) haben und wir mit den Methoden der Stichprobentheorie nicht auf die Grundgesamtheit aller Einheiten, sondern faktisch nur auf die Gesamtheit der auskunftsbereiten Einheiten schließen, die evtl. ganz anders strukturiert sein kann als die Masse der Nichtauskunftsbereiten und damit auch als die Grundgesamtheit insgesamt.

Man kann nonresponse also auch einen Fall von self selection (der Befragten) auffassen: es ist nicht (mehr) die Stichprobe, die der Statistiker gezogen hatte. Die self selection ist z.B. systematisch beim "feedback effect": Die Teilnahmebereitschaft bei einer freiwilligen Meinungsbefragung ist abhängig davon, ob man zum Gegenstand der Befragung eine positive oder negative Meinung hat, und die Stichprobe kann dann (weil mehr Bereitschaft ist bei positiver Meinung) systematisch verzerrt sein kann zugunsten der positiven Meinung.

Angenehmer ist natürlich der Fall zufälliger Ausfälle oder von "missingness completely at random" (MCAR),[106] weil dann (gerade wegen der "Zufälligkeit") eine Verzerrung nicht eintritt. Zwar sind Schätzungen jetzt nicht per se systematisch verzerrt, wohl aber sind es faktisch Schätzungen auf Basis eines geringeren Stichprobenumfangs n, wobei n – wie auch gleich an der entsprechenden Formel gezeigt wird – entscheidend ist für den Stichprobenfehler, der *das* Maß für die Güte einer Stichprobe ist. Der Stichprobenumfang n ist auch relevanter als die response rate r. Bei n = 1000 und r = 0,35 ist der Umfang faktisch n* = 350 < n, was besser ist als ein größeres r = 0,40 bei n = 800 (dann ist der faktische Stichprobenumfang n* = 320).

Eine "sample selection bias"[107] kann auch auftreten, wenn man die "falschen" Einheiten auswählt oder auswählt aus einem "falschen" sampling frame wofür das "klassische" Beispiel ein massiver Prognosefehler bei der U.S. Präsidentschaftswahl 1948 war, wo man eine Telefonbefragung durchführte (Besitzer von Telefonen unterschieden sich seinerzeit noch systematisch von Nichtbesitzern) und den Sieg des republikanischen Kandidaten voraussagte, dann aber Truman (Demokrat) gewann.

Das alles zeigt, dass es bei einer Stichprobe[108] vor allem darauf ankommt, eine (potenzielle) Verzerrung zu vermeiden. Aus zwei Gründen lässt man bei der Auswahl den Zufall entscheiden

- um eine ***Verzerrung*** zu vermeiden, die entstünde, wenn man systematisch (in

dierte Unternehmen ja nicht mehr zur Zielgesamtheit gehört.

[104] Nur "echte" Nichtbeantwortung (z.B. Antwortverweigerung) verlangt Korrekturmaßnahmen wenn mit ihr eine Verzerrung der Ergebnisse droht. Vgl. auch im Anhang die Bemerkung zur "Hochrechnung".

[105] Wenn Schätzungen von Daten möglich sind und z.B. Differenzen zwischen (geschätzten) vollständigen und den erhobenen unvollständigen Daten bestehen kann man versuchen, die Auswirkung der "missingness" (auf Parameter bzgl. der Verteilung der interessierenden Variablen) abzuschätzen.

[106] Man unterscheidet MCAR und MNAR (missingness not at random) sowie MAR (at random) was in gewisser Weise zwischen MCAR und MNAR liegt. Auch bei Zufälligkeit, also MCAR ist eine missing data technique (MDT) angebracht und die Unterscheidung MCAR, MAR und MNAR ist jetzt dafür relevant, welche MDT angewendet werden sollte.

[107] Beispiel: Untersuchung des Zusammenhangs zwischen Verdienst Y und Dauer der Ausbildung X. Befragt man nur Berufstätige (B), ist anzunehmen dass B mit X und anderen Einflussfaktoren auf Y, z.B. Art der Tätigkeit, Position im Betrieb, Dauer der Betriebszugehörigkeit usw. die in der Störgröße zusammengefasst sind positiv korreliert ist, während dies alles bei Nichtberufstätigen nicht wirksam würde. Man könnte die Situation auch so deuten, dass B vergessen (omitted) wurde (es gibt keine Variation bezüglich B, weil nur Berufstätige betrachtet werden). und B andererseits X und Y beeinflusst.

[108] Wir haben ja auch im Zusammenhang mit "big data" festgestellt, dass trotz reichlich verfügbarer Daten weiter die Notwendigkeit besteht, Stichproben zu ziehen.

Abhängigkeit vom Untersuchungsgegenstand) nur ganz bestimmte Leute befragt ("Zufall" um eine Bevorzugung oder Benachteiligung zu vermeiden) und

- um sagen zu können, dass die ausgewählten Personen A, B und C "repräsentativ" für alle sind, die Ergebnisse also ***verallgemeinerungsfähig*** sind (hätte man A, B und C bewusst ausgewählt, dann würden die Ergebnisse nur für A, B und C gelten, hat man sie dagegen zufällig ausgewählt, dann gelten sie auch für andere Personen X, Y und Z, und zwar deshalb *weil der Zufall ja gerade darin besteht, dass es genauso gut sein kann, dass X, Y und Z, statt A, B und C ausgewählt werden.*

Was heißt aber, dass auch X, Y und Z "genauso gut" in die Auswahl gekommen wären? Das ist eine Frage nach der "Auswahlwahrscheinlichkeit" und der Grund dafür, dass "Zufallsauswahl" – zumindest bei einer "einfachen" (nicht geschichteten) Stichprobe definiert ist als "gleiche Auswahlwahrscheinlichkeit bei allen Einheiten der Grundgesamtheit".

Das ist etwas *ganz anderes als "willkürliche"*[109] Auswahl, die vorliegt im Falle des sog. "*convenience sample*" (oder synonym "accidental sample")[110], wo die bequeme Erreichbarkeit über die Auswahl entscheidet (typisches Beispiel: der Psychologie Professor bedient sich der Studenten seiner Vorlesung als Versuchspersonen).[111] Weil die Erreichbarkeit entscheidet ist die Auswahlwahrscheinlichkeit gerade *nicht* gleich, sie ist groß bei den Erreichbaren und null bei Nichterreichbaren.

Es gibt zwei sehr *verbreitete Missverständnisse* bei der Zufallsauswahl

- die Nichtunterscheidung von Zufall und Willkür (wobei unklar bleibt, welche Vorteile sich Statistiker vom Zufall versprechen) und
- der Fokus auf die Struktur von Stichprobe und Grundgesamtheit statt auf die Auswahlwahrscheinlichkeiten

Nach einer gängigen und hartnäckigen Vorstellung ist eine Stichprobe dann "repräsentativ", wenn sie ähnlich strukturiert ist wie die Grundgesamtheit. Weil ich mich andernorts [112] wiederholt mit der "***Repräsentativität***" auseinandergesetzt habe, kann ich mich hier kurz fassen:

1. Am einfachsten wird klar, dass dieses Konzept unbrauchbar ist, wenn man sich vorstellt, die Grundgesamtheit bestehe zu 50% aus Männern und zu 50% aus Frauen. Dann wäre eine Stichprobe von n = 8 Personen mit 4 Männern und 4 Frauen repräsentativ, aber eine von n = 1000 mit 508 Männern und nur 492 Frauen wäre es nicht (oder weniger)[113]
2. Dass A, B und C "repräsentativ" sind für alle, also auch für X, Y und Z, liegt nicht daran, dass dank A, B und C eine bestimmte Quote im gewünschtem Maße größer oder kleiner wird, sondern – wie gesagt – daran, dass statt A, B und C genauso gut X, Y und Z in die Auswahl hätten gelangen können.[114]

Die "Repräsentativität" R stellt nicht auf das ab, was eigentlich wichtig ist, nämlich

- der Stichprobenumfang n und die Homogenität der Grundgesamtheit (gemessen

[109] Auch die falsche, aber wohl "intuitiv" naheliegende Gleichsetzung von "zufällig" und "willkürlich" ist ein verbreitetes Hindernis, Statistik zu verstehen.

[110] Im der deutschen Literatur wird das "Auswahl aufs Geratewohl" oder "willkürliche Auswahl" genannt.

[111] Man konstruiert dann (bezeichnenderweise "after the fact") eine "Repräsentativität", nicht für die gesamte Bevölkerung der USA, sondern für den Teil, der hinsichtlich Alter, sowie ethnische und soziale Herkunft den typischen Studierenden entspricht. Hier liegen gleichzeitig mehrerer Missverständnisse der Zufallsauswahl vor, bei der ja u.a. gerade auch a priori (also vor Ziehung der Stichprobe) die Auswahlwahrscheinlichkeiten festliegen.

[112] besonders detailliert zuletzt in dem Text "Stichprobenumfang und Repräsentativität", der auch auf dieser Homepage zum Download zur Verfügung steht, bzw. als Diskussionsbeitrag Nr. 187 des Fachbereichs Wirtschaftswissenschaft der Universität Duisburg-Essen (Campus Essen).

[113] Wie sähe es bei einem quantitativen Merkmal X, etwa dem Einkommen aus? Soll man sich bei der Struktur in Gestalt von "Quoten" an einer Klasseneinteilung orientieren, wie 0 bis unter 500, 500 bis unter 1000 usw., oder muss es eine feinere Klasseneinteilung für die Quoten sein und wie fein muss sie sein?

[114] In diesem Sinne sind A, B und C tatsächlich "stellvertretend" auch für X, Y und Z. Sie wären es nicht, wenn z.B. die Auswahlchance von X, Y und Z größer oder geringer wäre als die von A, B und C.

an der Varianz σ^2 von X in der Grundgesamtheit), und
- die Wahrscheinlichkeit, in die Stichprobe zu gelangen,[115]

weil man sich nur an den "Quoten" orientiert und sie ist auch unbrauchbar, weil es kein Maß für R gibt.

Ganz anders ist es beim ***Stichprobenfehler*** (z.B. dem des Mittelwerts $\bar{x}$), an den man sich als Statistiker orientiert. Er ist definiert als $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ und man sieht, dass n, aber auch die Varianz σ_x^2 (oder einfach σ^2) von x eine Rolle spielen und man erhält bei gegebenem n und σ auch einen konkreten Zahlenwert für $\sigma_{\bar{x}}$.

Dagegen kann niemand ein exaktes Maß für die "Repräsentativität" R angeben. Es gibt keine Berechnungsformel, um zu bestimmen, wie groß R ist. Und dass die Varianz σ^2 relevant ist, sieht man wie folgt: angenommen alle Einheiten der Grundgesamtheit sind bezüglich X gleich; dann ist $\sigma^2 = 0$ und eine Stichprobe von n = 1 reicht aus,[116] um exakt (mit einem Stichprobenfehler von null) auf die Grundgesamtheit zu schließen (weil sich ja dann alle anderen Personen bezüglich X von der einen, einzigen ausgewählten Person gar nicht unterscheiden).

Hinzu kommt, dass wegen der *Zufalls*auswahl die Stichproben und damit auch die "Quoten" wegen des Zufalls ganz unterschiedlich ausfallen können, eine Stichprobe könnte danach "repräsentativer" sein als eine andere, obgleich beide gleichermaßen aus der gleichen Grundgesamtheit zufällig gezogen wurden und bei gleichem n und σ_x auch einen gleich großen Stichprobenfehler haben.

Wir haben es hier wieder mit dem *Fehler* zu tun, *aus einer Wahrscheinlichkeit etwas für eine konkrete Beobachtung folgern zu wollen* (ganz nach Art des gambler's mistake).

Der "Stichprobenfehler " ist, wie die Stichprobenverteilung (vgl. Abschn. 6b), aus der er abgeleitet ist, eine *Wahrscheinlichkeitsaussage* (was gerade durch die Zufälligkeit der Auswahl – und nur durch sie – möglich ist), also eine Aussage über *die Gesamtheit aller aus der gleichen Grundgesamtheit zu ziehenden Stichproben* gleichen Umfangs, und *nicht – wie die "Repräsentativität" – eine Aussage* über *eine einzelne konkrete Stichprobe*.

Gleichwohl steht im Denken vieler Menschen die Beurteilung einer Stichprobe an den Quoten hoch im Kurs. Man spricht in US Statistikbüchern auch von einem "representative sample", was bei uns als Quotenauswahlverfahren vor allem in der Markt- und Meinungsforschung bekannt und aus praktischen Gründen beliebt ist.[117]

Dadurch, dass bei diesem Verfahren die Frage, welche Leute im einzelnen befragt werden, sofern sie zur Erfüllung der Quote passend sind, nicht nach dem Zufall entschieden wird, kann ein bias durch Bevorzugung (z.B. wegen besserer Erreichbarkeit oderAntwortbereitschaft usw.) von Einheiten entstehen, und damit ein biased sample. Auf die Quotenauswahl ist also die Wahrscheinlichkeitsrechnung nicht anwendbar.

Wir kommen nun zu einer anderen Art von biased sample, wo die *Verzerrung nicht durch eine nichtzufällige Auswahl, sondern durch Strukturveränderungen* der Gesamtheit im Zeitablauf entstanden ist. Ein Beispiel hierfür ist der – bereits erwähnte – survivor bias, aufgrund des *systematischen* Einflusses unterschiedlich große Überlebensfähigkeit der Einheiten.

Ein Ignorieren der Survivor Bias ist der von W. Krämer kritisierte Schluss, Gemälde seien eine lohnende Geldanlage, weil Gemälde alter Meister stets hohe Preise in Auktionen erzielen, denn in den Auktionen erscheinen nicht die vielen mittelmäßigen Gemälde längst vergessener Maler, sondern bevorzugt die "besseren", so dass die "Überlebenden" nicht repräsentativ sind für *alle* (alten) Gemälde.[118]

[115] Ohne solche Auswahlwahrscheinlichkeiten benennen zu können, kann man auch keine Stichprobenverteilung, deren Standardabweichung der "Stichprobenfehler" ist kennen.

[116] So etwas ist bei der "Repräsentativität" nicht vorgesehen. Es gibt keine Quoten beim Nenner n = 1.

[117] Es hat also durchaus eine Existenzberechtigung, aber man kann mit ihm nicht Konfidenzintervalle und statistische Test durchführen, weil auf eine solche Auswahl die Wahrscheinlichkeitsrechnung nicht anwendbar ist. Das wird gerne übersehen bei Autoren, die wie Rebecca Warner und Timothy Urdan dem "representative sample" ähnliche Qualitäten zusprechen wie dem "random sample" (Zufallsauswahl).

[118] "…die Stichprobe ist deutlich zugunsten hoher Preissprünge verzerrt; genauso können wir auch zei-

Angenommen in einer Gesamtheit gibt es anfänglich zu gleichen Anteilen F_0 "Fitte" und U_0 "Unfitte" (zur Zeit 0 ist also $F_0 = U_0$) mit einem Variablenwert (z.B. Rentabilität) von $Y_F = 100$ und $Y_U = 50$, so dass $Y_F = 2Y_U$. Wie verändert sich mit der Zahl x der Perioden (z.B. Jahre) das Verhältnis F_x/U_x der Fitten zu den Unfitten und der durchschnittliche Wert von Y, den wir μ nennen wollen? Nehmen wir zur Vereinfachung konstante (für alle x gleiche) einjährige Sterbewahrscheinlichkeiten q und damit auch Überlebenswahrscheinlichkeiten p = 1 – q an; dann haben wir nach x Perioden (im Alter von x) $F_x = F_0(p_F)^x$ Fitte und $U_x = U_0(p_U)^x$ Unfitte, und der Anteil $\alpha_x = F_x/(F_x + U_x)$ der Fitten, der ursprünglich 50% ($\alpha_x = 0{,}5$) war, ist jetzt $\alpha_x = \frac{\beta^x}{1+\beta^x} = \frac{1}{1+1/\beta^x}$ und hängt nur ab vom konstanten Verhältnis $\beta = p_F/p_U$ der beiden Überlebenswahrscheinlichkeiten, nicht aber die Überlebenswahrscheinlichkeiten p_F und p_U selber. Wenn β nur einigermaßen groß ist, hat man schnell die Situation, dass sich die Überlebenden praktisch zu 100% aus den Fitten zusammensetzen.

Man erhält für α_x in Abhängigkeit von x und β die folgenden Werte

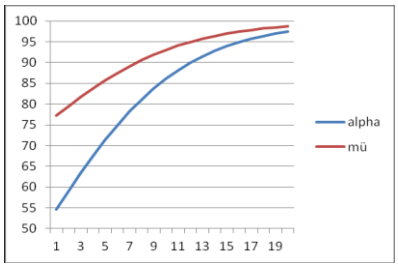
x	$\beta = 1{,}2$	$\beta = 1{,}5$	$\beta = 2$
0	½ = 0,5	0,5	0,5
1	0,54545	0,6	2/3 = 0,67
5	0,71333	0,88364	0,96969
10	0,86095	0,98295	0,99902

Ist die Überlebenswahrscheinlichkeit der Fitten um 20% höher als die der Unfitten besteht das Sample schon nach 5 Perioden zu über 70% nur noch aus den Fitten.

Betrachten wir nun das mittlere Y, im Zeitablauf, also $\mu = 100 - \frac{50}{\beta^x + 1} = 50(\alpha_x + 1)$, so dass man sich je nach der Größe von β ausgehend von $\mu = (100+50)/2 = 75$ bei x = 0 mehr oder weniger schnell an 100 annähert.

	α_x (in%) und $\Rightarrow$ μ	
x	$\beta = 1{,}2$	$\beta = 1{,}5$
0	50 $\Rightarrow$ 75	50 $\Rightarrow$ 75
1	54,54 $\Rightarrow$77,2727	60 $\Rightarrow$ 80
5	71,333 $\Rightarrow$85,6667	88,364 $\Rightarrow$ 94,182
10	86,095 $\Rightarrow$93,0475	98,295 $\Rightarrow$ 9,91475

oder graphisch veranschaulicht (dort ist $\beta = 1{,}2$)



Die Veränderung in Form eines größer werdenden μ ist eine "unechte", d.h. strukturell bedingte (durch Zu- und Abgänge zwischen den Erhebungszeitpunkten, wobei wir es hier nur mit Abgängen von Unfitten haben) Veränderung. Aus solchen Gründen kann z.B. die Rentabilität in einer Branche allein durch das Ausscheiden der Grenzanbieter" zunehmen.[119]

Weil "echte" und "unechte" Veränderungen, oft schwer oder gar nicht zu unterscheiden sind kann es nützlich sein, eine (aufwändige) Panelbefragung (der gleichen Einheiten im Zeitablauf) durchzuführen, im Unterschied zu einer bloßen Zeitreihenbetrachtung ("repeated observations") mit einer sich laufend ändernden Gesamtheit von jeweils anwesenden Einheiten.

b) Stichprobenverteilung und Likelihoodfunktion als Mittel der Inferenz

Wie kann man von einer Stichprobe und ihren Kennzahlen (statistics)[120] wie z.B. einem arithmetischen Mittel $\bar{x}$ oder einem Anteil p auf die entsprechende Größe μ bzw. π in der Grundgesamtheit (Größen, die wir allgemein "Parameter" nennen wollen) schließen? [121] Wir wollen hierzu kurz zwei oft in der Statistik angewendete und sehr grundlegende "Denkmuster" darstellen:

- man nimmt die Werte für die unbekannten Parameter (oder den Wert für den unbekannten Parameter) an, bei denen die konkrete (die mit ihr berechnete statistic) am wahrscheinlichsten ist (***Maximum Likelihood [ML] Methode***), und

gen (wir fragen nur die Gewinner), dass Lottoscheine gute Kapitalanlagen sind" (W. Krämer, So lügt man mit Statistik, Neuausgabe 2011, S. 111).

[119] Mit $x = \ln(9)/\ln(\beta)$ kann man auch leicht ausrechnen, bei welchem x die Gesamtheit schon zu 90% nur noch aus den Fitten besteht: bei $\beta = 1{,}2$ ist das erst bei x = 12,05 erreicht bei $\beta = 2$ (doppelt so hohe Überlebenswahrscheinlichkeit) aber schon bei x = 3,17.

[120] auch Schätz- oder Stichprobenfunktion genannt.

[121] Ich wüsste nicht, wie man die Frage beantworten könnte, wenn man keine (Zufalls-) Stichprobe hätte und deshalb auch die Wahrscheinlichkeitsrechnung nicht anwenden könnte.

- man untersucht die Verteilung einer statistic (p, $\bar{x}$), die sich ergibt wenn man *alle* überhaupt *möglichen Stichproben* gleichen Umfangs n, aus einer Grundgesamtheit mit π bzw. μ ziehen würde (das ist das Konzept der ***Stichprobenverteilung*** oder sampling distribution).

Was die ML Methode betrifft, so stelle man sich vor, man habe eine Stichprobe von n = 3 gezogen und dabei x = 1 Männer und n-x = 2 Frauen gezählt, also einen Anteil p = x/n = 1/3. Man kann sich das vorstellen als Ziehen aus einer Urne (*mit* Zurücklegen, so dass die Urne praktisch unendlich groß ist) mit blauen (Männer) und roten (Frauen) Kugeln interpretieren. Aus welcher der folgenden vier Urnen (Grundgesamtheiten) mit jeweils drei Kugeln[122] könnte man die gegebene Stichprobe O O O (wo p =1/3 ist) mit welcher Wahrscheinlichkeit gezogen haben?

Urne 1	Urne 2	Urne 3	Urne 4
O O O	O O O	O O O	O O O
π = 0	π = 1/3	π = 2/3	π = 1

π ist der Anteil der blauen Kugeln in der Urne

Man sieht sofort, dass es Urne 1 und 4 nicht sein können, aus der die Stichprobe gezogen worden ist, denn wir haben eine blaue und zwei rote Kugeln gezogen und Urne 1 enthält keine blaue Kugel (woher also die eine blaue Kugel in der Stichprobe?) und Urne 4 keine roten Kugeln (wir haben aber zwei roten Kugeln gezogen.

Bei den beiden anderen Urnen ist es prinzipiell möglich, die Stichprobe zu erhalten. Da wir *mit* Zurücklegen ziehen kann man natürlich auch dann zwei blaue (oder zwei rote) Kugeln ziehen, wenn die Urne jeweils nur eine blaue bzw. rote Kugel enthält. Die entsprechenden Wahrscheinlichkeiten errechnen sich mit der Likelihood Funktion[123]

$$\lambda = \binom{n}{x}\pi^x(1-\pi)^{n-x} = \binom{3}{1}\pi^1(1-\pi)^2 = 3\pi(1\text{-}\pi)^2.$$

Man erhält damit die folgenden Werte für λ:

Urne 1	Urne 2	Urne 3	Urne 4
λ = 0	λ = 4/9 = 0,44	λ = 2/9 = 0,22	λ = 0

wobei wir den Wert von jeweils 0 bei Urne 1 und 4 bereits erklärt haben. Man beachte, dass die Summe von 4/9 = 0,444 und 2/9 = 0,222 kleiner als 1 ist. Wir haben ja auch über Likelihoods (für im Prinzip beliebig viele Werte von π) addiert, nicht über Wahrscheinlichkeiten für x = 0, 1, …, n bei gegebenem n und λ.

Es ist unschwer zu sehen, dass die Likelihood Funktion ihr Maximum bei π = 1/3 hat: Aus $\frac{d\lambda}{d\pi} = 3 - 12\pi + 9\pi^2 = 0$ oder 1-π(4-3π) = 0 folgt π = 1/3. Bei π = 1/3 ist also die gegebene Stichprobe am wahrscheinlichsten. Es erscheint damit sinnvoll, den Wert p = $\hat{\pi}$ = 1/3 der Stichprobe als Schätzwert für π zu nehmen, es ist der Maximum Likelihood (ML) -Schätzer" für π.

Für die Schätzfunktion p = $\hat{\pi}$ = x/n (Anteil der "Erfolge" [im Beispiel der blauen Kugeln] bei einer Anzahl x von "Erfolgen") kann man auch das Konzept der Stichprobenverteilung demonstrieren. Bei Stichproben vom Umfang n, ist die ***Stichprobenverteilung***

- *von x* die Binomialverteilung mit E(X) = nπ und der Varianz $\sigma_x^2 = n\pi(1\text{-}\pi)$,
- *von p = x/n* die relativierte Binomialverteilung mit E(p) = E(X)/n = π und der Varianz $\sigma_p^2 = \sigma_x^2/n^2 = \pi(1\text{-}\pi)/n$.

Bei Stichproben vom Umfang n = 10 aus einer Grundgesamtheit mit π =1/3 erhält man mit der Formel $\binom{10}{x}\pi^x(1-\pi)^{n-x}$

x	p = x/n	Wahrsch.
0	0	0,0173
1	0,1	0,0867
2	0,2	0,1951
3	0,3	0,2601
4	0,4	0,4552
5	0,5	0,1366
usw. bis x = 10		

Bei 1,73% der Stichproben ist p = 0, bei 19,51% ist es p = 0,2. Es gibt Stichproben, bei denen p < π = 1/3 (etwa p = 0,2), aber auch solche, bei denen p größer ist, als π, wie z.B 0,4.

[122] Dass es nur drei sind ist nicht notwendig, sondern soll nur der Verständlichkeit dienen.

[123] Die Formel ist bekannt als *Wahrscheinlichkeits*-funktion f der Binomialverteilung. Es ist f(x|n,π) wenn π gegeben ist und wir die Wahrscheinlichkeit bei n und x = 0, 1,…, n suchen. Es ist eine *Likelihood*-funktion λ(π|n,x), wenn n und x gegeben ist und wir π variieren, mit 0 ≤ π ≤ 1. Man spricht von "Likelihood" statt "Wahrscheinlichkeit", weil π, im Unterschied zu x, keine Zufallsvariable ist und man nur bei einer Zufallsvariable von einer Wahrscheinlichkeit sprechen kann. Die oben zitierte Definition der Likelihood als bedingte Wahrscheinlichkeit P()

Likelihood = P(evidence | hypothesis)

gilt auch jetzt in Gestalt von

P(Stichprobe | angenommener Wert eines Parameters).

eine Verteilung mit $E(p) = \pi = 1/3$ und $\sigma_p^2 = \pi(1-\pi)/n = 0{,}222/n = 0{,}0222$.

Dass wir hier bei x = 0 nicht eine Wahrscheinlichkeit von null (sondern 0,0173) haben mag überraschen. Aber oben haben wir gefragt: kann man eine (x = 1) blaue Kugel aus einer Urne ziehen, in der es keine blauen Kugeln gibt (π = 0)? Jetzt ist aber die Frage, kann man x = 0 blaue Kugeln aus einer Urne ziehen, in der der Anteil blauer Kugeln $\pi = 1/3$ ist? Oben war es auch genauer gesagt eine Likelihood.

Es ist interessant zu sehen, dass

- π und p Variablen von ganz unterschiedlicher Art sind: die *Quote* π (etwa der Anteil blauer Kugeln in der Grundgesamtheit) hat hier einen und nur einen Wert, nämlich 1/3, aber p (die entsprechende Quote in der Stichprobe) ist eine *Zufalls*variable, von der wir nur wissen, dass sie *verschiedene Werte annehmen* kann, aber nicht welchen bei einer konkreten Stichprobe (wir kennen nur die *Wahrscheinlichkeiten* für die möglichen Wert von p);
- es keinen Grund gibt, warum $p = \pi$ sein sollte; die einzelnen Werte für p können sehr verschieden von π sein,[124] aber "im Mittel" ist p gleich 1/3, denn $E(p) = \pi$ (man sagt, π wird mit p "erwartungstreu" geschätzt),[125]
- es bei der Varianz $\sigma_p^2 = \pi(1-\pi)/n$ der Stichprobenverteilung von p auf n ankommt: Sie wird bei größerem n kleiner (sie ist 0,0222 bei n = 10 und 0,222/30 = 0,007407 bei n =30), d.h. auch, dass der Stichprobenfehler σ_p mit wachsendem n immer kleiner wird.[126]

Bei n = 30 und $\pi = 1/3$ hätten wir die folgende Stichprobenverteilung von x bzw. p:

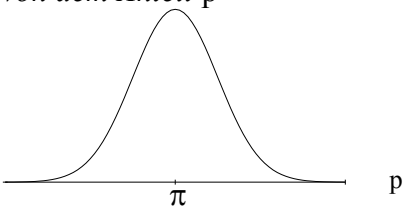
x	p = x/n	Wahrscheinl.
0	0	0,0000052
3	0,1	0,0026466
6	0,2	0,0483842
usw. bis x = 30 (p = 1)		

usw.

Bei x = 10 hätten wir p = 10/30 = 1/3 und hier müsste entsprechend die Wahrscheinlichkeit am größten sein. In der Tat ist die Wahrscheinlichkeit auch bei x = 9 (also p = 0,3) 0,1457, bei x = 10 (also p = 1/3) 0,1530 und dann bei x = 11 wieder kleiner, nämlich 0,1391.8

Nach. den sog. Grenzwertsätzen nähert sich mit $n \to \infty$ die *Stichprobenverteilung*

- *von dem Anteil* p



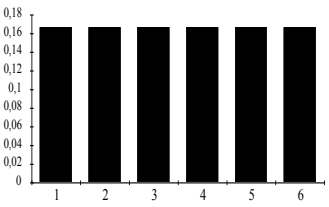
der Normalverteilung mit mit $E(p) = \pi$ und der Standardabweichung

$\sigma_p = = \sqrt{\frac{\pi(1-\pi)}{n}}$ (das wäre dann auch der Stichprobenfehler von p als Schätzwert für π)

- *von dem arithmetischen Mittel* $\bar{x}$ der Normalverteilung mit $E(\bar{x}) = \mu$ und der Standardabweichung (das ist der oben bereits genannte Stichprobenfehler von $\bar{x}$)

(8) $\sigma_{\bar{x}} = = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$, wobei μ der Mittel- bzw. Erwartungswert und σ^2 die Varianz von X in der Grundgesamtheit ist.

Ein verbreitetes Missverständnis ist die Behauptung, dass x in der Grundgesamtheit normalverteilt sein müsse. Ist $\bar{x}$ die durchschnittliche Augenzahl beim n-maligem Werfen eines Würfels), dann ist die Zufallsvariable X in der Grundgesamtheit wie folgt verteilt



mit dem Erwartungswert (wir haben hier ja eine Zufallsvariable und keine endliche Grundgesamtheit und deshalb keinen Mittelwert von $\mu = 3{,}5$

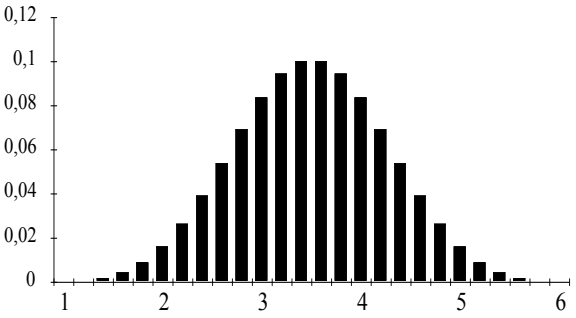
und der Varianz von $\sigma^2 = 13/6 = 2{,}167$ (und damit der Standardabweichung von $\sigma = 1{,}472$), so dass von Normalverteilung keine Rede sein kann.

[124] Bei keiner ist p genau gleich $\pi = 1/3$. Streng genommen wäre hier also keine einzige Stichprobe "repräsentativ". Wie man sieht verkennt der übliche Begriff der "Repräsentativität" auch ganz, dass p und π Variablen ganz unterschiedlicher Art sind. Es ist allerdings zuzugeben, dass entsprechende Vorstellungen von einer "repräsentativen" Stichprobe der sog. "Hochrechnung" zugrundeliegen (mehr dazu unter den Ergänzungen im Abschn. 8).

[125] Es gilt $E(p) = \pi$ bei *jedem* (also auch einem kleinen) n und nicht erst bei $n \to \infty$ (dann wäre sie nur *asymptotisch* erwartungstreu).

[126] Die Schätzung von π mit p ist "konsistent", d.h. man nähert sich, ganz im Sinne des "gesunden Menschenverstands", umso mehr dem wahren Wert je größer die Stichprobe ist.

Zieht man n = 5 mal aus dieser Grundgesamtheit mit $\mu = 3{,}5$ und $\sigma = 1{,}472$ (d.h. wirft man 5 mal mit einem Würfel, oder einmal mit 5 Würfeln), was man natürlich beliebig oft machen kann,[127] und bildet man jeweils die mittlere Augenzahl $\bar{x}$, dann ist die Stichprobenverteilung von $\bar{x}$ bei n = 5:



Das sieht schon etwas nach einer Normalverteilung aus und wir haben hier $\mu = 3{,}5$ und $\sigma_{\bar{x}} = \sqrt{2{,}167/5} = 0{,}658$.

Das Konzept der Stichprobenverteilung ist von einem nicht zu überschätzenden Wert. Es erlaubt uns von der Grundgesamtheit auf die Stichprobe (genauer: die Gesamtheit der möglichen Stichproben) zu schließen. Wir haben hier

1. *zwei Verteilungen*, bei denen jeweils x an der Abszisse steht, nämlich die
   - ♦ Verteilung von x in der Grundgesamtheit (siehe oben) mit $\mu = 3{,}5$ und die
   - ♦ Verteilung von x in einer Stichprobe, die z.B. bei n = 5 so aussehen könnte:

x	2	3	4	5	damit ist
h_x	1/5	1/5	2/5	1/5	$\bar{x} = 3{,}6$

   mit h_x als relativer Häufigkeit

2. *eine Verteilung*, bei der $\bar{x}$ an der Abszisse steht, nämlich die Stichprobenverteilung von $\bar{x}$; sie ist sozusagen das Bindeglied zwischen den beiden unter 1 genannten Verteilungen und erst mit ihr ist es möglich, ein Konfidenzintervall für μ zu schätzen oder – was damit eng verwandt ist – eine Hypothese über μ zu "testen" (wir kennen nur $\bar{x}$, nicht aber μ und deshalb kann es über μ nur Vermutungen, also Hypothesen geben).

[127] deshalb haben wir es beim folgenden Bild ja auch mit einer Wahrscheinlichkeitsverteilung zu tun.

Bei der Normalverteilung ist z.B. die Wahrscheinlichkeit 95,45%, ein x im Bereich von $\mu \pm 2\sigma$ zu erhalten. Das bedeutet, angewendet auf die Stichprobenverteilung: wenn $\bar{x}$ bei n = 5 normalverteilt ist[128] mit $\mu = 3{,}5$ und $\sigma_{\bar{x}} = 0{,}6583$, dann kann man bei einer Stichprobe mit $\bar{x} = 3{,}6$ mit 95,45% Wahrscheinlichkeit[129] vermuten, dass μ in den Grenzen $\bar{x} \pm 2 \cdot \sigma_{\bar{x}}$ liegt, also zwischen $3{,}6 - 2 \cdot 0{,}6583 = 2{,}28$ und $3{,}6 + 2 \cdot 0{,}6583 = 4{,}92$.

Weil ein hypothetisch angenommener Wert von $\mu = \mu_0 = 4$ noch innerhalb dieses Konfidenzintervalls liegt, würde man diese Hypothese annehmen (= nicht ablehnen), aber z.B. die Hypothese $\mu = 5$ ablehnen (= verwerfen), weil 5 außerhalb des Intervalls liegt, also ein Wert von 5 oder mehr nicht sehr wahrscheinlich wäre.

H_0: $\mu = \mu_0$

Konfidenzintervall (KI)	Testentscheidung
μ_0 liegt innerhalb der Grenzen des KI	H_0 annehmen (= "nicht signifikant")
μ_0 liegt außerhalb der Grenzen des KI	H_0 ablehnen (= " signifikant")

Zwischen der Betrachtung von Anteilen p (wie oben bei der Erklärung der ML-Methode) und von Mittelwerten besteht folgender Zusammenhang: wenn es nur zwei Ausprägungen gibt, etwa blaue und rote Kugeln und man setzt x = 1 bei blau und x = 0 bei rot, dann ist der Mittelwert $\Sigma x_i/n = p$, weil dann nämlich die Summe Σx_i die Anzahl der blauen Kugeln darstellt.

Noch kurz zu *"Likelihood" und "Wahrscheinlichkeit"*: Wir haben bei der Frage, aus welcher Urne die Stichprobe gezogen sein könnte, den Parameter π variiert bei gegebener Stichprobe (n = 3, x = 1) und bei der Stichprobenverteilung (was eine Wahrscheinlichkeitsverteilung ist) umgekehrt x bei gegebenem π variiert:

	gegeben	variiert
Likelihood	x	π
Stichprobenverteilung	π	x

[128] Von einer guten Approximation kann man allerdings erst ab n = 30 sprechen. Wir nehmen hier nur den Fall n = 5, weil wir dies auch schon vorher (insbesondere bei der letzten Abbildung) getan haben.

[129] wir haben die krumme Zahl von 95,45% nur genommen, weil man dann mit 2 leichter rechnen kann. Üblicher wäre 95% (und damit eine Irrtumswahrscheinlichkeit [= ein "Signifikanzniveau]) von 5%.

n ist stets gegeben. X ist eine Zufallsvariable, π nicht. Die Terminologie ist hier also voll im Einklang mit den im Zusammenhang mit dem Bayesschen Theorem eingeführten Begriffen.

c) Signifikanz, power und Effektstärke

Es ist stets zu bedenken, dass man es beim Schätzen und Testen, wie bei der "dahinter stehenden" Stichprobenverteilung mit Wahrscheinlichkeitsaussagen zu tun hat, die sich nicht auf eine konkrete Stichprobe beziehen, sondern auf die Gesamtheit der möglichen Stichproben vom Umfang n, die man aus der gleichen Grundgesamtheit ziehen könnte. In jedem Fall (Schätzen und Testen) ist und bleibt die Grundgesamtheit unbekannt, weil wir aus ihr ja nur eine Stichprobe gezogen haben und alle Überlegungen drehen sich immer nur um die Frage, wie *wahrscheinlich* eine Stichprobe, wie die gezogene ist, bzw. *wäre* (bei einer hypothetisch angenommenen Grundgesamtheit).

Viele glauben, man habe bei Ablehnung der Nullhypothese H_0 (also der Entscheidung "signifikant") gezeigt, dass die H_0 falsch ist. Aber

> Mit einem Test wird nicht *festgestellt*, ob eine Hypothese H richtig oder falsch *ist*, sondern es wird *entschieden*, ob man sie für richtig oder falsch *halten soll*, und

- es geht also nicht um Verifikation oder Falsifikation, sondern um eine Entscheidung über eine Hypothese und
- man stützt sich bei dieser Entscheidung auf den Zahlenwert T*, den eine Teststatistik T annimmt und die Wahrscheinlichkeit für $T \geq T^*$,
- aber dazu braucht man eine Stichprobenverteilung und eine exakt formulierte H, die Nullhypothese H_0,[130]
- H_0 besagt i.d.R. "nur Zufall", "kein systematischer Einfluss",
- wenn man H_0 ablehnt, hat man damit nur gezeigt, dass das Stichprobenergebnis *bei Geltung von H_0* wenig wahrscheinlich *wäre*, also mehr als nur Zufall im Spiel sein könnte

  Lehnt man bei der Regressionsgerade $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ die Hypothese H_0: $\beta = 0$ ab ($\hat{\beta}$ ist "signifikant"), könnte mit x ein systematischer Einfluss auf y gegeben sein (es könnten aber auch neben x ganz andere systematische Einflüsse im Spiel sein, die sogar relevanter sein können als x)[131]

- Dass es beim statistischen Test um eine *Entscheidung* geht und nicht darum, festzustellen, ob eine Hypothese H richtig oder falsch ist, sieht man auch daran, dass es zwei Arten von Fehlern gibt:

  1. die richtige Hypothese wird verworfen (Fehler erster Art oder α- Fehler) und
  2. die falsche Hypothese wird angenommen (Fehler zweiter Art oder β- Fehler),

  während es bei einer *Feststellung* nur einen Fehler gibt, dass nämlich das, was man festzustellen glaubte, nicht zutrifft, also "falsch" ist). Hinzu kommt, dass der Fehler erster Art und der der Fehler zweiter Art nicht zusammenhanglos nebeneinander stehen und dass in der Praxis die beiden Fehler (hinsichtlich ihrer Konsequenzen) als unterschiedlich gravierend bewertet werden können.

Wenn die Alternativhypothese H_1 als logischer Gegensatz zur Nullhypothese H_0 formuliert ist, etwa H_0: $\mu = \mu_0 = 100$ (oder bei zwei Stichproben $\mu_1 - \mu_2 = 0$ d.h. kein Effekt) und H_1: $\mu \neq 100$ (bzw. $\mu_1 - \mu_2 \neq 0$ Effekt vorhanden) bedeutet Ablehnung von H_0 ipso facto Annahme von H_1 und umgekehrt.

H_0: $\mu = \mu_0$

Hypothesen	
H_0: nur Zufall, kein Effekt	H_1 systematischer Einfluss (Effekt)
Testentscheidung	
H_0 ablehnen = H_1 annehmen H_1 ablehnen = H_0 annehmen	

Was die Entscheidung faktisch bedeutet hängt davon ab, welche Hypothese richtig (und damit auch welche falsch) ist.

[130] Ein Test ist ein "als-ob" Verfahren: man tut so als ob H_0 richtig wäre und fragt sich wie wahrscheinlich dann die vorliegende Stichprobe wäre.

[131] Für einen solchen Fall findet man gerade in diesem Abschn. 6c ein Beispiel: wovon hing es ab, ob man den Untergang der Titanic überlebte oder nicht?

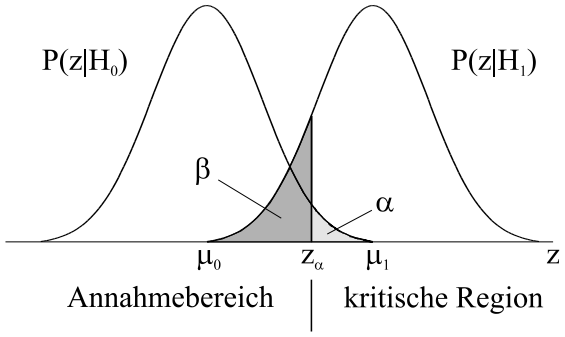
Weil der Fehler 1. Art (auch "α-Fehler", weil er höchstens mit der vorgegebenen Wahrscheinlichkeit α begangen wird) auch Annahme der falschen H_1 bedeutet, wird er auch als "false claim" bezeichnet. Entsprechend leuchtet es ein, dass der Fehler 2. Art (β-Fehler) im Nichterkennen eines Unterschieds (Effekts) besteht (der Effekt besteht ja, weil H_1 richtig ist).

state of nature → action ↓	H_0 ablehnen H_1 annehmen	H_0 annehmen H_1 ablehnen
H_0 richtig = H_1 falsch	α-Fehler false claim	1 - α (kein Fehler)*
H_0 falsch = H_1 richtig	1- β power*	β-Fehler missed effect

* es gibt zwei Arten von Fehlern, aber auch zwei Arten, etwas richtig zu machen; dabei ist 1-β ein wertvolleres Ergebnis als 1-α: es bedeutet, einen bestehenden Einfluss (Effekt) nachgewiesen zu haben (während 1-a "absence of evidence" bedeutet, was gerne als evidence of absence missverstanden wird).

Die Wahrscheinlichkeit 1 - β ist bekannt als Macht (***power***) eines Tests, oder in Deutsch auch "Trennschärfe" oder Teststärke genannt, und sie ein Maß für die Fähigkeit einen (in der Grundgesamtheit) bestehenden Unterschied auch zu erkennen (bzw. eine falsche H_0 abzulehnen). Es ist wichtig, zu sehen:

Es gibt nicht ***die*** power: 1- β hängt ab von n, von der Wahl von α und davon, wie unterschiedlich H_0 und H_1 sind. Faktisch ist davon aber nur n ist gestaltbar.



Nur wenn die beiden Glockenkurven zu einer "zusammenfallen" ist β = 1-α, also 1-β = α. In dem Maße, in dem sie sich voneinander entfernen wird 1-β größer.

Mit Festlegung von α (üblich sind 10%, 5% oder 1%) liegt der "kritische Wert" (kW), was z_α in der Abbildung[132] entspricht (im folgenden Zahlenbeispiel wäre $\bar{x}_\alpha$ statt z_α richtig für kW) fest, der den Annahmebereiche (gemeint ist Annahme von H_0) vom Ablehnungsbereich trennt.

Wie groß dann die power 1-β ist hängt auch davon ab, wie unterschiedlich H_0 und H_1 sind (wie groß also der Abstand zwischen den beiden Glockenkurven ist). Es ist klar, dass ein größerer Unterschied auch leichter als solcher erkannt wird.

Das folgende Rechenbeispiel macht die Zusammenhänge deutlich. Wir rechnen[133] mit H_0: μ = 100, σ = 30 sowie H_1: μ > 100 (einseitiger Test), σ = 30 und nehmen unterschiedliche Werte für α, n und den Unterschied zwischen H_0 und H_1 an:

α = 0,05 (5%, einseitig)

n	kW	power bei (H_1: μ = μ_1)		
		μ_1 = 105	μ_1 = 110	μ_1 = 115
50	106,98	0,319	0,761	0,970
100	104,93	0,500	0,954	0,996
200	103,49	0,758	0,999	≈ 1

α = 0,01 (1%, einseitig)

n	kW	power bei		
		μ_1 = 105	μ_1 = 110	μ_1 = 115
50	109,87	0,425	0,501	0,587
100	106,98	0,455	0,843	0,996
200	104,93	0,501	0,991	≈ 1

Der Stichprobenumfang n ist also eine ganz einflussreiche Stellschraube: will man einen "Effekt" nachweisen, so muss man n groß wählen:[134] Wie groß $\bar{x}$ sein muss, um den Wert für signifikant größer als 100 zu erklären hängt vom Signifikanzniveau α und vom Stichprobenumfang n ab. Bei α = 0,01 und n = 50 ist erst ein $\bar{x}$ > 109,87 signifikant, aber bei n = 200 ist es schon nur 104,93.

Und bei einem Signifikanzniveau von 5% (statt 1%) wären es schon die Werte > 106,98 und > 103,49.

Kleine Unterschiede sind also schon dann signifikant, wenn α und n groß sind.

[132] Z ist hier großgeschrieben, weil es eine Zufallsvariable ist. Es gibt zwei Stichprobenverteilungen für die Prüfgröße Z, einmal bei Geltung von H_0 und dann bei Geltung von H_1, wo μ = μ_1 angenommen wird.

[133] Einzelheiten der Rechnung mögen hier nicht interessieren. Es kommt mehr darauf an, zu erkennen, wovon kW und die power abhängen.

[134] Bei einem kleinem n wird das nicht gelingen, weil sich bei einer kleinen Stichprobe der Zufall stärker auswirken kann.

Die power 1- β eines Tests hängt auch ab von α. Es gibt eine Art *trade-off zwischen α und β* und damit auch zwischen α und 1-β. Man könnte die power 1-β auch dadurch größer (und β kleiner) machen, dass man ein größeres α (etwa 10% statt 5%) in Kauf nimmt. Das würde dann aber auch die Wahrscheinlichkeit für false claims vergrößern. Wie sehr α und β zusammenhängen wird auch wie folgt deutlich:

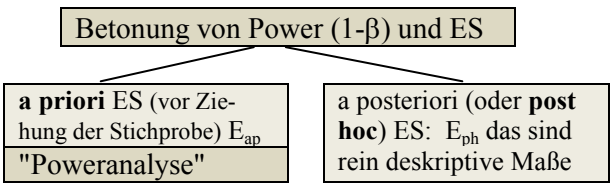
	H_0 ablehnen (= F_0 nicht heiraten)	H_0 annehmen (= F_0 heiraten)
H_0 richtig: F_0 ist die "richtige" Frau	α-Fehler: die richtige Frau nicht heiraten	
H_0 falsch: F_0 ist die "falsche" Frau	power	β-Fehler: die falsche Frau heiraten

Dem "ewigen Junggesellen" mag es gelingen, den Fehler 2. Art (β) zu vermeiden, er wird damit aber wahrscheinlich öfter den Fehler 1. Art begehen und dies später als alter Mann bereuen. Entsprechend wäre es auch nicht rational, um jeden Preis den Fehler 1. Art zu vermeiden und sich zu früh zu binden. Die power 1- β bestünde hier darin, die "falsche" Frau F_0 zu meiden, was ja auch heißt, zu erkennen, dass F_1 die "Richtige" ist und F_1 zu heiraten.

Man kann geltend machen dass es nicht (oder nicht allein) die bloße ja/nein Entscheidung eines Hypothesentests ist, die primär von Interesse sein sollte, sondern

- die *power* 1-β und die
- ***Effektstärke*** (Effektgröße; effect size)

Für die stärkere Beachtung der Effektstärke (ES), der power und auch des Konfidenzintervalls haben v.a. die Bücher von Ellis und Cumming[135] plädiert und sich damit offenbar weitgehend durchgesetzt:



Bei der (sehr in Mode gekommenen) sog. Poweranalyse geht es darum, den Stichprobenumfang n unter Berücksichtigung der folgenden Größen 1 bis 3 zu planen (es geht um vier Größen die alle eng zusammenhängen, in der Weise, dass je drei von ihnen die vierte bestimmen):

1. das Signifikanzniveau α (auch "Testniveau")
2. E_{ap} die a priori Effektstärke (im Unterschied zur post hoc Effektstärke E_{ph},
3. 1 - β, die power, und
4. n, der Stichprobenumfang.

Was die power betrifft so haben wir gezeigt, dass sie entscheidend von der Unterschiedlichkeit zwischen H_0 und H_1 abhängt

Im Zahlenbeispiel haben wir gesehen, dass man – was auch sehr plausibel ist – bei gegebenem n und α einen großen Unterschied zwischen H_0 und H_1 eher erkennt als einen kleinen Unterschied. Bei α = 5% und n = 200 erkennt man den relativ großen Unterschied von μ_1 = 115 und μ_0 100 mit fast 100% Wahrscheinlichkeit, den kleineren zwischen 105 und 100 aber nur mit einer Wahrscheinlichkeit von ca. 76%;

Die Unterschiedlichkeit von H_0: $\mu = \mu_0$ und H_1: $\mu = \mu_1$ wird als a priori Effektstärke ins Spiel gebracht mit $E_{ap} = \frac{|\mu_0 - \mu_1|}{\sigma}$, oder ähnlich konstruierten Größen, wobei jedoch zu beachten ist, dass

- alle Größen in E_{ap} Annahmen über die Grundgesamtheit,[136] bzw. Setzungen darstellen, was man für "praktisch signifikant" halte will, und dass
- es meist leicht ist für H_0 einen konkreten Wert anzunehmen, etwa bei der Regression von y auf x H_0: β = 0 (kein Einfluss von x) vs. H_1: β ≠ 0 (Einfluss von x), es aber kaum möglich ist, auch für H_1 einen exakten Wert anzugeben; denn mit welcher Begründung sollte man H_1: β = 0,8 oder β = 1,3 annehmen?

Man hat also bei E_{ap} als einer maßgeblichen Größe im Rahmen der Poweranalyse wenig "Objektives" in der Hand. Nun zur nicht weniger problematischen *Effektstärke post hoc* E_{ph} als Maß zur Beurteilung eines Stichprobenergebnisses, worum es ja bei dem ganzen Streit um die Relevanz der Testentscheidung[137] geht:

[135] vgl. Fußnote 81, Seite 25.

[136] Es heißt, dass man zwischen einer statistical und einer practical significance unterscheiden müsse (und die Testentscheidung nur eine statistical significance wiederspiegle) und dass $|\mu_0 - \mu_1|$ Ausdruck dessen sei, was man als "praktisch signifikant" ansehen möchte. Das ist aber naturgemäß etwas "Geschmackssache".

[137] Das Ergebnis, dass etwa β_k "signifikant " ist oder nicht ist ja auch eine Aussage über ein Stichprobener-

Es ist zwar richtig, dass auch dann, wenn in zwei Fällen gleichermaßen H_0 abgelehnt wird, also ein Effekt (d.h. mehr als nur Zufall) gegeben sein dürfte, dieser Effekt quantitativ gleichwohl von sehr unterschiedlicher "Bedeutung" sein könnte. Aber daraus folgt noch nicht, wie das Maß Eph beschafffen sein soll mit dem man diesen Effekt quantifizieren will.

Wir zeigen das im Folgenden am Beispiel des Untergangs der Titanic (das wir dem Buch von Ellis entnommen haben), wonach es für das Überleben der Schiffskatastrophe eine Rolle spielte in welcher (Schiffs-) Klasse man sich als Passagier befand:

	died (D)	sur-vived	Σ
first class (C)	122	203	325
third class	528	178	706
Σ	650	381	1031

Die t_{n-2} verteilte Prüfgröße t für den t-Test für zwei unabhängige Stichproben ("tea for two") ist $t = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, wobei

$a = p_1 = P(D|C) = 122/325 = 0{,}38$ [138]

$b = p_2 = P(D|\overline{C}) = 528/706 = 0{,}75$ und

$p = 650/1031$ $n_1 = 325$ und $n_2 = 706$ ist.

Man erhält t = 11,5 und die χ_1^2 verteilte[139] Prüfgröße $\chi^2 = t^2 = 132{,}5$. Das Ergebnis ist hochsignifikant bei $\alpha = 1\%$ (nach dem t und dem [äquivalenten] χ^2 Test).

Das Beispiel zeigt auch, dass wir mit einem Test der H_0 nur gezeigt haben, dass mehr als nur Zufall im Spiel sein könnte und dass Geld (Klassenzugehörigkeit C) eine Rolle spielen könnte. Das heißt aber nicht, dass nicht auch ganz andere systematische Einflüsse als C im Spiel sein könnten, die auch sogar im höheren Maße relevant sein könnten als C. Genau das dürfte hier der Fall sein.

Mit F = Frauen und Kinder und $\overline{F}$ = Männer erhält man.

	died (D)	sur-vived	Σ
Frauen/Kinder(F)	145	249	394
Männer	505	132	637
Σ	650	381	1031

und jetzt ist $p_1 = P(D|F) = 0{,}37$, $p_2 = P(D|\overline{F}) = 0{,}79$ und t = -13,7 was auch wieder hochsignifikant (1%) wäre. Aber hinsichtlich der Testentscheidung wird kein Unterschied gemacht und der größere absolute Wert 13,7 vs. 11,5 wird nicht weiter beachtet.[140]

Es ist deshalb verständlich, dass man nach einem Maß sucht, dass zum Ausdruck bringt, dass vielleicht Geschlecht und Alter (Kinder!) bei der Titanic wichtiger für das Überleben waren als der Umstand, ob man eine teuere oder billigere Klasse gebucht hat. Und das soll die Effektstärke E_{ph} messen. Es gibt jedoch sehr viele Maße der ES[141] Sie beruhen

- auf Differenzen zwischen bzw. Quotienten von bedingten Wahrscheinlichkeiten (**d-Maße**) – etwa wie oben – p_2-p_1= 0,79 – 0,37 oder auf
- Assoziations-, Kontingenz- oder Korrelationskoeffizienten (**r-Maße**).

d-Maße angewendet auf das Titanic Beispiel:

E_{ph} von C	E_{ph} von F*
relative risk	
$\frac{r_1}{r_2} = \frac{P(D\|C)}{P(D\|\overline{C})} = \frac{0{,}38}{0{,}75} = 1{,}97$	2,15
odds ratio	
$\frac{r_1/(1-r_1)}{r_2/(1-r_2)} = \frac{2{,}97}{0{,}60} = 4{,}95$	$\frac{3{,}83}{0{,}58} = 6{,}57$
Vierfelderkorrelation (Assoziation)	
$\phi_{CD} = -0{,}3585$	$\phi_{FD} = -0{,}4276$

* jeweils analog definiert mit F statt C

Nach diesen drei E_{ph}-Maße von Typ d-Maße, ist der Effekt von F bezüglich sterben/überleben (also bezüglich D) größer als der von C.

gebnis, bloß nach Meinung von Ellis und Cumming usw. keine sehr aussagefähige.

[138] mit a, b, c, d nehmen wir Bezug auf die Formel für die Vierfelderkorrelation ϕ (siehe Seite 19). Man beachte, dass der Quotient ad/bc die odds ratio (OR; Verhältnis der odds) darstellt; denn wie man leicht sieht, sind die odds für die Gruppe X = 1 a/b und die für die andere Gruppe (X = 0) c/d. Gern benutzt wird als ES Maß auch log(OR).

[139] χ^2 verteilt mit einem Freiheitsgrad.

[140] Wie man sieht kann die gleiche Testentscheidung eine ganz unterschiedliche post hoc Effektstärke (E_{ph}) bedeuten. Es mag also schon sinnvoll sein auch ein Maß der E_{ph} mitzuteilen. Aus dieser Einsicht aber gleich – wie erwähnt – eine "New Statistics" (Fußnote 81) zu machen dürfte jedoch sehr übertrieben sein.

[141] nach Ellis hat man über 70 solche Maße gezählt.

Wir kommen auf das Beispiel noch einmal zurück indem wir zeigen, dass die Korrelation (bzw. Assoziation weil es sich hier um 0-1 Variablen handelt) zwischen D und C also $\phi_{CD} \neq 0$ nicht um eine Scheinkorrelation handelt.

Was E_{ph} betrifft ist zu bedenken

1. es gibt viele E_{ph} Maße und es ist nicht immer klar, wie man E_{ph} messen soll, wenn man z.B. bei nichtparametrischen Tests mit Rangsummen arbeitet,
2. dass für E_{ph} (*nach* Ziehung der Stichprobe) Maße der deskriptiven Statistik benutzt werden, bei denen an keiner Stelle erkennbar wird, dass wir von einer Stichprobe vom Umfang n auf eine Grundgesamtheit vom Umfang N schließen. [142]

"New Statistics" bedeutet, dass man sich bei der Beschreibung des empirischen Befunds sowohl bestimmter Elemente der induktiven als auch der deskriptiven Statistik bedient :

	ist abhängig von
1. Testentscheidung*	n (Stichprobenumfang) und α (Signifikanzniveau)
2. power und Effektstärke a priori (E_{ap})	wie 1 *und* davon, wie unterschiedlich H_0 und H_1 sind (z.B. von $\lvert\mu_0-\mu_1\rvert$), also E_{ap}, was eine Einschätzung der "praktischen" Signifikanz bedeutet und oft schwer zu begründen ist (etwa β = 0,3 statt nur $\beta \neq 0$)
3. Effektstärke E_{ph}	nur den in der Stichprobe festgestellten relativen Häufigkeiten

* und Konfidenzintervall:
Grüne Felder betreffen den Schluss von der Stichprobe auf die Grundgesamtheit; beim violetten Feld geht es nur um die Stichprobe.

Mit Maßen der Effektstärke (post hoc) haben wir uns praktisch in das Gebiet der deskriptiven Statistik begeben, in dem ein Datensatz (Stichprobe oder Totalerhebung) mit geeigneten Kennzahlen ("Maßen") z.B. für die Messung der Streuung, Konzentration, des Zusammenhangs usw. beschrieben wird. Auf Probleme. geeignete Maße zu konstruieren gehen wir kurz im Anhang ein. Im folgenden Abschnitt 7 geht es nur um Aggregationsprobleme bei solchen Maßen.

[142] Wir haben oben nur Maße für dichotome Variablen (wie sie beim Titanic-Beispiel allein relevant sind) vorgestellt. Es gibt auch Maße für E_{ph}, bei metrisch skalierten Merkmalen, wie die (mit einer "gepoolten" Standardabweichung standardisierte) Differenz zwischen zwei arithmetischen Mitteln oder der Korrelationskoeffizient

7. Aggregationsprobleme und sog. "Paradoxien"

Deskriptive Maße wie Mittelwerte, "Quoten", Korrelationskoeffizienten usw. können sich auf Teilgesamtheiten oder auf durch Aggregation über Teilgesamtheiten gebildete übergeordnete größere Gesamtheiten beziehen. Intuitiv erwartet man oft, dass das Maß für die größere Gesamtheiten ein Mittel aus den entsprechenden Maßen für die Teilgesamtheiten ist. Das ist nicht immer so und darauf beruhen einige "Paradoxien", die in Wahrheit oft keine Widersprüchlichkeiten sind, sondern sich leicht erklären lassen.

a) Will-Rogers Paradoxon [143]

Will Rogers (1879 - 1935), ein amerikanischer Entertainer und Schauspieler, behauptete dass der Umzug von Einwohnern Oklahomas (X) nach Kalifornien (Y), also X → Y in *beiden* Staaten die durchschnittliche Intelligenz erhöht habe.[144] Dazu ein Zahlenbeispiel mit drei bzw. vier Personen wovon eine den Standort wechselt

vorher	
Y	X
2000	4000
3000	6000
4000	8000
	5000
Mittelwerte	
3000	5750

→

nachher	
Y	X
2000	4000
3000	6000
4000	8000
5000	
Mittelwerte	
3500	6000

Offensichtlich ist 3500 > 3000 und 6000 > 5750. Der Grund für die Paradoxie, dass sich Mittelwert bei X *und* Y erhöht hat ist, dass das Mittel der in X gebliebenen n_{x1} Einheiten $\bar{x}_{01}$ (im Beispiel 6000) größer ist als das Mittel der n_{x2} aus X abgewanderten Einheiten $\bar{x}_{02}$ (im Beispiel 5000), und $\bar{x}_{02}$ größer ist das Mittel der schon n_{y1} in Y vorhandenen Einheiten $\bar{y}_{01}$ (im Beispiel 3000):

$$\bar{y}_{01} < \bar{x}_{02} < \bar{x}_{01}.$$

Jeder Wert für $\bar{x}_{02}$ zwischen $\bar{y}_{01}$ = 3000 und $\bar{x}_{01}$ = 6000 hätte also den hier interessieren-

[143] Den Hinweis auf die Sache und das Beispiel – bei dem es offensichtlich nicht um IQs geht – verdanke ich einer im Internet verfügbaren pdf-Datei (vom 27. 4. 2012) von Klaus Dürrschnabel (Hochschule Karlsruhe f. Technik und Wirtschaft).

[144] Die naive Vermutung wäre: wenn der Durchschnitt in einem Staat gestiegen ist, muss er im anderen gefallen sein (weil es ja die gleichen Leute sind, die jetzt in X fehlen und in Y hinzugekommen sind).

den Effekt erzeugt.[145] Man kann die Ungleichung umkehren, also von $\bar{y}_{01} > \bar{x}_{02} > \bar{x}_{01}$ ausgehen, etwa mit (8000 > 5000 > 4000) und erhält so ein *Sinken* des Mittelwerts bei X *und* Y, also den gegenteiligen Effekt:

vorher	
Y	X
8000	4000
	5000
Mittelwerte	
8000	4500

→

nachher	
Y	X
8000	4000
5000	
Mittelwerte	
6500	4000

Fazit: Es gibt hier nicht notwendig so etwas wie kommunizierende Röhren, wonach ein Mittel ab- und ein anderes zunehmen muss. Das wäre nur dann der Fall, wenn wir zwei Ungleichungen hätten, etwa $\bar{y}_{01} < \bar{x}_{02}$ aber $\bar{x}_{02} > \bar{x}_{01}$ wie in der folgenden Situation

vorher	
Y	X
2000	4000
	5000
Mittelwerte	
2000	4500

→

nachher	
Y	X
2000	4000
5000	
Mittelwerte	
3500	4000

Hier nimmt $\bar{y}$ zu, aber $\bar{x}$ ab; und bei $\bar{y}_{01} > \bar{x}_{02}$ aber $\bar{x}_{02} < \bar{x}_{01}$ nimmt $\bar{y}$ ab und $\bar{x}$ zu.

Das Will Rogers Paradoxon setzt also eine bestimmte Ungleichung voraus. Es ist kein allgemeines Phänomen.

b) Simpson Paradoxon und Scheinkorrelation

Unter diesem Namen finden sich in der Literatur sehr viele Darstellungen,[146] die auf dem ersten Blick verschieden aussehen, aber trotzdem letzten Endes auf das Gleiche hinauslaufen. Im Grundsatz liegt ein Aggregationsproblem vor, bzw. wir haben es mit einer nicht beim Zusammenhang zwischen X und Y explizit beachteten dritten Variable (auch covariate oder confounder) Z zu tun, was die Sache in eine Verwandtschaft mit der Scheinkorrelation rückt.

Betont man den Aggregationsaspekt, so läuft es darauf hinaus, dass für das Gesamtaggregat (mit den $n_1 + n_2$ Einheiten) bezüglich des Zusammenhangs (der Korrelation) zwischen X und Y etwas anderes gelten kann als für die Teilaggregate 1 (mit n_1 Einheiten) und 2 (mit n_2 Einheiten).[147] So wird das Simpson Paradoxon häufig – und auch im Folgenden – erklärt.

Einfacher als mit Korrelationen bzw. Assoziationen dürfte es sein, das Paradoxon im Falle von Verhältniszahlen, wie z.B. Todesraten zu erklären, wie das im folgenden Abschn. 7c.

Im folgenden Beispiel betrachten wir den Zusammenhang zwischen Rauchen und Überleben einer Krebsbehandlung über eine bestimmte Anzahl von Jahren. Es sei R = (Anzahl der) Raucher, NR = Nichtraucher, Ü = Überlebende, NÜ = Nichtüberlebende und man habe zwei Gruppen (jüngere = Alter x zwischen 55 und 64, ältere mit $65 \le x \le 74$) und die folgenden Vierfeldertafeln

1. Jüngere		
	R	NR
NÜ	51	40
Ü	64	81

2. Ältere		
	R	NR
NÜ	29	101
Ü	7	28

Bei den Jüngeren und den Älteren ist der Anteil der Überlebenden (Ü) unter den R geringer (55% denn 64/(51+64) = 0,5565 und 19,4% also 7/36) als bei den NR (dort 66,9% und 21,7%): Raucher (R) überleben also seltener als NR. Rauchen (R) ist mit NÜ "assoziiert" und NR mit Überleben Ü. Für die Vierfelderkorrelationen ϕ erhält man bei den Jüngeren $\phi = +\,0{,}1159$ und bei den Älteren $\phi = +\,0{,}0228$. Legt man nun die Tafeln zusammen, so erhält man für *alle* Personen ($55 \le x \le 74$) *zusammen* die folgende Tafel 3

3. Jüngere *und* Ältere		
	R	NR
NÜ	80	141
Ü	71	109
$\phi = -\,0{,}033315$		

47% der Raucher überleben, aber bei den Nichtrauchern sind es weniger, nämlich 44% und die

[145] Auf die Anzahl der beteiligten Einheiten, also den n_{x1} und n_{y1} nichtwandernden Personen und den n_{x2} wandernden Personen kommt es nicht an, so dass man auch einfach $n_{x1} = n_{y1} = n_{x2} = 1$ annehmen darf.

[146] Die Sache ist sehr viel bekannter als das eben erwähnte (und mir früher auch nicht bekannte) Paradoxon von Will-Rogers. Für das (fiktive) Zahlenbeispiel verwenden wir wieder die im Internet zu findende Datei von K. Dürrschnabel, auf die wir uns auch schon beim Will-Rogers Paradoxon gestützt haben.

[147] Es besteht ein Unterschied zwischen einer Untersuchung mit k Aggregaten mit jeweils n_1, n_2, n_k Einheiten und einer Untersuchung mit den einzelnen n Individuen ($n = n_1 + n_2 + \ldots + n_k$). Das Problem ist unter dem Namen *ecological fallacy* bekannt und das Simpson Paradoxon gilt als Spezialfall hiervon. Wir gehen darauf in Abschn. 7d ein.

Vierfelderkorrelation ist jetzt negativ, so dass R mit NÜ statt mit Ü assoziiert ist.

Man kann diese Umkehrung der Verhältnisse, also diese "Paradoxie" auch mit einer dritten dichotomen (zwei Ausprägungen jung/alt oder J/A) Variable "Alter" erklären. Der Zusammenhang: R/NR ↔ Ü/NÜ entsteht, weil J/A und R/NR assoziiert sind und andererseits aber auch J/A mit Ü/NÜ, wie die folgenden beiden Vierfeldertafeln zeigen

4. J/A ↔ R/NR

	R	NR
J	115	121
A	36	129
ϕ = + 0,2733		

5. J/A ↔ Ü/NÜ

	Ü	NÜ
J	145	91
A	35	130
ϕ = + 0,3980		

48,7% (denn 115/(115+121) = 0,487) der Jüngeren sind Raucher aber nur 21,8% der Älteren und entsprechend überleben 61,4% der Jüngeren aber nur 21,2% der Älteren. Sowohl R/NR als auch Ü/NÜ sind positiv korreliert mit dem Alter.[148]

Dass es auf die sich in Tab. 4 und 5 ausdrückende Korrelationen ankommt, wird deutlich, wenn man einmal annimmt wir hätten mit dem Faktor k ≠ 0 die folgende bei den Jüngeren und Älteren gleiche Struktur:

1a. Jüngere

	R	NR
NÜ	51	40
Ü	64	81

2a. Ältere

	R	NR
NÜ	k*51	k*40
Ü	k*64	k*28

Man hätte jetzt in Tab. 1a und 2a und in der entsprechend veränderten Tab. 3 für das Gesamtaggregat die gleiche Vierfelderkorrelation in Höhe von ϕ = + 0,1159. Das Simpson Paradoxon könnte also nicht Platz greifen. Zugleich erhielte man aber auch anstelle der Tab. 4 und 5 die folgenden Vierfeldertafeln:

4a. J/A ↔ R/NR

	R	NR
J	115	121
A	k*115	k*121

5a. J/A ↔ Ü/NÜ

	Ü	NÜ
J	145	91
A	k*145	k*91

und es ist unschwer zu sehen, dass dies bedeutet, dass die Korrelationen, zwischen dem Alter (J/A) und sowohl R/NR als auch Ü/NÜ verschwinden, man also bei 4a und 5a jeweils ϕ = 0 erhält.

Die Situation beim Simpson Paradoxon ist damit nicht prinzipiell völlig verschieden von der einer Scheinkorrelation. Dort korrelieren zwei Variablen (so, wie hier R/NR mit Ü/NÜ) miteinander, und zwar *nur* deshalb, weil sie beide mit einer dritten Variable (hier wäre es J/A) korrelieren.

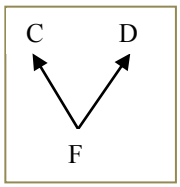
Bei Scheinkorrelation verschwinden die partiellen Assoziationen.[149] Man kann auch umgekehrt Assoziationen in den Teilgesamtheiten aber keine Assoziation insgesamt haben:

8	4
10	5
insgesamt ϕ = 0	

7	2
5	4
Gruppe 1 ϕ = 0,236	

1	2
5	2
Gruppe 2 ϕ = - 0.5	

Wir kommen noch einmal auf das Titanic Beispiel zurück: Ist die Korrelation (bzw. Assoziation) zwischen dem Überleben D und der Schiffsklasse C also ϕ_{CD} = -0,3585 ≠ 0 nicht vielleicht eine Scheinkorrelation?



Es könnte sein, dass das Geschlecht nicht nur mit dem Überleben (D) sondern auch mit der Schiffklasse (C) korreliert, weil in der billigen dritten Klasse viele Männer waren.

Der Zusammenhang war aber gering $\phi_{FC} = \phi_{CF}$ = 0,1108 und ϕ_{CD} = - 0,4276. Läge eine Scheinkorrelation zwischen C und D über F vor, dann müsste $\phi_{CF}\phi_{FD} \approx \phi_{CD}$ sein. Das Produkt ist aber 0,0474 während ϕ_{CD} = -0,3585 ist. Es müssten dazu auch die partiellen Assoziationen zwischen C und D jeweils um Null herum liegen. Die ϕ_{CD} sind aber partiell (speziell) bei den Männern - 0,1799 und bei Frauen und Kindern - 0,5550.

Es scheint also einen echten (kausalen) Zusammenhang zwischen Geld (Schiffsklasse) und Überleben D gegeben zu haben, auch wenn der Faktor Geschlecht (Frauen/Kinder vs. Männer) für D vielleicht bedeutsamer gewesen sein mag (eine größere Effektstärke hatte).

Dazu passt auch dass die relative risks r_1/r_2 bei Frauen und Kindern mit 0,046 und Männern mit 0,805 sehr unterschiedlich waren:

[148] Dabei kann die Korrelation zwischen Alter und Überleben auch gar nicht mit dem Rauchen zusammenhängen, sondern einfach nur Ausdruck der allgemeinen Verschlechterung des Gesundheitszustands mit zunehmendem Alter sein.

[149] Im obigen Beispiel verschwinden sie in den Teilgesamtheiten J und A zwar nicht, aber haben nur ein anderes Vorzeichen als im Gesamtaggregat.

$$\frac{P(D|CF)}{P(D|\overline{C}F)} = \frac{4/150}{141/244} = \frac{0{,}03}{0{,}58} = 0{,}046 \text{ und}$$

$$\frac{P(D|C\overline{F})}{P(D|\overline{C}\overline{F})} = \frac{118/175}{387/462} = \frac{0{,}67}{0{,}84} = 0{,}805.$$

Das Risiko umzukommen war also *bei beiden Geschlechtern* für first-class Passagiere geringer als für third-class Passagiere (bei Frauen mit 0,03 zu 0,58 noch ausgeprägter als bei Männern mit 0,67 zu 0,84).

c) *Simpson Paradoxon und Strukturunterschiede*

Viel bekannter als mit Korrelationen (in der Gesamtheit einerseits und ihren Teilgesamtheiten andererseits) ist das Simpson Paradoxon im Falle von Verhältniszahlen (Raten, Quoten usw.), wo das Paradoxon dann wohl auch deutlich leichter zu verstehen ist.

Ein klassisches Beispiel, das in der einschlägigen Literatur immer wieder zitiert wird, ist die Klage gegen die Universität von Berkley wegen Benachteiligung von Frauen bei der Zulassung zum Studium im Jahr 1974. Der Anteil Q der abgewiesenen Bewerbungen um einen Studienplatz war bei den Frauen größer als bei den Männern ($Q_F > Q_M$). Dabei zeigte sich aber, dass sich die Studentinnen überdurchschnittlich stark an solchen Fachbereichen (z.B. den geisteswissenschaftlichen) um eine Zulassung bemühten, an denen generell viele Bewerber abgewiesen wurden. Die fachbereichsspezifischen Ablehnungsquoten waren bei den Männern sogar durchwegs größer, so dass $Q_{iM} > Q_{iF}$ für alle i = 1,…, n Fachbereiche galt.[150]

Ein Unterschied in der Gesamtheit kann auch dann auftreten, wenn die entsprechenden Quoten in den Teilgesamtheiten alle gleich sind wie der folgende fiktive Vergleich der Sterberaten (oder Todesraten) von Priestern und Bergarbeitern zeigt. Dabei bedeutet J Jüngere (z.B. unter 50 Jahre, also x < 50) und Ä Ältere (≥ 50), ferner D Gestorbene und L Lebende (jeweils in dem Jahr, auf das sich die Statistik bezieht, und in der betreffenden Altersklasse)

1. Priester		
	L	D
J	1000	10
Ä	9000	540
Σ	10000	550

2. Bergarbeiter		
	L	D
J	9000	90
Ä	1000	60
Σ	10000	150

Die "rohe" (als Mittel über alle Altersklassen) Sterberate der Priester ist mit 550/10000 = 0,055 geringer als die der Bergarbeiter, die nur 0,015 beträgt, was aber nicht heißt, dass Bergarbeiter der gesündere Beruf ist. Wie man sieht, sind die altersspezifischen Sterberaten bei beiden Berufen gleich 1% bei den jüngeren und 6% bei den Älteren. Der Unterschied in der rohen Todesrate m = D/L kommt allein durch die Altersstruktur zustande:[151] jung sind nur 10% der Priester aber 90% der Bergarbeiter und entsprechend alt 90% und 10%. Somit ist m offenbar ein gewogenes arithmetisches Mittel

$m_P = 0{,}055 = 0{,}1 \cdot 0{,}01 + 0{,}9 \cdot 0{,}06$ (Priester)

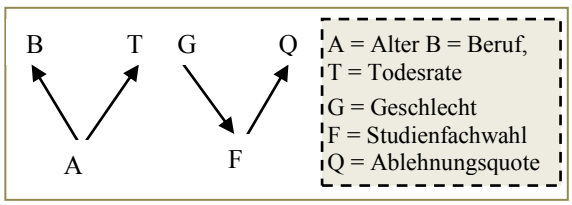
$m_B = 0{,}015 = 0{,}9 \cdot 0{,}01 + 0{,}1 \cdot 0{,}06$ (Bergarbeiter).[152]

Generell ist eine Verhältniszahl V = Z/N ein gewogenes Mittel der speziellen Verhältniszahlen $V_i = Z_i/N_i$ mit $Z = \Sigma Z_i$ und $N = \Sigma N_i$ und zwar

- ein arithmetisches Mittel mit den Gewichten N_i/N oder ein
- harmonisches Mittel mit den Gewichten Z_i/Z.

Wir gehen auf weitere Zusammenhänge noch einmal kurz im Anhang (S. 51unten) ein.

Man kann die Beispiele mit der höheren Ablehnungsquote von Frauen in Berkley und der höheren Sterblichkeit der Bergarbeiter auch als Scheinkorrelationen interpretieren:



[150] Auch hier ist natürlich eine störende dritte Variable zwischen X (Geschlecht) und Y (Zulassung) im Spiel, nämlich F, die Art des Fachbereichs. Es besteht ein Zusammenhang zwischen F und Y (Fachbereiche mit relativ vielen/wenigen Ablehnungen) und auch einer zwischen X und F (Frauen bewerben sich mehr an Fachbereichen mit relativ vielen Ablehnungen).

[151] Auf der gleichen Ebene steht die Feststellung eines besonders hohen Sterberisikos von Studenten ("the most dangerous profession", "lowest average age of death" Nach einer von der Columbia Univ. New York im Internet zitierten Studie von 1685).

[152] Wie man leicht sieht, sind bei zwei gleichen altersspezifischen Todesraten (1% und 6%) die rohen Todesraten m_P und m_B nur dann gleich, wenn auch die Altersstruktur gleich ist.

d) Individuen oder Gesamtheiten als Beobachtungseinheiten: "ecological fallacy"

Unter der ecological fallacy (EF) versteht man die in der Regel unzutreffende Erwartung, dass die Variablen X und Y genauso korrelieren müssten, wenn sich die Daten statt auf Individuen (k = 1,…, n) auf I zusammenfassende Einheiten (z.B. Städte o.ä., daher "ökologisch" = räumlich) beziehen.[153] Wir gehen im Folgenden von Städten und ihren Einwohnern aus.

Angenommen die i-te Stadt (i = 1,…,I) hat n_i Einwohner und die Mittelwerte bezüglich X und Y seien, $\bar{x}_i$ und $\bar{y}_i$, dann sind die Gesamtmittel $\bar{x} = \sum_i n_i \bar{x}_i / n$ mit $n = \sum_i n_i$ und $\bar{y} = \sum_i n_i \bar{y}_i / n$. Für die n-fache "ökologische" Kovarianz (E steht im Folgenden für "ecological", was *zwischen* [between] den Städten bedeutet) erhält man dann

$$E_{xy} = \sum_i n_i (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) = \\ = \sum_i n_i \bar{x}_i \bar{y}_i - n \cdot \bar{x} \cdot \bar{y},$$

und für die (n-fachen)Varianzen

$$E_{xx} = \sum_i n_i (\bar{x}_i - \bar{x})^2 = \sum_i n_i \bar{x}_i^2 - (\bar{x})^2$$

und analog $E_{yy} = \sum_i n_i (\bar{y}_i - \bar{y})^2$.

Dann ist die auf der individuellen Ebene berechnete Kovarianz[154] innerhalb (within) der ökologischen Einheiten

$$W_{xy} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)$$

$= \sum_i \sum_j x_{ij} y_{ij} - \sum_i n_i \bar{x}_i \bar{y}_i$ und für die totale Kovarianz gilt $T_{xy} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})(y_{ij} - \bar{y}) = \sum_i \sum_j x_{ij} y_{ij} - n \cdot \bar{x} \cdot \bar{y}$. Man sieht, dass gilt $T_{xy} = E_{xy} + W_{xy}$ und entsprechende Zusammenhänge gelten auch für die Varianzen, so dass man für die Korrelationen erhält

$\rho_T = \dfrac{T_{xy}}{\sqrt{T_{xx} T_{yy}}}$ (auf Basis individueller Daten) und

$\rho_E = \dfrac{C^E_{xy}}{\sqrt{V^E_{xx} V^E_{yy}}}$ (auf Basis der Daten für die Städte).

Mit der analog definierten Korrelation ρ_W (within) erhält man dann als "Endergebnis"

$$(9)\quad \rho_T = \rho_E \sqrt{\frac{E_{xx} E_{yy}}{T_{xx} T_{yy}}} + \rho_W \sqrt{\frac{W_{xx} W_{yy}}{T_{xx} T_{yy}}} \\ = e\rho_E + w\rho_W$$

womit ρ_T eine Linearkombination aus ρ_E und ρ_W ist, nicht aber einfach ein gewogenes arithmetisches Mittel weil sich die Gewichte e und w nicht zu 1 addieren (e + w ≠ 1). Aus Gl.9 wird auch klar, dass ρ_E und ρ_T sehr wohl ein unterschiedliches Vorzeichen haben können, so dass man sich mit ρ_E statt ρ_T auch in der Richtung des Zusammenhangs (X und Y verändern sich gleichsinnig oder gegensinnig) irren kann.

Wie man sieht wird der ökologische Fehler u.a. dann *nicht* begangen, d.h. es ist $\rho_E = \rho_T$, wenn W_{xx} und/oder W_{yy} null sind, was ja bedeutet, dass sich die Einwohner einer Stadt hinsichtlich X und/oder Y nicht unterscheiden. Anders gesagt, eine Ursache für die ecological fallacy ist die Streuung der Variablen *innerhalb* der Städte und es leuchtet unmittelbar ein, dass man ohne diese Streuung mit den Daten der I Städte (also mit ρ_E) das gleiche Ergebnis erhielte, wie mit den Daten der n Einwohnern (also mit ρ_T).

Dass sich ρ_E und ρ_T unterscheiden liegt auch an ρ_W. Wie das zu verstehen ist, wird deutlich wenn man beachtet, dass ρ_E und ρ_W selbst wieder Linearkombinationen (auch hier wieder mit Gewichten, die sich nicht zu 1 addieren) von Korrelationen darstellen, was wir gleich im Fall von nur I = 2 Städten leicht sehen werden.

Wir wollen nun zeigen, dass die ecological fallacy mit dem Simpson Paradoxon verwandt ist (im Sinne von Abschnitt 7b). Dabei geht es um den Zusammenhang zwischen ρ_W und ρ_T

---

[153] "the correlation between individual variables is deduced from the correlation of the variables collected for the group to which the individuals belong" (Piamtados / Byar / Green). Wir gehen davon aus, dass die I Städte eine Zerlegung (partition) darstellen, d.h. eine Einheit (ein Individuum) gehört zu einer und nur einer Stadt und es gibt keine Einheiten, die zu keiner der I Städte gehören. Mit "Aggregaten" müssen nicht notwendig räumlich ("ökologische") Einheiten wie z.B. Gemeinden gemeint sein.

[154] Zur Vereinfachung erwähnen wir nicht weiter den Zusatz "n-fach" bzw." n_i fach".

und nicht mehr, wie bei der ecological fallacy um den Zusammenhang zwischen ρ_E und ρ_T. Nehmen wir dazu ohne Beschränkung der Allgemeinheit nur I = 2 Städte (Teilgesamtheiten) an, dann gilt

$$E_{xy} = n_1(\bar{x}_1 - \bar{x})(\bar{y}_1 - \bar{y}) + n_2(\bar{x}_2 - \bar{x})(\bar{y}_2 - \bar{y})$$
$$= E_{xy}^{(1)} + E_{xy}^{(2)}$$

für die Gesamtheit bzw. die beiden Städte getrennt. Entsprechend gilt für die Situation innerhalb der beiden Städte

$$W_{xy} = \sum_{j=1}^{n_i}(x_{1j} - \bar{x}_1)(y_{1j} - \bar{y}_1) + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)(y_{2j} - \bar{y}_2) = W_{xy}^{(1)} + W_{xy}^{(2)}$$

Auch hier gelten wieder $T_{xy} = E_{xy} + W_{xy}$ und die entsprechenden Formeln für die Varianzen E_{xx}, E_{yy}, W_{xx} und W_{yy} die ebenfalls jeweils Summen darstellen, bezogen auf die beiden Städte, z.B. $E_{xx} = E_{xx}^{(1)} + E_{xx}^{(2)}$ analog $W_{xx} = W_{xx}^{(1)} + W_{xx}^{(2)}$.

Mit den Korrelationen innerhalb der beiden Städten (mit n_1 und n_2 Einwohnern)

$$\rho_w^{(1)} = \frac{W_{xy}^{(1)}}{\sqrt{W_{xx}^{(1)}W_{yy}^{(1)}}} \text{ und } \rho_w^{(2)} = \frac{W_{xy}^{(2)}}{\sqrt{W_{xx}^{(2)}W_{yy}^{(2)}}}$$

erhält man eine zusammengefasste "*interne*" Korrelation als Linearkombination der Korrelationen *innerhalb* der beiden Städte

$$(10)\quad \rho_w = \frac{\sqrt{W_{xx}^{(1)}W_{yy}^{(1)}}}{\sqrt{W_{xx}W_{yy}}}\rho_w^{(1)} + \frac{\sqrt{W_{xx}^{(2)}W_{yy}^{(2)}}}{\sqrt{W_{xx}W_{yy}}}\rho_w^{(2)}$$

und für die ökologische Korrelation zwischen den Städten

$$(11)\quad \rho_E = \frac{\sqrt{E_{xx}^{(1)}E_{yy}^{(1)}}}{\sqrt{E_{xx}E_{yy}}}\rho_E^{(1)} + \frac{\sqrt{E_{xx}^{(2)}E_{yy}^{(2)}}}{\sqrt{E_{xx}E_{yy}}}\rho_E^{(2)},$$

wobei sich dieser Ausdruck jedoch wegen

$$\rho_E^{(1)} = \frac{E_{xy}^{(1)}}{\sqrt{E_{xx}^{(1)}E_{yy}^{(1)}}} = \frac{n_1(\bar{x}_1 - \bar{x})(\bar{y}_1 - \bar{y})}{\sqrt{n_1(\bar{x}_1 - \bar{x})^2 n_1(\bar{y}_1 - \bar{y})^2}} = 1$$

und $\rho_E^{(2)} = 1$ vereinfacht zu

$$(12)\quad \rho_E = \frac{\sqrt{E_{xx}^{(1)}E_{yy}^{(1)}} + \sqrt{E_{xx}^{(2)}E_{yy}^{(2)}}}{\sqrt{E_{xx}E_{yy}}} = \frac{\varepsilon_1 + \varepsilon_2}{\sqrt{E_{xx}E_{yy}}}.$$

Was hier die Größen ε_1 und ε_1 bedeuten, wird aus der Gleichung für $\rho_E^{(1)}$ deutlich. Danach ist $\varepsilon_1 + \varepsilon_2 =$
$= n_1(\bar{x}_1 - \bar{x})(\bar{y}_1 - \bar{y}) + n_2(\bar{x}_2 - \bar{x})(\bar{y}_2 - \bar{y})$,
aber der Nenner $\sqrt{(E_{xx}^{(1)} + E_{xx}^{(2)})(E_{yy}^{(1)} + E_{yy}^{(2)})}$ ist leider etwas komplizierter, nämlich die Wurzel aus dem Produkt
$(n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2)(n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2)$
was sich jedoch auch hier wieder vereinfacht zu $n^2(\bar{x}_1^2 + \bar{x}_2^2 - \bar{x}^2)(\bar{y}_1^2 + \bar{y}_2^2 - \bar{y}^2)$.

Es ist klar, dass es für ρ_E entscheidend darauf ankommt, dass die Mittelwerte (bezüglich X und/oder Y) der Städte untereinander und damit vom jeweiligen Gesamtmittel abweichen.

Nebenüberlegung (bis ♦): Wir prüfen nun ob $|\rho_E| \le 1$ ist, wie bei einer Korrelation erforderlich. Man sieht schnell, dass ρ_E betragsmäßig nicht größer als 1 werden kann, denn der quadrierte Zähler von ρ_E beträgt $\varepsilon_1^2 + 2\varepsilon_1\varepsilon_2 + \varepsilon_2^2$ und der quadrierte Nenner ist $(E_{xx}^{(1)} + E_{xx}^{(2)})(E_{yy}^{(1)} + E_{yy}^{(2)}) = \varepsilon_1^2 + \varepsilon_2^2 + E_{xx}^{(1)}E_{yy}^{(2)} + E_{xx}^{(2)}E_{yy}^{(1)}$, was nicht kleiner ist als der Zähler, da
$E_{xx}^{(1)}E_{yy}^{(2)} + E_{xx}^{(2)}E_{yy}^{(1)} \ge 2\varepsilon_1\varepsilon_2 = 2\sqrt{E_{xx}^{(1)}E_{yy}^{(2)}E_{xx}^{(2)}E_{yy}^{(1)}}$
so lange $\left(\sqrt{E_{xx}^{(1)}E_{yy}^{(2)}} - \sqrt{E_{xx}^{(2)}E_{yy}^{(1)}}\right)^2 \ge 0$ ist. ♦

Um abschließend wieder zur Korrelation ρ_T auf Basis aller $n_1 + n_2$ Individuen zu gelangen gilt auch hier wieder Gl. (9)
$\rho_T = (T_{xx}T_{yy})^{-1/2}\left[(\varepsilon_1 + \varepsilon_2) + \sqrt{W_{xx}W_{yy}}\rho_W\right]$ wobei wir auf den ersten Ausdruck in der eckigen Klammer, nämlich $\varepsilon_1 + \varepsilon_2$ bereits ausführlich eingegangen sind (Gl. 12) und nach obiger Gl. 10 für den zweiten Ausdruck gilt
$\sqrt{W_{xx}W_{yy}}\rho_w = \sqrt{W_{xx}^{(1)}W_{yy}^{(1)}}\rho_w^{(1)} + \sqrt{W_{xx}^{(2)}W_{yy}^{(2)}}\rho_w^{(2)}$
Die Korrelationen der Teilgesamtheiten $\rho_w^{(1)}$ und $\rho_w^{(2)}$ einerseits und ρ_T andererseits können also ganz im Sinne von Simpsons Paradoxon sehr unterschiedlich sein und es gibt keinen Grund anzunehmen, dass z.B. ρ_T ein Mittel aus $\rho_w^{(1)}$ und $\rho_w^{(2)}$ sein müsste. Somit

kann man durchaus Simpsons Paradoxon als Spezialfall der ecological fallacy auffassen.

Die ganze Betrachtung erinnert stark an die bekannte Zerlegung der Varianz σ^2 bei einer klassierten Verteilung in eine externe und eine interne Varianz.[155] Dabei gilt ja

$$\sigma^2 = \sum h_i(\mu_i - \mu)^2 + \sum h_i\sigma_i^2 = V_{ext} + V_{int},$$

wobei der erste Summand (V_{ext}) die externe und der zweite (V_{int}) die interne Varianz darstellt. Wie man sieht ist die interne Varianz ein arithmetisches Mittel der Varianzen σ_i^2 innerhalb der Schichten ($\Sigma h_i = 1$), aber auch V_{ext} ein solches Mittel nämlich $h_I(\mu_1-\mu)^2 + \ldots + h_I(\mu_I-\mu)^2$.

Ähnlich gilt auch bei der ecological fallacy $T_{xx} = E_{xx} + W_{xx}$ mit $E_{xx} = \sum n_i(\bar{x}_i - \bar{x})^2$ und $W_{xx} = \sum n_i\sigma_i^2$ (T_{yy} analog). Mit Korrelationen statt Varianzen ist die Situation jedoch nicht so einfach: zwar setzt sich

- auch die within-Korrelation ρ_W aus (stadtspezifischen inneren) Korrelationen $\rho_w^{(1)}$ und $\rho_w^{(1)}$ zusammen und auch
- die ecological (between) Korrelation ρ_E kann nach Gl. 11 als Linearkombination der entsprechenden Korrelationen $\rho_E^{(1)}$ und $\rho_E^{(2)}$ geschrieben werden (Gl. 11),

aber die Gewichte addieren sich nicht zu 1 (im Unterschied zu den Gewichten $h_i = n_i/n$), wie bei den Varianzen. Das gleiche gilt auch nach Gl. 9 für ρ_T im Verhältnis zu ρ_E und auch ρ_W.

8. Zusammenfassung und Ergänzungen

8.1. Zusammenfassung

Wir haben gesehen:

- Um eine statistische Methode wirklich zu verstehen und bei Statistiken Fehlinterpretationen zu vermeiden, kommen wir meist nicht darum herum, uns die Dinge sorgfältig zurechtzulegen und genau zu überlegen; wir können insbesondere mit einer spontanen "intuitiven" Einschätzung völlig danebenliegen, wofür das Ziegenproblem in Abschn. 2e ein Beispiel ist.

  Es ist auch sinnvoll, sich früh und eingehend mit Wahrscheinlichkeitsrechnung und Inferenz (Schließen von einer Stichprobe) zu beschäftigen, auch wenn die Entwicklung der Statistik etwas anders verlaufen ist und in der Praxis nach wie vor die Deskriptive Statistik im Vordergrund steht. Denn heutzutage ist Statistik mehr und mehr in Analysen auf der Basis von stochastischen Modellen.[156]

- Es ist offenbar schwierig, für ein Problem ***die relevante bedingte Wahrscheinlichkeit P(A/B) zu bestimmen*** (was ist A und was B im konkreten Fall) und daraus richtige Schlüsse zu ziehen. Darum ist es nützlich, das Bayessche Theorem zu studieren, was zugleich als ein Modell für das "Schätzen" und Lernen aus der Erfahrung gesehen werden kann.

- ***Wahrscheinlichkeitsaussagen*** (Interpretation von berechneten Wahrscheinlichkeiten) sind von fundamental anderer Art als erhobene Daten für einzelne Einheiten (z.B. einzelne Personen), weil sie sich nicht darauf beziehen, was sich bei einer konkreten Person ereignen kann, sondern auf die (hypothetische) Gesamtheit aller unter den gleichen Bedingungen möglichen Ereignisse (viele Missverständnisse, nicht nur die gambler's fallacy beruhen darauf, dass dieser Unterschied nicht gesehen wird); [157]

[155] Man benutzt diesen Zusammenhang mit interner und externer Varianz z.B. bei der Herleitung des Schichtungseffekts (gewinn in Gestalt von mehr Genauigkeit oder geringerer Anzahl der zu befragenden Einheiten, also geringerem Stichprobenumfang) einer geschichteten Stichprobe. Vgl. v. d. Lippe, Induktive Statistik, Formeln, Aufgaben, Klausurtraining, 5. Aufl. 1999, S. 110

[156] Auch hier gilt leider wohl, dass wir nicht darum herum kommen, uns auch in schwierige Dinge zu vertiefen um Statistik zu verstehen.

[157] In diesem Sinne ist z.B. auch der "Stichprobenfehler", wie die Stichprobenverteilung, aus der er abgeleitet ist, ganz anders als die "Repräsentativität" nicht eine Aussage über *eine konkrete Stichprobe* (die ja auch wegen der Zufallsauswahl zufällig ganz anders ausfallen kann), sondern über *die Gesamtheit der aus der gleichen Grundgesamtheit zu ziehenden Stichproben* gleichen Umfangs. Auch Missverständnisse darüber, was eine "signifikante" Testentscheidung eigent-

- Aussagen über ***Ursachen*** können i.d.R. nicht sicher durch die Berechnung von Korrelationen aufgrund von Beobachtungsdaten gewonnen werden, weil hier alternative Möglichkeiten der (auch indirekten über eine dritte Variable) Verursachung (z.B. in Gestalt einer Scheinkorrelation) i.d.R. nicht ausgeschlossen werden können. Diesem Zweck, sicherzustellen, dass alternative Erklärungen ausgeschlossen werden können, dient im Experiment die "Kontrolle" sonstiger Einflüsse und die Randomisierung (Abschn. 4a), und bei der Analyse von Beobachtungsdaten die Prüfung der Modellannahmen über die Störgrößen (Abschn. 4c).
- Statistik "verstehen", wie es im Titel dieses Papiers heißt, bedeutet vor allem, mehr und mehr Zusammenhänge zwischen Methoden zu erkennen und das hinter ihnen bestehende *System* zu sehen.
- Zwar steht auch bei "***Big Data***" die Erwartung, gerade durch die Massenhaftigkeit von Zahleninformationen zu neuen Erkenntnissen zu gelangen im Fokus, ähnlich wie traditionell in der Statistik. Gleichwohl dürfte Big Data in vieler Hinsicht eher das glatte Gegenteil von Statistik sein.
- Es gibt bestimmte Denkmuster und Begrifflichkeiten die in der Statistik immer wieder auftreten, wie z.B. die Unterscheidung zwischen ***systematisch*** (erklärt) und ***zufällig*** (residual) und eine darauf aufbauende Varianzzerlegung, oder die Bestimmung von Wahrscheinlichkeiten bei wiederholten ***unabhängigen*** Zügen (Stichproben) aus identisch verteilten Grundgesamtheiten.
- Fundamental ist auch für das "Schätzen" und "Testen" das Konzept der ***Stichprobenverteilung*** (z.B. einer Teststatistik) bei Geltung einer Hypothese und das ***Maximum Likelihood*** Prinzip, für unbekannte Parameter der Grundgesamtheit die Werte anzunehmen, bei denen das konkret beobachtete Stichprobenergebnis am wahrscheinlichsten ist.
- Ein zentrales Instrument ist auch das Modell eines datenerzeugenden Prozesses auf Basis einer Gleichung oder eines Gleichungssystems, wobei ***stochastische Modelle*** deutlich Vorteile haben gegenüber nichtstochastischen Modellen, weil sie es erlauben, die Güte der Anpassung des Modells an die Daten zu beurteilen. Im Rahmen solcher Modelle gilt es oft, Gewichte $g_1, \ldots, g_m$ einer Linearkombination von Variablen wie $\tilde{y}_i = g_1 y_{1i} + g_2 y_{2i} + \ldots + g_m y_{mi}$ so zu bestimmen, dass die Variable $\tilde{y}$ [meist unter einigen Nebenbedingungen] in bestimmter Weise "optimal" ist.[158]
- Der ***Zufall als Auswahlprinzip*** bei (echten) Stichproben soll nicht nur sicherstellen, dass keine Verzerrung durch Bevorzugung (z.B. von besonders leicht erreichbaren Einheiten) oder Benachteiligung von Einheiten eintritt, sondern auch die "Repräsentativität" in dem Sinne gewährleisten, dass die ausgewählten Einheiten auch die nicht ausgewählten repräsentieren (was sie nur dann tun, wenn auch die nicht ausgewählten Einheiten die gleiche Auswahlchance hatten wie die ausgewählten). Nur die Zufallsauswahl erlaubt die Anwendung der Wahrscheinlichkeitsrechnung und die Berechnung des ***Stichprobenfehlers*** als Gütemaß. "***Repräsentativität***" im Sinne von gleichen (oder ähnlichen) Quoten in der Stichprobe wie in der Grundgesamtheit ist dagegen ein unbrauchbares Konzept.[159]
- Ein statistischer Test wird oft als Verifikation oder Falsifikation einer Aussage über die Realität missverstanden. Aber mit einem Test wird nicht *festgestellt*, ob

lich bedeutet, beruhen oft auf ein Verkennen des Charakters einer Wahrscheinlichkeitsaussage.

[158] Es kann z.B. gefordert werden, dass sie eine minimale Varianz hat, oder dass mit einer anderen Linearkombination möglichst hoch korreliert. Die meisten Verfahren der "multivariaten Analyse" beruhen auf solchen Überlegungen.

[159] Hierauf wird jedoch in gewisser Weise implizit Bezug genommen bei der sog. Hochrechnung" (dazu mehr unten im Teil Ergänzungen 8.2c).

eine Hypothese richtig oder falsch *ist*, sondern es wird *entschieden*, ob man sie für richtig oder falsch *halten soll*. Die bisher verbreitete einseitige Betonung der Testsentscheidungen ("signifikant") bei der Darstellung empirischer Befunde hat sicher erhebliche Nachteile. Aber eine jetzt in Mode gekommene Fokussierung auf die ***power*** oder Maße der ***Effektstärke*** ist auch nicht unproblematisch. Hinzu kommt, dass die genannten Kriterien untereinander zusammenhängen und auch ganz entscheidend vom Stichprobenumfang bestimmt werden. Es wird auch oft vergessen, dass statistische Tests nicht nur Hypothesen über Modell***parameter*** sondern auch solche über Modell***voraussetzungen*** (die für die Schätzung eines Modells notwendig sind) betreffen, insbesondere Annahmen über die Störgrößen.

- Abschn. 7 zeigte, dass Beziehung zwischen Kennzahlen (statistics), wie Mittelwerte, Quoten, Korrelationskoeffizienten usw. für ***Teilaggregate*** (Teilgesamtheiten) ***und*** den entsprechenden Kennzahlen für das ***Gesamtaggregat*** oft sehr viel verwickelter sind als man zunächst denken mag.[160] Auf hier naheliegende Fehlschlüsse beziehen sich diverse Paradoxien und fallacies, wobei aber kaum bekannte Zusammenhänge zwischen ihnen bestehen und oft eine Paradoxie auf eine andere zurückgeführt werden kann. So kann man z.B. das Simpson Paradoxon als einen relativ einfachen Spezialfall der doch recht komplexen ecological fallacy auffassen.
- Zwar haben wir einige verbreitete Argumentationsmuster in der induktiven Statistik dargestellt, aber wir konnten ähnlich grundlegende Überlegungen aus der deskriptiven Statistik aus Platzgründen hier nicht unterbringen. Nur in einem Fall, was die sog. "Axiome" betrifft, wollen wir im Folgenden kurz etwas ergänzend nachtragen (Abschn. 8.2a).

8.2. Ergänzungen

a) Axiome in der Deskriptiven Statistik
b) Messung und Bereinigung von Struktureffekten (Details zu Abschn. 7)
c) Hochrechnung
d) Endogenität und Simultaneität
e) Einige nützliche Transformationen
f) Witze über Statistik

a) Axiome in der Deskriptiven Statistik

Es ist allgemein bekannt, dass es verschiedene Mittelwerte M_x einer Variable x gibt. Man kann z.B. das arithmetische Mittel $\bar{x}$ oder den Median (Zentralwert) berechnen, oder das (weniger bekannte) geometrische und harmonische Mittel, oder man könnte auch eine ganz neue Mittelwertformel vorschlagen. Mit welchen Argumenten kann man sich für oder gegen eine Formel aussprechen oder sich für eine besonders vorteilhafte entscheiden? Hierzu dienen "Axiome".[161]

Von einem Mittelwert ist es z.B. sinnvoll, zu fordern, dass er zwischen dem kleinsten und dem größten Wert liegt $x_{min} \le M_x \le x_{max}$, wie das Wort "*Mittel*"wert ja schon sagt. Ist M_x größer als der größte x-Wert oder kleiner als der kleinste x-Wert, so wäre dies nicht "sinnvoll". Axiome sind exakt gefasste Aussagen darüber, was als sinnvoll oder sinnlos anzusehen ist, und weil sie exakt gefasst sind, kann man beweisen ob eine Formel ein Axiom erfüllt oder nicht. Ein ebenfalls vernünftiges Axiom ist, dass M_x berechnet mit den Datenvektor $[x_1 \;\; x_2 \;\; (x_3+\Delta)]$ größer (wenn $\Delta > 0$), bzw. kleiner (wenn $\Delta < 0$) sein sollte als M_x beim Vektor $[x_1 \;\; x_2 \;\; x_3]$.[162]

Axiome erlauben es auch Klassen von Maßzahlen zu unterscheiden, z.B. Streuungsmaße (S) von Maßen der Konzentration (K). Einzusehen,

[160] Verwickelt ist insbesondere der Zusammenhang zwischen Korrelationen, die *auf Basis von Aggregaten* (z.B. Städte eines Landes) berechnet wurden und den entsprechenden Berechnungen *auf Basis der individuellen Daten* der Personen (ecological fallacy).

[161] Axiome als Beurteilungsmaßstäbe sind traditionell sehr üblich bei Preisindexzahlen und Maßen der Konzentration (Ungleichheit). In meinem Buch "Deskriptive Statistik" (UTB Reihe Bd. 1632) 1993 habe ich versucht, für die meisten Maße Axiome zu definieren. Da das Buch nicht wieder aufgelegt wurde, steht es auf dieser Website frei zum Download zur Verfügung.

[162] Diese Forderung ist das Monotonie-Axiom: Wenn alle x-Werte gleich sind bis auf einen x-Wert, sollte der Mittelwert nicht gleich sein, sondern diesem Unterschied Rechnung tragen (hinsichtlich der *Richtung*, d.h. größer oder kleiner werden; es ist nicht gesagt, um *wie viel* größer oder kleiner).

dass mit S und K unterschiedliche Sachverhalte gemessen werden ist wohl für einige schwer, zumal man ja beides berechnen kann, wenn man eine Häufigkeitsverteilung von x hat.[163]

Ich hörte neulich einen Vortrag eines Professors M aus L, bei dem die angewandten statistischen Methoden nicht über den Stoff der ersten beiden Vorlesungen Statistik I hinausgingen (Median, Quartilsabstand etc.). Noch befremdlicher war jedoch, dass M eine Lorenzkurve und das Disparitätsmaß von Gini d_G bei der Variable "wöchentliche Arbeitszeit in Stunden" betrachtete. Dabei sollte eigentlich klar sein dass relative Konzentration (= Disparität) nur sinnvoll ist bei einem extensiven (summierbaren) Merkmal, wie Vermögen, nicht aber wenn x intensiv ist (z.B. bei der Intelligenz) oder wenn für x nur ein sehr begrenzter Wertebereich möglich ist.

Nicht bei jeder Art von "Ungleichheit" kann man die Lorenzkurve zeichnen oder d_G berechnen. Man kann es beim Vermögen, aber beim IQ oder bei der Arbeitszeit wäre es Unsinn.[164]

Wenn einer ein Vermögen von 50.000 und der andere von 150.000 hat, kann man sinnvoll von "zusammen 200.000" sprechen, oder sich auch vorzustellen, dass einer allein 200.000 und der andere 0 hat. Vermögen ist ein extensives Merkmal. Aber es ist Unsinn, zu meinen, dass wenn einer einen IQ von 110 und der andere einen von 130 hat, dass sie dann zusammen 240 haben, oder dass es auch genauso gut sein könnte, dass einer 0 und der andere 240 hat.

Es ist abwegig, bei der Arbeitszeit die Lorenzkurve zu betrachten. Selbst wenn es einen Kuchen der Gesamtarbeitszeit gäbe, kann man sich ja davon nicht beliebig große Scheiben abschneiden. Die Scheibe kann schon mal nicht größer als 168 (= 24 mal 7) Stunden wöchentlich sein und was würde auch $d_G = 1$ bedeuten? Es hieße doch, dass einer allein etliche Millionen Stunden in der Woche arbeitet (wie soll das gehen?) und alle anderen 0 Stunden.[165]

b) Messung und Bereinigung von Struktureffekten (Ergänzung zu Abschn. 7)

Abgesehen vom Assoziations- und Korrelationskoeffizient haben wir in Abschn. 7 nur solche Maße betrachtet, bei denen das Maß für das Gesamtaggregat eine Linearkombination der entsprechenden Maße für die Teilgesamtheiten ist, wie bei einer Beziehungszahl (z.B. der Todesrate)

$$Q = \frac{X}{Y} = \sum Q_j g_{yj} \text{ mit } Q_j = \frac{x_j}{y_j},\ X = \Sigma x_j,\ Y = \Sigma y_j \text{ und } g_{yj} = y_j / Y \text{ (Gewichte)}$$

mit j = 1, 2, …, J oder dem Mittelwert

$$\bar{x} = \sum \bar{x}_j h_j \text{ mit den relativen Häufigkeiten } h_j \text{ und } \Sigma h_j = 1$$

Wenn es nur zwei Gesamtheiten (Aggregate) gibt, A und B, die jeweils aus J Teilaggregaten bestehen, kann man Differenzen zur Kennzeichnung ihres Unterschieds bilden.[166] Die Differenz $\Delta Q = Q_A - Q_B$ hat zwei Komponenten, eine "echte" Veränderung E und eine Strukturveränderung S.

Mit $Q_A = \sum_j Q_j^A g_{yj_2}^A$ und Q_B entsprechend ist

$$\Delta Q = \sum_j \left(Q_j^A - Q_j^B\right) g_{yj}^A + \sum_j \left(g_{yj}^A - g_{yj}^B\right) Q_j^B = E + S$$

oder

$$\Delta Q = \sum_j \left(Q_j^A - Q_j^B\right) g_{yj}^B + \sum_j \left(g_{yj}^A - g_{yj}^B\right) Q_j^A = E^* + S^*.$$

Es kann also bei $Q_A \neq Q_B$ für alle j = 1, …, J gelten:

$Q_j^A - Q_j^B = 0$	**S**: kein echter, nur Strukturunterschied	Priester/Bergarbeiterbeispiel
$g_{yj}^A - g_{yj}^B = 0$	**E**: Echter Unterschied	hierfür oben kein Beispiel gebracht

Man beachte:
sowohl in S bzw. S als auch E* bzw. S* werden Differenzen gewichtet, aber die Q-Gewichte bei S bzw. S* summieren sich (anders als die g-Gewichte bei E bzw. E*) nicht zu 1.

[163] Dass der Unterschied zwischen absoluter und relativer Konzentration (=Disparität) vielen nicht klar ist (obgleich auch er darin besteht, dass andere Axiome gefordert werden und damit die Aussage entsprechender Maße [z.B. Herfindahl vs. Gini Index] eine andere ist) ist bekannt, aber ungewöhnlich ist es schon, dass man den Unterschied zwischen Streuung und Disparität nicht kennt. Es gibt eine Überschneidung insofern, als der Variationskoeffizient für die Messung sowohl der relativen Streuung, als auch der Disparität benutzt wird (wo er jedoch ein meist für die Disparitätsmessung gefordertes Transferaxiom nicht erfüllt). Mehr dazu in v. d. Lippe, Deskriptive Statistik, S. 171.

[164] Bei Intelligenz und Arbeitszeit ist ein Streuungsmaß das allein richtige Maß, um zu zeigen, dass hier große Unterschiede zwischen den Menschen bestehen.

[165] Die Lorenzkurven von Prof. Dr. M. waren – wie zu erwarten war – entsprechend auch alle nicht weit von der Gleichverteilungsgeraden entfernt,

[166] Zu dieser Betrachtung v. d. Lippe, Deskriptive Statistik, S. 335. Bei I > 2 Teilgesamtheiten sind die Zusammenhänge natürlich etwas komplizierter.

Beim arithmetischen Mittel gilt analog für $\Delta\bar{x} = \bar{x}_A - \bar{x}_B$ die Zerlegung in E und S

$$\Delta\bar{x} = \sum\left(\bar{x}_j^A - \bar{x}_j^B\right)\cdot h_j^A + \sum\left(h_j^A - h_j^B\right)\cdot \bar{x}_j^B,$$

bzw. in E* und S*

$$\Delta\bar{x} = \sum\left(\bar{x}_j^A - \bar{x}_j^B\right)\cdot h_j^B + \sum\left(h_j^A - h_j^B\right)\cdot \bar{x}_j^A.$$

Das Muster der Zerlegung in E/S bzw. E*/S* und der Gewichtung dabei ist also bei $\Delta\bar{x}$ ganz analog dem von ΔQ.

Die Gleichungen für $\Delta\bar{x}$ erklären auch das Simpson Paradoxon, insbesondere können auch alle Differenzen in den Teilaggregaten negativ sein und $\Delta\bar{x}$ positiv sein (oder umgekehrt). Dass sich der reine Struktureffekt ausdrückt in $\sum \bar{x}_j^B h_j^A - \bar{x}_B \neq 0$ bzw. in $\sum \bar{x}_j^A h_j^B - \bar{x}_A \neq 0$ rechtfertigt die sog. "*Standardisierung*",[167] d.h. die Berechnung von Mittelwerten (oder auch Beziehungszahlen) bei *Zugrundelegung der gleichen Gewichte*. Die standardisierten Maße sind also bereinigt von einem evtl. bestehenden Struktureffekt.

c) Hochrechnung

Im Deutschen wird dieses Wort gebraucht für eine Prognose (z.B. am Wahlabend), was aber eigentlich eine Punktschätzung ([point] estimation) der Stimmenzahl für die Parteien ist. In der Literatur spricht man auch von Hochrechnung, wenn es gilt, eine absolute Größe N_i (z.B. eine Anzahl) in der Grundgesamtheit aufgrund eines Anteils $p = n_i/n$ in der Stichprobe zu schätzen. Das geschieht meist durch Multiplikation von p mit dem reziproken Auswahlsatz, also $pN/n = N_i$.[168] Man nennt das "***freie Hochrechnung***". Auch das ist nur eine spezielle Punktschätzung und wird in anderen Ländern oft gar nicht gesondert thematisiert.

Wir haben deshalb als Deutsche notorisch Schwierigkeiten, das Wort "Hochrechnung" zu übersetzen. Es ist klar, dass es nicht "high calculation" heißen kann, aber wie heißt es dann?

Man spricht auch von Hochrechnung bei "Korrekturen" an den erhobenen Daten (also an den Stichprobenwerten) mit dem Ziel, dass so bestimmte Quoten in der Stichprobe den entsprechenden Quoten (oder Eckzahlen) der Grundgesamtheit "angepasst" werden. Weil das meist in Form von höher- oder niedriger-"gewichten" von Angaben erfolgt, wird auch oft von "Gewichtung" (weighting) anstelle von "Hochrechnung" gesprochen. Aber weil man in der Statistik Gewichtungen auf Schritt und Tritt und in den verschiedensten Zusammenhängen begegnet, ist auch dies kein sonderlich klarer Begriff.

Vom Grundgedanken her ähneln solche Bestrebungen dem von uns kritisierten Konzept der "Repräsentativität", wofür es keine Formel und kein Maß gibt, und das auch nicht dem Prinzip der Zufallsauswahl gerecht wird. Aber in Gestalt von "Hochrechnungen" haben sich solche Gedanken wohl durchgesetzt, so dass solche Anpassungen, Korrekturen oder "Gewichtungen" wohl stets gängige Praxis sein werden.

Dabei sind zwei Punkte zu bedenken, die zur Zufallsauswahl oft kritisch vermerkt werden

- sie kann dazu führen, dass *zufällig* wichtige Einheiten der Grundgesamtheit nicht in der Stichprobe erscheinen, und
- wir haben es in der Praxis oft mit erheblichen nonresponse Quoten zu tun, so dass der Gedanke, Zufallsauswahl stelle per se Repräsentativität sicher, vielleicht doch zu theoretisch und puristisch ist.

Dem ersten Punkt kann und sollte durch eine *geschichtete* Stichprobe Rechnung getragen werden. Der zweite mag in der Tat ex post Anpassungen an vorgegebene Strukturen nach Art von Hochrechnungen rechtfertigen.

Von "***gebundenen***" im Unterschied zu freien ***Hochrechnungen*** ist auch die Rede, wenn man versucht, bekannte Zusammenhänge zwischen Variablen in der Grundgesamtheit (z.B. Differenzen $\mu_X - \mu_Y$ oder Verhältnisse [ratios] μ_X/μ_Y auf die Stichprobe zu übertragen). Das mag zur Verbesserung der Stichprobe beitragen. In der englischen Literatur findet man hierzu Begriffe, wie "ratio calibration" und "difference estimation".

d) Endogenität und Simultaneität (zu 4c)

Mit dem Hintergrund einer einfachen Überlegung aus der Ökonometrie dürfte der Endogenitätsfehler verständlicher werden: im folgenden Modell mit zwei Gleichungen

(a) $C_t = \alpha + \beta Y_t + u_t$

[167] Von "Standardisierung" spricht man auch bei der z-Transformation und der Standardnormalverteilung.
[168] Dahinter steckt natürlich die Gleichsetzung von p (Stichprobe) und $\pi = N_i/N$ (Grundgesamtheit).

(b) $Y_t = C_t + I_t$

besteht eine Interdependenz zwischen C und Y, die also beide *endogene* Variablen sind (nur I ist hier exogen). Nach der Verhaltensgleichung (a) ist $Y \rightarrow C$, andererseits ist aber nach der Definitionsgleichung (b) auch $C \rightarrow Y$. Die Schätzung der marginalen Konsumquote mit[169]

$\hat{\beta}_{OLS} = s_{CY}/s_y^2$ ist nicht konsistent, weil die Kovarianz σ_{YU} nicht null ist. Man erkennt das, wenn man die reduzierte Form bildet

$C_t = \pi_0 + \pi_1 I_t + v_t$ und $Y_t = \pi_0 + \pi_2 I_t + v_t$ mit

$$\pi_0 = \frac{\alpha}{1-\beta}, \pi_1 = \frac{\beta}{1-\beta}, \pi_2 = \frac{1}{1-\beta} \text{ und } v_t = \pi_2 u_t .$$

Wegen

$$Y_t - E(Y_t) = \frac{1}{1-\beta}(I_t - E(I_t)) + \frac{1}{1-\beta}(u_t - E(u_t))$$

und der Exogenität von I_t (so dass $\sigma_I^2 = 0$) ist σ_{YU}

$$\sigma_{YU} = \text{cov}(Y_t u_t) = \frac{1}{1-\beta} E(u_t - E(u_t))^2 = \frac{\sigma^2}{1-\beta} > 0,$$

so dass Y und u miteinander korreliert sind. "Richtig" (konsistent) wird β geschätzt mit der indirekten Methode der kleinsten Quadrate (ILS) oder – was hier auf das gleiche hinausläuft – der Methode der Instrumentvariablen (IV):

$\hat{\beta}_{ILS} = \hat{\beta}_{IV} = s_{CI}/s_{YI}$ statt mit $\hat{\beta}_{OLS} = s_{CY}/s_y^2$, was β systematisch (auch bei $n \rightarrow \infty$) überschätzt.

e) Einige nützliche Transformationen

Wir präsentieren im Folgenden vier Transformationen, die ersten drei sind lineare und die vierte ist eine nichtlineare:

1. Mit der folgenden Lineartransformation kann man eine Variable $x > 0$ in eine Variable y transformieren, die auf den Wertebereich von 0 bis 1 "normiert" ist (also $0 \le y \le 1$)

   $$y = \left(-\frac{x_u}{x_o - x_u}\right) + \left(\frac{1}{x_o - x_u}\right) \cdot x = \alpha + \beta \cdot x$$

   wenn x_u der kleineste und x_o der größte x-Wert ist.

   Der erste Koeffizient α bewirkt eine Verschiebung des Nullpunkts (so dass man für $x = x_u$ den Wert $y = 0$ erhält und die zweite Größe (β) ist die Steigung dieser Lineartransformation und dies betrifft die Maßeinheit auf der y-Skala. Die Differenz $x_2 - x_1$ wird zur folgenden y-Differenz[170]

   $$y_2 - y_1 = \left(\frac{1}{x_o - x_u}\right) \cdot (x_2 - x_1) = \beta(x_2 - x_1)$$

   und wie man sieht erhält man für $x = x_o$ den Wert $y = 1$.

2. Wenn x auch negative Werte annehmen kann ($x_u < 0$), wird mit der Transformation

   $$y = \left(-\frac{x_o + x_u}{x_o - x_u}\right) + \left(\frac{2}{x_o - x_u}\right) \cdot x = \alpha^* + \beta^* \cdot x$$

   eine Variable y erzeugen, deren Werte zwischen -1 und +1 liegen. Für $x = x_u$ erhält man $y = -1$ für $x = x_o$ ist $y=+1$, und bei einem x Wert von $x = (x_o + x_u)/2$, also beim sog. midpoint ist $y = 0$.

3. Sehr bekannt ist die sog z-Transformation

   $$z = \frac{x - \mu_x}{\sigma_x} = \left(-\frac{\mu_x}{\sigma_x}\right) + \left(\frac{1}{\sigma_x}\right) \cdot x = \alpha_z + \beta_z \cdot x$$

   mit der man eine Variablen mit Mittelwert $\mu_x = \mu$ und Standardabweichung $\sigma_x = \sigma$ in eine Variable mit einem Mittel von 0 und einer Standardabweichung von 1 transformiert, denn $\mu_z = 0$ und $\sigma_z = 1$.

   Auch dies ist eine Lineartransformation. Die Größen α_z und β_z betreffen wieder den Nullpunkt und den Maßstab, der gestaucht ($\beta < 1$) bzw. getreckt ($\beta > 1$) wird. Durch Bildung der Differenz (also von deviation scores) $x - \mu$ wird der "neue" (also der auf der z Skala) Mittelwert 0. Die anschließende Division durch σ_x erzeugt eine Variable, die in Einheiten der Standardabweichung gemessen ist. Diese Art von z-Transformation ist beliebt um verschiedene Gesamtheiten (oder Messwerte) zu vergleichen, bei denen sich die interessierende Variable auf einem unterschiedlichen Niveau (also mit unterschiedlich hohem μ und damit auch σ) bewegt. Sie setzt nicht voraus, wie oft fälschlich gedacht wird, dass die Variable X normalverteilt ist.

   Die Umkehrung $z \rightarrow x$ ist ebenfalls eine lineare Transformation $x = \mu + z\sigma$.

4. Eine Transformation, die ebenfalls unter dem Namen z-Transformation bekannt ist, wird bei

[169] OLS steht für "ordinary least squares". Für Varianzen und Kovarianzen der *Stichprobe* steht auch hier – wie allgemein üblich – s (lateinische Buchstaben) und σ (griechisch) für die entsprechenden Größen der *Grundgesamtheit*.

[170] Die Subskripte u und o stehen für unten und oben.

Tests von Korrelationskoeffizienten r benutzt, wenn die Hypothese über den entsprechenden Parameter ρ in der Grundgesamtheit lautet $\rho \neq 0$ (statt $\rho = 0$). In diesem Fall ist nämlich die Stichprobenverteilung von r nicht mehr symmetrisch, sondern linkssteil (wenn $\rho < 0$), bzw. rechtssteil (wenn $\rho > 0$) zumal ja stets $-1 \le r \le +1$ gelten muss (r kann also nicht beliebig klein oder beliebig groß werden),

Die *nichtlineare* Transformation

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \rightarrow r = \frac{e^{2z}-1}{e^{2z}+1}$$

nach R. A. Fisher erlaubt es jetzt, gleichwohl approximativ mit einer Normalverteilung (für z, nicht für r) zu rechnen.

f) Witze über Statistik

Falsche Bezugsgröße (und falsche verbale Verkleidung von Rechenergebnissen)

In der Mathematik ist es kein Unterschied, ob man mit $\varepsilon = P/T$ oder mit $1/\varepsilon = \varepsilon^{-1} = T/P$ operiert. In der verbalen Erläuterung von Statistiken kann das aber sehr wohl der Fall sein. Angenommen, man habe per Erhebung in einem Zeitraum von T = 24 Stunden P = 420 tödliche Verkehrsunfälle festgestellt. T war vorgegeben und P hat sich so ergeben. Die Relation $\varepsilon = P/T = 420/24 = 17{,}5$ Verkehrstote pro Tag ist eine sinnvolle Aussage und entspricht dem, was erhoben wurde. Viele glauben aber, es dadurch verständlicher zu machen, dass sie mit T/P = 0,057 Stunden pro Toten rechnen und dann sagen, dass alle 3,428 (= 0,057·60) Minuten ein Mensch überfahren wird (was dann zur Frage verleitet: "wie hält der/die das überhaupt aus?"). Jetzt ist P (und nicht wie bei der Datenerhebung der Tag mit seinen T = 24 Stunden) die fest vorgegebene Bezugsgröße und das Zeitintervall zwischen den Unfällen die variable Größe.[171] Mathematisch ist T/P und P/T gleichwertig, aber sprachlich liegt hier durchaus ein Unterschied vor.

Auch operativ (hinsichtlich zu treffender Maßnahmen) ist P/T die interessante Größe, und nicht T/P. Denn es geht nicht darum. das ohnehin nicht (wie T/P fälschlich impliziert) konstante Zeitintervall zwischen tödlichen Unfällen zu verlängern, sondern die Anzahl der Unfälle (egal zu welcher Tageszeit sie passieren) zu verringern.

[171] Alle 3,4 Minuten suggeriert auch eine uhrwerksartige Regelmäßigkeit, die in der Realität gar nicht gegeben ist. Unfälle passieren ja nicht nach jeweils gleich langen Intervallen.

Falsches Merkmal

Ein anderer altbekannter Scherz ist: man hat beim Schuss auf eine Wildgans einmal zu hoch und einmal zu niedrig geschossen und ein Statistiker wird dann sagen: im Mittel habe man aber richtig geschossen.

Der Witz liegt hier darin, dass die interessierende Variable falsch definiert ist. Richtig wäre: man hat zweimal *nicht getroffen* (also *daneben*geschossen, egal ob zu hoch oder zu niedrig) und daraus wird nicht im Mittel ein Treffer. Die interessierende Variable ist also nicht, wie hoch geschossen wird, sondern ob man getroffen oder nicht getroffen hat.

Falsch gesetzte Bedingung

Der folgende Text, den ich im Internet gefunden habe, ist vielleicht weniger ein Witz als ein Beispiel für eine falsche Schlussweise mit bedingten Häufigkeiten (Wahrscheinlichkeiten):

> "15% aller Verkehrsunfälle geschehen unter Alkoholeinfluss, d.h. wenn wir alle besoffenen Schweine von der Straße holen, könnten jedes Jahr 15% aller Unfälle vermeidbar sein.
>
> "Wirklich? Und wenn wir alle nichtbesoffenen Schweine von der Straße holen wären sogar 85% aller Unfälle vermeidbar!"

Was hier vorliegt ist eine Vertauschung der Konditionen (Bedingungen) wie sie auch in Abschn. 2d betrachtet wurde.

Es sei T = betrunken sein und V = Verkehrsunfall. Weil sich die 15% und 85% zusammen auf die Gesamtzahl der Verkehrsunfälle beziehen besagen die Zahlen

$P(T|V) = 0{,}15$ und $P(\overline{T}|V) = 0{,}85$, nicht aber wie groß $P(V|T)$ und $P(V|\overline{T})$ sind, was aber eigentlich interessiert. Die Rechnung mit den angeblich zu 100% vermeidbaren Verkehrsunfällen geht so: $P(V) - P(T|V) - P(\overline{T}|V) = 0$

Das wäre aber nur richtig, wenn $P(V) = P(T|V) + P(\overline{T}|V)$ wäre,[172] tatsächlich ist aber $P(V) = P(VT) + P(V\overline{T})$, oder $P(V) = P(V|T)P(T) + P(V|\overline{T})P(\overline{T})$, was aber keineswegs 100% ist. Mit dem Bayesschen Theorem erhält man

[172] Hinzu kommt, dass Unfall (V) und Verkehrsteilnehmer (die man "von der Straße holen" möchte) gleichgesetzt wird und damit P(V) praktisch 1 gesetzt wird. Aber nicht jeder Verkehrsteilnehmer (ob betrunken oder nicht) verursacht einen Unfall.

$P(V|T) = P(T|V)P(V)/P(T)$ und $P(V|\overline{T}) = P(\overline{T}|V)P(V)/P(\overline{T})$. Auch ohne P(V) zu kennen könnte man die Relation

$$\frac{P(V|T)}{P(V|\overline{T})} = \frac{P(T|V)}{P(\overline{T}|V)}\frac{P(\overline{T})}{P(T)} = \frac{0{,}15 \cdot P(\overline{T})}{0{,}85 \cdot P(T)}$$

auch ohne P(V) zu kennen bestimmen. Weil es sehr viel wahrscheinlicher ist, nicht betrunken als betrunken Auto zu fahren – (nehmen wir einmal eine Relation von 9:1 an bei $P(\overline{T})/P(T)$ – ist die entscheidende Relation für die Gefährlichkeit der Trunkenheit am Steuer nicht 0,15/0,85, sondern neunmal so groß.

Zwei weitere Beispiele für die Verwechslung der Likelihood mit der a posteriori Wahrscheinlichkeit[173]

1. Werbeaktion mit missverstandener Statistik

Bei Dewdney (S. 112) findet man das folgende ähnliche Beispiel: "Eine Reihe von Elektrizitätsgesellschaften der USA startete vor einiger Zeit eine Kampagne, in der herausgestellt wurde, wie vorteilhaft sich eine gute Straßenbeleuchtung auf die Verhütung der Kriminalität auswirkt. In den groß aufgemachten Annoncen wurde festgestellt, daß in den USA 96 Prozent der innerörtlichen Straßen schlecht beleuchtet sind und dort 88 Prozent der Verbrechen verübt wurde. Für viele Leser ging daraus eine eindeutige Korrelation hervor: Dank der Elektrizität passiert nicht noch mehr."

Dem Text sind folgende Angaben zu entnehmen wenn L = gute und $\overline{L}$ keine (schlechte) Beleuchtung und V = Verbrechen bedeutet:

	$\overline{V}$	V	Summe
L		$0{,}12 \cdot P(V)$	0,04
$\overline{L}$		$0{,}88 \cdot P(V)$	0,96
Summe	$P(\overline{V})$	P(V)	1

Wir kennen also $P(L|V) = 0{,}12$ und $P(\overline{L}|V) = 0{,}88$, aber nicht – was eigentlich interessiert – $P(V|L)$ und $P(V|\overline{L})$. Man kann aber leicht mit dem Bayesschen Theorem sehen

$$P(V|L) = \frac{0{,}12 \cdot P(V)}{0{,}04} \text{ und } P(V|\overline{L}) = \frac{0{,}88 \cdot P(V)}{0{,}96},$$

so dass mit $P(V|L)/P(V|\overline{L}) = 3{,}27$, eigentlich Verbrechen auf beleuchteten Straßen mehr als dreimal so wahrscheinlich sind, wie auf unbeleuchteten Straßen.

Offenbar ist den Machern der Werbekampagne (aber auch den Lesern) gar nicht aufgefallen, dass die Statistik nicht für, sondern gegen ihr Anliegen spricht. Natürlich ist wohl auch hier wieder eine Korrelation (zwischen L und V) durch eine dritte Variable (Nähe zum Stadtzentrum) zu erklären.

2. Noch einmal Verkehrsunfälle

Dieses Beispiel wird in vielen einschlägigen Büchern erwähnt: Es passieren mehr Autounfälle A in der näheren Umgebung des Wohnorts (N) als weiter entfernt: $P(N|A) > P(\overline{N}|A)$.

Was aber beabsichtigt ist zu zeigen, ist die angeblich größere Gefährlichkeit der Wohnortsnähe (sie drückt sich durch die Unfallneigung A, nicht durch die Entfernung N aus); gefragt ist also ob P(A|N) besonders groß ist. Analog zur obigen Gleichung für $P(V|T)/P(V|\overline{T})$ gilt jetzt

$$\frac{P(A|N)}{P(A|\overline{N})} = \frac{P(N|A)}{P(\overline{N}|A)}\frac{P(\overline{N})}{P(N)} \text{ und da } P(\overline{N}) \text{ i.d.R.}$$

sehr viel keiner ist als P(N) (die meisten Fahrten finden ja in Wohnortnähe statt) kann trotz $P(N|A) > P(\overline{N}|A)$ gelten $P(A|N) > P(A|\overline{N})$.

Literatur

Es gibt – wie eingangs erwähnt – unzählige Bücher, die auf der Welle von Darrel Huff "How to Lie with Statistics" reiten, und die alle – sehr zu meinem Erstaunen – enorm erfolgreich sind. Von den deutschen Büchern dieser Art habe ich hier nur Bos-bach/Korff und Krämer aufgeführt.[174]

Ähnlich beliebt sind auch die deutlich niveauvolleren Bücher (noch überwiegend englisch und noch nicht so sehr auch in deutscher Sprache) darüber, dass unsere Verständnisprobleme mit Statistik mit dem geringen Überlebenswert (unter Urzeit-Bedingungen) des rich-

[173] Man kann hier auch von der *base rates fallacy* sprechen (vgl. oben S. 7), die darin besteht, die zwischen den Likelihood und den a posteriori Wahrscheinlichkeit stehenden a priori Wahrscheinlichkeiten (= base rates) zu ignorieren

[174] Das Spiel, der Statistik "Lügen" anzudichten wird sich wohl immer größter Beliebtheit erfreuen. In meinem Papier "Statistik und Manipulation" (auch auf dieser Homepage) habe ich versucht zu zeigen, warum ich davon nichts halte. Wie wenig all das Gerede über "Lügen" durchdacht ist, wird schon dann deutlich, wenn man einmal jemand, der von "Lügen" spricht, fragt, was denn die Wahrheit ist. In vielen Fällen (z.B. die "wahre" Arbeitslosenquote) wird man keine Antwort bekommen.

tigen Operierens mit Wahrscheinlichkeiten, also evolutionsgeschichtlich zu erklären sind (vgl. Fußnote 2).

Best J., Damned Lies and Statistics, Untangling Numbers from the Media, Politicians, and Activists, University of California Press 2001, 2012

Bakeman R. & B.F. Robinson, Understanding Statistics, Mahwah (NJ), London 2005

Bochenski, I. M. Die zeitgenössischen Denkmethoden, 5. Aufl., München 1971 (1. Aufl. 1954)

Bosbach G. u. J. J. Korff, Lügen mit Zahlen: Wie wir mit Statistiken manipuliert werden, München 2011

Bram U., Thinking Statistically, Kindle Edition 2011

Campbell S., Flaws and Fallacies in Statistical Thinking, Englewood Cliffs (Prentice Hall) 1974, 2002

Crosby A. W., The Measure of Reality: Quantification and the Western Society, 1250 – 1800, Cambridge University Press, Cambridge (UK), 1997

Cumming G., Understanding The New Statistics: Effect Sizes, Confidence Intervals and Meta-Analysis, New York 2013

DeHaene S., The Number Sense, How the Mind Creates Mathematics, Oxford Univ. Press 2011

Devlin K., The Math Gene. How Mathematical Thinking Evolved And Why Numbers Are Like Gossip, London 2000

Dewdney A. K., 200% of Nothing: An Eye Opening Tour through the Twists and Turns of Math Abuse an innumeracy, New York (J. Wiley) 1993, deutsch: 200 Prozent von nichts, Die geheimen Tricks der Statistik und andere Schwindeleien mit Zahlen, Basel etc. (Birkhäuser) 1994

Ellis P. D., The Essential Guide to Effect Sizes, Cambridge University Press, Cambridge (UK), 2010

Fung K., Numbers Rule Your World, The Hidden Influence of Probabilities and Statistics Mac Graw Hill 2010

Fung K., Numbersense: How to Use Bid Data to Your Advantage, Mc Graw Hill 2013

Goldacre, B., Bad Science, London 2006

Hooke, R., How to Tell Liars from the Statisticians, New York 1983

Huck, S, W. Statistical Misconceptions, New York u. London 2012

Huff, D. How to Lie with Statistics, New York (Norton) 1954

Jaffe A. J. and H. F. Spirer , Misused Statistics, New York (Marcel Decker) 1983; Neuauflage mit den Autoren Herbert A. Spirer, Louise Spirer und Abram J. Jaffe 1998

Kida T., Don't Believe Everything, New York (Prometheus) 2006

Krämer W., So lügt man mit Statistik, Neuausgabe München u. Zürich (Piper) 2011

Kimble G., How to Use (and Misuse) Statistics, Upper Saddle River NJ (Prentice Hall) 1978

Mayer-Schönberger V. & K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, London 2013

Paulos J. A., Innumeracy, New York 1988

Piamtados S., Byar D.P. & S. B. Green, The Ecological Fallacy, American Journal of Epidemiology Vol. 127/N. 5 (1988), p. 893

Pinker S., How the Mind Works, London 1999 (Penguin books)

Porter T. M., Trust in Numbers: The Pursuit of Objectivity in Science and Public Life, Princeton University Press, Princeton NJ, 1995

Savage, L. The Foundations of Statistics, 2nd ed. New York 1972

Silver N., The Signal and the Noise: Why so Many Predictions Fail but Some Don't, Penguin 20121

Steen L. A., Mathematics and Democracy: The Case for Quantitative Literacy, Princeton NJ 2001

Stevens J., Intermediate Statistics, A Modern Approach, 3rd ed., New York 2013

Sutherland S., Irrationality, Constable & Co, 1992, posthum neu aufgelegt 2007

Taleb, N. N. Fooled by Randomness, The Hidden Role of Chance in Life and in the Markets, 2nd ed., London 2007 (Penguin books)

Urdan T., Statitics in Plain English, 3rd. ed. New York 2012

Wang C., Sense and Nonsense of Statistical Inference, New York (Marcel Decker) 1993

Warner R., Applied Statistics. From Bivariate Through Multivariate Techniques, Los Angeles (Sage Publications) 2012

Oben zitierte unveröffentlichte Texte von mir (alle auf dieser Homepage: www.von-der-Lippe.org)

- Statistik für Schaumschläger (2011)
- Was tun, wenn einem eine Statistik nicht passt? Nützliche Tipps, wenn man gegen statistische Daten und Analysen eines Redners argumentieren möchte (2013) und
- Statistik und Manipulation
  Vortrag bei einer Tagung der Geschäftsstelle des Wissenschaftsrats (2013)

Benutzt, bzw. erwähnt wurde auch zwei meiner Bücher, die ebenfalls auf dieser Homepage zum Download zur Verfügung stehen:

- "Deskriptive Statistik" (UTB Reihe Bd. 1632) 1993 und
- "Induktive Statistik", Formel, Aufgaben, Klausurtraining, 5. Aufl. München u. Wien (Oldenbourg Verlag, 1999