

Das simpsonsche Paradoxon

So kommt Licht ins Dunkel

Timm Grams, Fulda, 13.07.2010 (28.02.2011), <http://www.hs-fulda.de/~grams>

Inhaltsverzeichnis

| | |
|---|---|
| Das simpsonsche Paradoxon am Beispiel..... | 1 |
| Das Paradoxon der bedingten Wahrscheinlichkeit..... | 2 |
| Repräsentation mittels Balkendiagramm | 3 |
| Repräsentation mittels Vektordiagramm..... | 3 |
| Weitere Beispiele | 4 |
| Xenophobie | 4 |
| Frauendiskriminierung..... | 5 |
| Quellen | 7 |

Das simpsonsche Paradoxon am Beispiel

Zwei Medikamente, wir wollen sie mit X und Y bezeichnen, sollen in einem Experiment hinsichtlich ihrer Wirksamkeit verglichen werden. Die Gesamtpopulation der Versuchspersonen wird aufgeteilt in einen Teil, der das Medikament X verabreicht bekommt. Der andere Teil der Population bekommt das Medikament Y. Weiterhin werden die Ergebnisse für Männer und Frauen separat erfasst. Die Tabelle 1 fasst die Zählzahlen dieses fiktiven Experiments zusammen (Havil, 2009, Tabellen 2.5 und 2.6).

Tabelle 1 Vergleich zweier Medikamente hinsichtlich Wirksamkeit

| | Männer | | Frauen | |
|---|---------|--------|---------|--------|
| | Wirksam | Gesamt | Wirksam | Gesamt |
| X | 40 | 200 | 85 | 100 |
| Y | 30 | 200 | 300 | 400 |

Wir haben es also mit einer Gesamtpopulation zu tun, die auf dreierlei Weisen in zwei Teilmengen zerlegt wird. Seien A , B und C Teilmengen der Gesamtpopulation und $\neg A$, $\neg B$ und $\neg C$ die jeweiligen Komplementärmenge. Diese Teilmengen sind Ereignisse im Sinne der Wahrscheinlichkeitsrechnung. Es ist bequem, sich diese Ereignisse als Aussagevariablen zu denken. (Genau genommen sind es Prädikate für die Subjekte der Population.) Ein Ereignis ist dann die Mengen all jener Subjekte, für die die entsprechende Aussage wahr ist.

- A ist die *Effektvariable*. Sie bezeichnet die Tatsache, dass der untersuchte Effekt eintritt. In unserem Beispiel steht A also für die Gültigkeit der Aussage „Das Medikament ist wirksam“ bzw. für die Teilmenge all jener Testpersonen, für die diese Aussage richtig ist. Mit $\neg A$ wird die Negation bezeichnet und auch die Teilmenge der Testpersonen, bei denen sich keine Medikamentenwirksamkeit hat feststellen lassen.
- B ist die *Zielvariable*. Sie bewirkt eine Unterteilung der Gesamtpopulation im Hinblick auf den Untersuchungsgegenstand. In unserem Beispiel wird diese Unterteilung vorgenommen nach dem verabreichten Medikament. Die Variable B zeichnet die Personen aus, die das Medikament X verabreicht bekommen haben. Die Negation $\neg B$ steht für die Personen mit dem Medikament Y.
- C bezeichnet eine *Einflussvariable*. Im Einführungsbeispiel bewirkt sie die Unterscheidung nach Mann und Frau: Für Männer gilt C und für Frauen $\neg C$.

Wenn wir anstelle der Zahlen in Tabelle 1 Kleinbuchstaben als Variablenbezeichner einführen, kommen wir zum Schema der Tabelle 2. Man beachte, dass in der zweiten und der vier-

ten Spalte der Tabelle Summen stehen! Beispielsweise ist b gleich der Anzahl aller Personen, für die B und C wahr ist, egal ob A oder $-A$ gilt.

Tabelle 2 Tabellenschema

| | C | | -C | |
|----|---|---|----|---|
| | A | | A | |
| B | a | b | e | f |
| -B | c | d | g | h |

Die Medikamentenwirksamkeit lässt sich nun recht übersichtlich mit bedingten Wahrscheinlichkeiten darstellen.

Das Medikament X wirkt bei Männern mit der Wahrscheinlichkeit $P(A|BC) = a/b = 20\%$.

Das Medikament Y wirkt bei Männern mit der Wahrscheinlichkeit $P(A|-BC) = c/d = 15\%$.

Das Medikament X wirkt bei Frauen mit der Wahrscheinlichkeit $P(A|B-C) = e/f = 85\%$.

Das Medikament Y wirkt bei Frauen mit der Wahrscheinlichkeit $P(A|-B-C) = g/h = 75\%$.

Also: Das Medikament X ist dem Medikament Y sowohl bei den Männern als auch bei den Frauen hinsichtlich der Wirksamkeit überlegen.

Wird nicht nach Männern und Frauen differenziert, dann ergibt sich für die Wirksamkeit des Medikaments X die Wahrscheinlichkeit $P(A|B) = (a+e)/(b+f) = 42\%$ und für die Wirksamkeit des Medikaments Y der Wert $P(A|-B) = (c+g)/(d+h) = 55\%$. Bei Zusammenfassung der Zähl-daten für Frauen und Männer kommen wir also zu einem Urteil, das dem aus der Detailbe-trachtung widerspricht: Das Medikament Y scheint dem Medikament X überlegen zu sein.

Wir wollen nun offen legen, wie derartige *Aggregierungsfehler* zustande kommen. Wir begehen dabei verschiedene Pfade der Problemrepräsentation: Wahrscheinlichkeitsrechnung, Balkendiagramm und Vektordiagramm. Jede dieser Repräsentationen hat Vor- und Nachteile, je nach Datenlage. Der Mechanismus ist in allen Fällen natürlich derselbe. Wir beginnen mit elementarer Wahrscheinlichkeitsrechnung. Wer einen anderen Einstieg bevorzugt, lässt diesen Abschnitt einfach aus.

Im Tabellenkalkulationsblatt [Simpson.xls](#) sind für die Beispiele dieser Studie und für noch ein paar weitere die verschiedenen Repräsentationsformen realisiert.

Das Paradoxon der bedingten Wahrscheinlichkeit

Elementare Wahrscheinlichkeitsrechnung liefert die folgenden Zusammenhänge zwischen den bedingten Wahrscheinlichkeiten (Székely, 1990, S. 63 ff.):

$$P(A|B) = P(A|BC) \cdot P(C|B) + P(A|B-C) \cdot P(-C|B)$$

$$P(A|-B) = P(A|-BC) \cdot P(C|-B) + P(A|-B-C) \cdot P(-C|-B)$$

Die Wahrscheinlichkeit $P(A|B)$ entsteht als gewichtete Summe der Wahrscheinlichkeiten $P(A|BC)$ und $P(A|B-C)$, das sind die vergleichsweise höheren Werte für Medikament X. Und $P(A|-B)$ ist die gewichteten Summe der vergleichsweise niedrigeren Werte $P(A|-BC)$ und $P(A|-B-C)$ für das Medikament Y. Die Umkehrung der Verhältnisse kann sich nur aus den Gewichtungsfaktoren ergeben. Diese sind

$$P(C|B) = b/(b+f) = 2/3$$

$$P(-C|B) = f/(b+f) = 1/3$$

$$P(C|-B) = d/(d+h) = 1/3$$

$$P(-C|B) = h/(d+h) = 2/3$$

Jetzt wird die Sache klar: Die Frauen sind beim Medikament X unterrepräsentiert und beim Medikament Y überrepräsentiert. Da Frauen auf beide Medikamente deutlich besser als die Männer ansprechen, macht sich das zugunsten des Medikaments Y bemerkbar:

$$P(A|B) = 20\% \cdot 2/3 + 85\% \cdot 1/3 = 42\%$$

$$P(A|-B) = 15\% \cdot 1/3 + 75\% \cdot 2/3 = 55\%$$

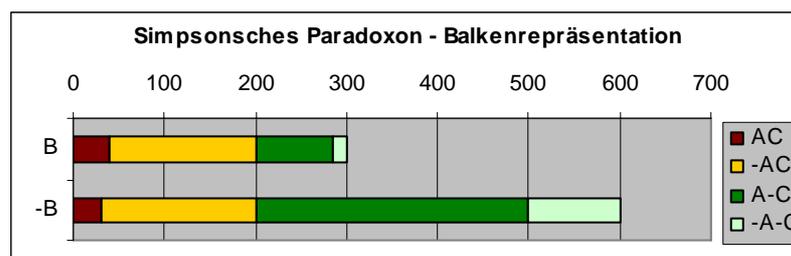
Repräsentation mittels Balkendiagramm

In diesem und im nächsten Abschnitt geht es um grafische Darstellungen der Zahlenverhältnisse aus dem ersten Abschnitt. Zunächst wird die Tabellenrepräsentation geringfügig geändert: Anstelle der Summen werden jetzt die Differenzen eingesetzt. Der Einfachheit halber machen wir das gleich mit den Zahlenwerten des einführenden Beispiels.

Tabelle 3 Tabelle für Balkenrepräsentation

| | C | | -C | |
|----|----|-----|-----|-----|
| | A | -A | A | -A |
| B | 40 | 160 | 85 | 15 |
| -B | 30 | 170 | 300 | 100 |

Diese Tabelle lässt sich gut als Diagramm gestapelter Balken repräsentieren.



Die braunen Felder repräsentieren C (Männer) und die grünen -C (Frauen). Die dunklen Teile sind die jeweiligen Anteile, für die auch A (Wirksamkeit des Medikaments) gilt. Wir sehen: Die relativen Anteile der dunklen Felder sind für B (Medikament X) jeweils größer als für -B (Medikament Y), wenn man die Sache separat für die braunen und für die grünen Felder sieht. Fasst man je Zeile die braunen und die grünen Felder zusammen, nehmen die dunklen Teilfelder im Fall -B einen größeren Anteil ein als im Fall B. Der Aggregationsfehler aufgrund der nicht ausbalancierten Gewichtung wird so augenfällig.

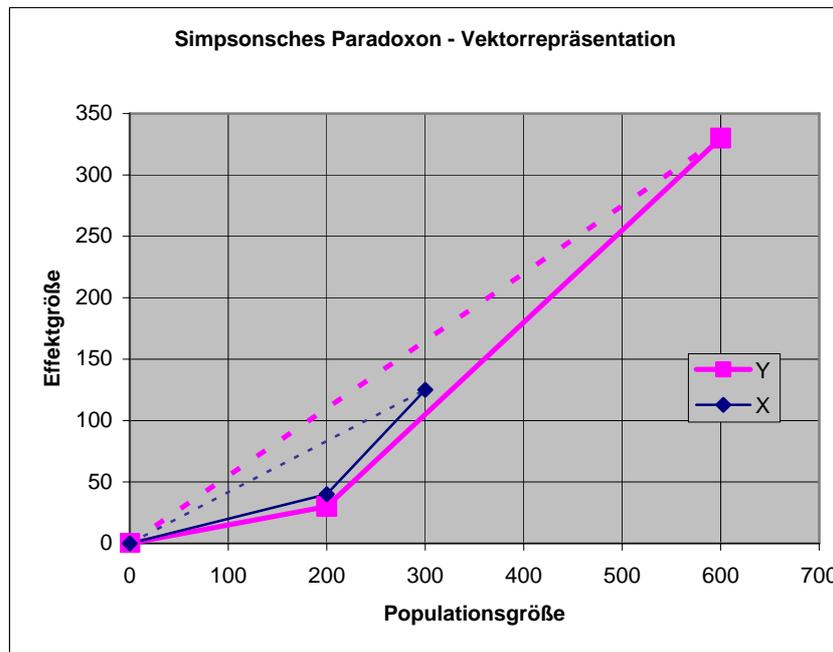
Repräsentation mittels Vektordiagramm

Die elegante Vektorrepräsentation des simpsonschen Paradoxons habe ich im Buch von Havil (2009) gefunden. Für jede Teilpopulation BC, B-C, -BC, -B-C wird der untersuchte Effekt durch einen Vektor dargestellt. Auf der x-Achse wird die Populationsgröße angetragen und auf der y-Achse der Effekt.

In unserem Beispiel sind das zunächst einmal die vier Vektoren $V(BC) = (200, 40)$, $V(B-C) = (100, 85)$, $V(-BC) = (200, 30)$ und $V(-B-C) = (400, 300)$. Die Vektoren der aggregierten Größen ergeben sich durch Vektoraddition: $V(B) = V(BC) + V(B-C) = (300, 125)$ und $V(-B) = V(-BC) + V(-B-C) = (600, 330)$. Die Steigungen der Vektoren entsprechen den jeweiligen relativen Effekten (Wirksamkeit).

Die dunkelblauen Vektoren gehören zu B (Medikament X) und die magentafarbenen zu -B (Medikament Y). Die Steigung des ersten blauen Vektors ist größer als die des ersten magentafarbenen. Dasselbe gilt für den jeweils zweiten Vektor. Bei den Steigungen der Summen-

vektoren kehren sich die Verhältnisse um. Hier sieht man ganz genau, wie das zustande kommt.



Weitere Beispiele

Eine Reihe von Beispielen zum simpsonschen Paradoxon enthält das Buch von Dubben und Beck-Bornholdt (2005). Ich beschränke mich hier auf mein Xenophobie-Beispiel und auf den wohl berühmtesten Fall zum Thema, den „Berkeley Sex Bias“-Fall.

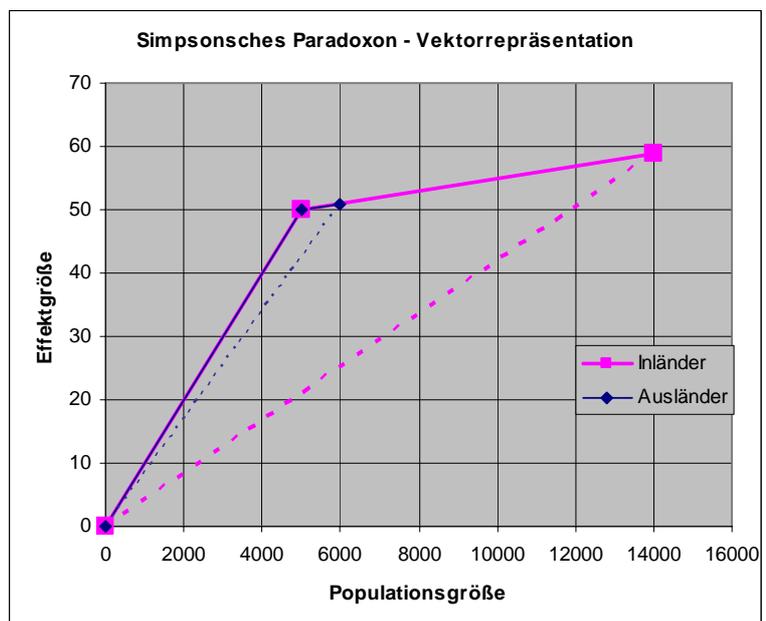
Xenophobie

Tabelle 4 enthält die Daten des [Xenophobie-Beispiels](#) in der oben vorgegebenen Standardform.

Hier empfiehlt sich die Vektorrepräsentation, da sich die Anzahl der Straftaten bezogen auf die Bevölkerung kaum als Wahrscheinlichkeit denken lässt. Denkbar ist ja, dass einzelne Subjekte mehrere Straftaten verüben.

Tabelle 4 Daten des Xenophobie-Beispiels

| | Aschental | | Innenstadt | |
|-----------|-----------------|-----------|-----------------|-----------|
| | Straftaten/Jahr | Einwohner | Straftaten/Jahr | Einwohner |
| Ausländer | 50 | 5000 | 1 | 1000 |
| Inländer | 50 | 5000 | 9 | 9000 |



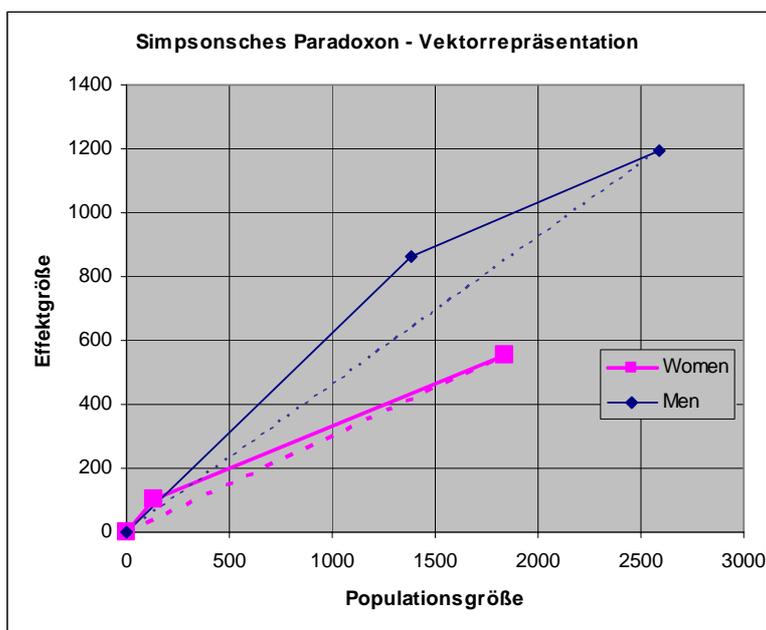
Frauendiskriminierung

Eines der bekanntesten Beispiele für das simpsonsche Paradoxon geht auf eine Diskriminierungsklage gegen die Universität von Kalifornien in Berkeley zurück. Es wurde darauf verwiesen, dass im Herbst 1973 die Aufnahmequote für Frauen im Schnitt niedriger lag als die für Männer. Aber bei genauerem Hinsehen entpuppte sich der Vorwurf als grundlos und es stellte sich heraus, dass die Frauen bevorzugt solche Fächer wählten, die geringe Aufnahmequoten hatten. Männer hingegen wählten überwiegend die weniger überlaufenen Fächer mit hohen Aufnahmequoten. Tabelle 5 ist eine Zusammenstellung der entsprechenden Daten (Wikipedia, Simpson's Paradox, Zugriff am 11. Juli 2010).

Tabelle 5 Bewerbungs- und Aufnahmezahlen Berkeley 1973

| | High admission rates (C) | | Low admission rates (-C) | |
|------------|--------------------------|------------|--------------------------|------------|
| | Admitted (A) | Applicants | Admitted (A) | Applicants |
| Men (B) | 864 | 1385 | 328 | 1205 |
| Women (-B) | 106 | 133 | 451 | 1702 |

Die folgende Grafik zeigt die zugehörige Vektorrepräsentation.



Die Wahrscheinlichkeitswerte sind in der Tabelle 6 wiedergegeben. Die hellen Felder sind direkt aus den Häufigkeiten ermittelte Werte. Grau unterlegt sind die Daten für die Gewichtungsfaktoren und die sich damit ergebenden Werte aufgrund der Formeln für bedingte Wahrscheinlichkeiten.

Tabelle 6 Wahrscheinlichkeitsinterpretation „Berkeley Sex Bias“

| | | | | | |
|-------------|-----|--------------|-----|------------|--------|
| $P(A BC)=$ | 62% | $P(A B-C)=$ | 27% | $P(A B)=$ | 46.02% |
| $P(A -BC)=$ | 79% | $P(A -B-C)=$ | 27% | $P(A -B)=$ | 30.33% |
| $P(C B)=$ | 53% | $P(-C B)=$ | 47% | $P(A B)=$ | 46.02% |
| $P(C -B)=$ | 7% | $P(-C -B)=$ | 93% | $P(A -B)=$ | 30.33% |

Was durch die Diskriminierungsklage und deren Aufarbeitung in Berkeley sichtbar wurde, ist auch anderswo zu beobachten: Frauen studieren tendenziell am wirtschaftlichen Bedarf vorbei: Je höher die Arbeitslosenquote eines Berufszweiges ist, desto mehr zieht er die Frauen an – gegenläufig zum Trend bei den Männern.

Bei Deutschlands Frauen sind Germanistik und Sozialwesen besonders beliebt. Und diese Fächer bereiten auf Berufe mit besonders hohen Arbeitslosenquoten vor. Männer zieht es hingegen vornehmlich zur Informatik, zu den Naturwissenschaften und zu den technischen Disziplinen. Dort sind die Aufnahmequoten günstiger – und auch die späteren Chancen auf eine Beschäftigung.

Quellen

Dubben, Hans-Hermann; Beck-Bornholdt, Hans-Peter: Mit an Wahrscheinlichkeit grenzender Sicherheit. Logisches Denken und Zufall. rororo science, Reinbek bei Hamburg 2005

Havi, Julian: Das gibt's doch nicht. Mathematische Rätsel. Spektrum Akademischer Verlag, Heidelberg 2009

Székely, Gábor J.: Paradoxa - klassische und neue Überraschungen aus Wahrscheinlichkeitsrechnung und mathematischer Statistik. Harri Deutsch, Thun und Frankfurt am Main 1990