ÜBER SIMPSONS PARADOXON

von Erhard Künzel, Eberfing

Zusammenfassung: Zuerst wird das klassische Simpsonparadoxon vorgestellt. Es folgt eine mathematische Analyse, unter welchen Bedingungen man mit diesem Paradoxon rechnen muß. Im dritten Absatz wird die "Auflösung" des Paradoxons nach der Literatur referiert, darauf wird diese Auffassung in geändertem Kontext geprüft und eine allgemeinere Interpretation vorgeschlagen.

Ziel dieses Aufsatzes ist es nicht, über gute, neue Erkenntnisse zu berichten. Vielmehr soll dieses merkwürdige Paradoxon Grund- und Leistungskursleitern bewußt gemacht werden. Es bietet sich zur Behandlung in der Übungsphase zu bedingten Wahrscheinlichkeiten an. Als "Paradoxen" hat es seinen eigenen Reiz. Der Clou besteht aber in den verschiedenen Interpretationsmöglichkeiten, die von den Rahmenbedingungen abhängen, die in das mathematische Modell nicht eingehen. Dadurch unterscheidet es sich auch vom Bertrandschen Paradoxon [1], bei dem die Paradoxie durch die Mathematisierung entsteht. Vielleicht kann der Artikel auch Grundlage für Facharbeiten oder Referate werden.

1. Ein klassisches Simpsonproblem

Ein Arzt will die Wirksamkeit einer neuen Behandlungsmethode (BII) mit der bisher üblichen Methode (BI) vergleichen. Er führt seine Versuche in zwei Orten α-ville und β-ville an Patienten durch. Σ-ville bezeichne beide Orte zusammen. Nach einem Jahr erhält er folgende Daten:

Tabelle 1:

Effiktivität der Behandlung	α-ville		β-ville		Σ-ville	
	BI	BII	BI	BII	BI	BII
effektiv	2	36	54	8	56	44
nicht-effektiv	8	54	36	2	44	56
Summe	10	90	90	10	100	100
relative Häufigkeit des Erfolgs	0.20	0.36	0.54	0.80	0.56	0.44

Sowohl in α-ville als auch in β-ville ist demnach die relative Häufigkeit des Behandlungserfolges bei der neuen Methode größer als bei der alten. Man möchte meinen, Methode II ist besser. Betrachtet man aber die zusammengefaßte Statistik (Σ-ville), so ist die Häufigkeit der erfolgreichen Behandlung bei Methode I größer. Ist die alte Methode I doch besser?

Mathematisierung: Sei Ω die Gesamtheit aller Patienten dieses Arztes aus α-ville oder β-ville, bei denen eine der beiden Methoden versucht wurde.

Ereignis A sei: >>Patient wohnt in α-ville.<<

$\bar{A}$ heißt dann: >>Patient wohnt in β-ville.<<

Ereignis B sei: >>Patient wird mit Behandlungsmethode BI verarztet.<<

$\bar{B}$ ist dann: >>Patient wird mit Methode BII behandelt.<<

Ereignis E trete ein, wenn die Behandlung erfolgreich ist.

Mit diesen Abkürzungen tritt das Interpretationsproblem von Simpson stets dann auf, wenn (h stehe für Häufigkeit)

$$h_{A\cap B}(E) < h_{A\cap\bar{B}}(E) \quad \text{und} \quad h_{\bar{A}\cap B}(E) < h_{\bar{A}\cap\bar{B}}(E) \quad \text{und} \quad h_B(E) > h_{\bar{B}}(E).$$

Allgemein: Für drei beliebige Ereignisse A, B, E ⊂ Ω nenne ich die gleichzeitige Gültigkeit der oberen drei Ungleichungen ein Simpsonparadoxon.

1. Einwand: "Es handelt sich um kein Paradoxon, da relative Häufigkeiten, aber keine Wahrscheinlichkeiten vorkommen. Die relativen Häufigkeiten streuen um die Wahrscheinlichkeitswerte. Mit Wahrscheinlichkeiten hätte man keine paradoxen Werte erhalten."

Dieser Einwand ist leicht zu entkräften. Man kann sich ohne weiteres jede Anzahl mit der gleichen beliebig großen Zahl multipliziert denken und damit bei beliebiger Signifikanz nachweisen, daß die drei Definitionsungleichungen des Simpsonparadoxon für relative Häufigkeiten nur mit den entsprechenden Ungleichungen für Wahrscheinlichkeiten verträglich sind. Auch kann man sich eine Paradoxie mit den Wahrscheinlichkeiten 0.2, 0.4, 0.6 und 0.8 konstruieren und diese als absolut gültig ansehen.

Beispiel: Betrachtet werden zwei Urnen U1 und U2, die jeweils mit roten und grünen Kugeln gefüllt sind. Beide Kugelsorten können als Aufschrift entweder noch einen Buchstaben "B" tragen oder nicht.

Tabelle 2: Urnenfüllungen

	Urne 1		Urne 2	
	B	kein B	B	kein B
rot	2	36	54	8
grün	8	54	36	2
Summe	10	90	90	10

Experiment: Wähle zuerst eine der beiden Urnen willkürlich aus (Wahrscheinlichkeit jeweils 0,5 / Ereignis A: "Urne 1 gewählt"), dann entnehme eine Kugel und notiere, ob ein Buchstabe drauf ist (Ereignis B: "mit Buchstaben") und die Farbe (Ereignis E: "Farbe ist rot"). Leicht verifiziert man für die drei "Definitionsungleichungen" mit:

$p_{A\cap B}(E) = 0.2$; $p_{A\cap\bar{B}}(E) = 0.4$; $p_{\bar{A}\cap B}(E) = 0.6$; $p_{\bar{A}\cap\bar{B}}(E) = 0.8$

und $p_B(E) = 0.56$; $p_{\bar{B}}(E) = 0.44$.

2. Einwand: "Die Zahlen sind ja so hingetrickst."
Tatsächlich taucht aber das Simpsonparadoxon in einer Statistik über die Diskriminierung von Frauen bei der Zulassung zu Graduiertenkursen auf dem Berkeley Campus auf [[3] nach [2] S. 141]. Da man als Lehrer aber gemeinhin kaum genügend echte Statistiken zur Verfügung hat, folgt vor der Entscheidung, welcher Methode man trauen darf, ein Exkurs über die Bedingungen, wann man ein Simpsonparadoxon erwarten kann.

2. Wann tritt das Simpsonparadoxon (kurz SP) auf?

Betrachtet werde ein beliebiger Ergebnisraum Ω und Ereignisse A, B, E $\subset \Omega$. Die Schnittmengen $A\cap B$, $\bar{A}\cap B$, $A\cap\bar{B}$ und $\bar{A}\cap\bar{B}$ seien nicht leer. Ausgangspunkt für unsere Betrachtungen: $h_{A\cap B}(E) < h_{A\cap\bar{B}}(E)$ und $h_{\bar{A}\cap B}(E) < h_{\bar{A}\cap\bar{B}}(E)$. Außerdem kann man $h_{A\cap B}(E) < h_{\bar{A}\cap B}(E)$ annehmen (sonst vertausche man A mit $\bar{A}$ d.h. im ersten Beispiel die Namen α-ville und β-ville).

Ein SP liegt genau dann vor, wenn $h_B(E) > h_{\bar{B}}(E)$ ist, d.h. wenn

$$\frac{|E\cap B|}{|B|} > \frac{|E\cap\bar{B}|}{|\bar{B}|}$$ d.h. wenn gilt:

$$\frac{h_{A\cap B}(E)\cdot h_B(A)\cdot|A| + h_{\bar{A}\cap B}(E)\cdot[1 - h_B(A)]\cdot|A|}{|A|} \qquad \text{[Gl. 1]}$$

$$> \frac{h_{A\cap\bar{B}}(E)\cdot h_{\bar{B}}(A)\cdot|\bar{A}| + h_{\bar{A}\cap B}(E)\cdot[1 - h_{\bar{A}\cap\bar{B}}(A)]\cdot|\bar{A}|}{|\bar{A}|}$$

Zur besseren Übersicht seien folgende Abkürzungen eingeführt.

$\alpha_1 = h_B(A)$, $\alpha_2 = h_{\bar{B}}(A)$;

$\omega_1 = h_{A\cap B}(E)$, $\omega_2 = h_{A\cap\bar{B}}(E)$, $\omega_3 = h_{\bar{A}\cap B}(E)$ und $\omega_4 = h_{\bar{A}\cap\bar{B}}(E)$

Damit wird aus [Gl. 1]:

$$\frac{\omega_1\alpha_1|A| + \omega_3(1 - \alpha_1)|A|}{|A|} > \frac{\omega_2\alpha_2|\bar{A}| + \omega_4(1 - \alpha_2)|\bar{A}|}{|\bar{A}|} \qquad \text{[Gl. 2]}$$

Satz 1:
Bei festen α_1, α_2 und ω_1 bis ω_4 ist das Eintreten des Simpsonparadoxons unabhängig von $|A|$ und $|\bar{A}|$.
Begründung: Kürzen liefert

$$\omega_1\alpha_1 + \omega_3(1 - \alpha_1) > \omega_2\alpha_2 + \omega_4(1 - \alpha_2). \qquad \text{[Gl. 3]}$$

Wir halten nun alle $\omega_1$bis ω_4 fest und studieren den Einfluß der α_1 und α_2, die wir dazu als völlig variabel in [0, 1] ansehen. Man erkennt, daß $\omega_1\alpha_1 + \omega_3(1 - \alpha_1)$ nach Wahl von α_1 irgend ein Element des Intervalls $[\omega_1, \omega_3]$ ist. Analog kann man durch α_2 den Wert beliebig variieren.

Also folgt mit [Gl. 3]: Man kann SP durch Wahl der α_i erzwingen, genau wenn $]\omega_1, \omega_3[\cap]\omega_2, \omega_4[\neq \{\}$ ist.

Drei Fälle sind denkbar, da nach Voraussetzung ist: $\omega_1 < \omega_2$, $\omega_3 < \omega_4$ und $\omega_1 < \omega_3$ ist.

1. $\omega_1 < \omega_2 < \omega_3 < \omega_4$ Hier ist die Schnittmenge nicht leer.
2. $\omega_1 < \omega_3 < \omega_2 < \omega_4$ Hier ist die Schnittmenge leer.
3. $\omega_1 < \omega_3 < \omega_4 < \omega_2$ Hier ist die Schnittmenge leer.

Satz 2:

Das Simpsonparadoxon ist durch die Wahl von α_1 und α_2 genau dann möglich, wenn $\omega_2 < \omega_3$ ist.

Ein Ziel dieses Abschnitts ist ja, die Konstruktion von Simpsonparadoxien zu ermöglichen. Da der Leser nun passende $\alpha_1$und α_2 zu vorgegebenen Omegas sicher selbst berechnen kann, wende ich mich lieber zwei übersichtlicheren und leichter handhabbaren Spezialfällen zu:

α_1 und α_2 seien komplementär, d.h. es gebe ein $\Delta\alpha$ so, daß $\alpha_1 = 0.5 - \Delta\alpha$ und $\alpha_2 = 0.5 + \Delta\alpha$ ist. Aus [Gl. 3]:

$$\omega_1(0.5 - \Delta\alpha) + \omega_3(0.5 - \Delta\alpha) > \omega_2(0.5 - \Delta\alpha) + \omega_4(0.5 - \Delta\alpha)$$

dh.

$$\Delta\alpha\,(-\omega_1 + \omega_3 - \omega_2 + \omega_4) > 0.5\,(-\omega_1 - \omega_3 + \omega_2 + \omega_4)$$

In der Situation, in der Simpson möglich ist ($\omega_1 < \omega_2 < \omega_3 < \omega_4$), heißt das:

$$\Delta\alpha > 0.5\,\frac{(\omega_2 - \omega_1) + (\omega_4 - \omega_3)}{(\omega_3 - \omega_1) + (\omega_4 - \omega_2)} \qquad [3]$$

Betrachtet man den Spezialfall äquidistanter Wahrscheinlichkeiten, d.h. es gebe ein $\Delta\omega$ so, daß $\omega_4 = \omega_3 + \Delta\omega = \omega_2 + \Delta\omega = \omega_1 + \Delta\omega$ ist, dann folgt:

$$\Delta\alpha > 0.5\,\frac{\Delta\omega + \Delta\omega}{2\Delta\omega + 2\Delta\omega} = 0.25$$

Satz 3:

Im Spezialfall komplementärer α_i und äquidistanter ω_i mit $\omega_1 < \omega_2 < \omega_3 < \omega_4$ tritt ein SP genau dann ein, wenn $\alpha_1 < 0.25$ und (folglich) $\alpha_2 > 0.75$ ist, unabhängig von allen anderen Größen.

Methode zur Konstruktion eines Simpsonparadoxons vor Schülern:

1. Wähle $|A|$ und $|\bar{A}|$ beliebig.
2. Wähle $\omega_1 < \omega_2 < \omega_3 < \omega_4$ äquidistant in]0, 1[.
3. Wähle $\alpha_1 < 0.25$ und α_2 komplementär.
4. Berechne die restlichen Häufigkeiten nach Tabelle 3.

Wenn nur α_1 und α_2 genügend extrem sind, dann kann man leicht ein SP erzeugen. Umgekehrt: Wenn nur ω_2 und ω_3 weit genug auseinander liegen - relativ $\omega_1$und ω_2 -, dann gibt es leicht ein SP.

Tabelle 3: Ausdruck für die totalen Häufigkeiten beim Simpsonparasoxon abhängig von den Variablen $|A|$, $|\bar{A}|$, α_1, α_2 und ω_1 bis ω_4.

	A		$\bar{A}$	
	B	$\bar{B}$	B	$\bar{B}$
E	$\omega_1\alpha_1\|A\|$	$\omega_2\alpha_2\|\bar{A}\|$	$\omega_3(1 - \alpha_1)\|A\|$	$\omega_4(1 - \alpha_2)\|\bar{A}\|$
$\bar{E}$	$(1 - \omega_1)\alpha_1\|A\|$	$(1 - \omega_2)\alpha_2\|\bar{A}\|$	$(1 - \omega_3)(1 - \alpha_1)\|A\|$	$(1 - \omega_4)(1 - \alpha_2)\|\bar{A}\|$
Su.	$\alpha_1\|A\|$	$\alpha_2\|\bar{A}\|$	$(1 - \alpha_1)\|A\|$	$(1 - \alpha_2)\|\bar{A}\|$

	Ω B	$\bar{B}$
E	$\omega_1\alpha_1\|A\| + \omega_3(1 - \alpha_1)\|A\|$	$\omega_2\alpha_2\|A\| + \omega_4(1 - \alpha_2)\|\bar{A}\|$
$\bar{E}$	$(1 - \omega_1)\alpha_1\|A\| + (1 - \omega_3)(1 - \alpha_1)\|\bar{A}\|$	$(1 - \omega_2)\alpha_2\|A\| + (1 - \omega_4)(1 - \alpha_2)\|\bar{A}\|$
Su.	$\|A\|$	$\|\bar{A}\|$

3. Interpretation nach Blyth

Welcher der relativen Häufigkeiten darf man nun zur Interpretation trauen? Kehren wir zum "Behandlungsproblem" zurück!

Nach Blyth [[4] in [2] S. 139] sind die Teilergebnisse verläßlicher als das Gesamtergebnis. Er begründet: Würde in α-ville nur Methode I verwendet, dann wären 20 Leute geheilt, in β-ville nur mit Methode I 60 Leute, macht insgesamt 80 Leute, das entspricht 40%. Nur mit Methode II wären es in α-ville 40, in β-ville 80, d.h. 120 Leute, entspricht 60%.

Oder anders: Das Ergebnis der Summe entsteht, weil die α-ville-Leute anscheinend schwerer erkrankt waren (schlechtere Behandlungserfolge bei beiden Methoden) und, wenn man die Methode II vorwiegend bei Schwerkranken heranzieht, so wird das Resultat kläglicher sein als das der schlechteren Methode bei leichter Erkrankten.

Kann man nun den Schluß ziehen, daß stets den Teilresultaten der Vorzug gegeben werden muß?

4. Sind Frauen hier benachteiligt?

Zur Analyse sei folgende Tabelle mit fast gleichen Zahlen, aber geändertem Kontext gegeben. Es werde untersucht, ob Frauen bei der Jobsuche im Berufsfeld X benachteiligt werden. Die Analyse wird für große Menschen (> 1.75m) und kleine (≤ 1.75m) getrennt durchgeführt. Annahme: 90% der Männer sind größer als 1.75m, aber 90% der Frauen sind kleiner.

Tabelle 4:

	Kleine Menschen		Große Menschen		Gesamt	
	Männer	Frauen	Männer	Frauen	Männer	Frauen
Job	20	360	540	80	560	440
Kein Job	80	540	360	20	440	560
rel. Häuf. für Erfolg	0.2	0.4	0.6	0.8	0.56	0.44

In der Gesamtheit gesehen werden Frauen benachteiligt. Trotzdem ist die Einstellungsquote sowohl bei kleinen, als auch bei großen Männern ungünstiger als bei den entsprechenden Frauen.

Interpretation:

Eine sinngemäße Übersetzung der Argumentation von Abschnitt 3 wäre:

Das Ergebnis der Summe entsteht, weil die Kleinen anscheinend schlechtere Einstellungsvoraussetzungen mitbringen (schlechtere Quoten bei Männern und Frauen), und wenn man Frauen vorwiegend unter den Kleinen heranzieht, so wird ihr Resultat kläglicher sein als das der Männer. Im Gegensatz zur letzten Aufgabe, wo - scheinbar? α_1 und α_2 willkürlich waren, ist das hier nicht der Fall! Männer sind nunmal im allgemeinen größer. Das Kriterium "besser Große", das entscheidend ist, benachteiligt Frauen stärker, als durch ihre besseren Teilwahrscheinlichkeiten ausgeglichen werden könnte. Hier wird man wohl die "Summenhäufigkeiten" interpretieren müssen, da an der Verknüpfung der Merkmale "Frau" und "eher klein" nichts geändert werden kann.

Also ist doch auf die Gesamtwahrscheinlichkeit mehr Verlaß? Übetragen wir zur Prüfung die gerade erfolgreiche Argumentation auf das Behandlungsproblem. Eine sinngemäße Abwandlung wäre:

"Das Kriterium "besser α-ville Patienten kurieren", benachteiligt Methode I stärker als durch die besseren Teilwahrscheinlichkeiten ausgeglichen werden könnte." Das ist nun tatsächlich

abstrus. Hier wird man wohl die Teilwahrscheinlichkeiten interpretieren, da die Verknüpfung "Methode 2" und "β-ville" jederzeit geändert werden kann.

Schlußfogerung:

Die Teilwahrscheinlichkeiten liefern das zu interpretierende Ergebnis, wenn die Verknüpfung der beiden Merkmale änderbar erscheint. Die Summenwahrscheinlichkeiten liefern das wichtigere Ergebnis, wenn der Zusammenhang der beiden Merkmale nicht veränderbar ist.

Ein letzes Beispiel soll dokumentieren, daß diese Interpretation sehr kontextsensibel ist:

Ein Reiseunternehmen möchte erforschen, ob die Werbung für einen Erholungsurlaub, den wir hier im Gegensatz zu einem Bildungsurlaub sehen, eher auf Männer oder auf Frauen abgestimmt werden soll. Das beauftragte demografische Institut macht Umfragen in zwei Kneipen, der "Grünen Zaubermühle" und dem "Goldenen Bräuhaus" und erhält folgendes Ergebnis:

Tabelle 5:

	Grüne Zaubermühle		Goldenes Bräuhaus		Gesamt	
	Männer	Frauen	Männer	Frauen	Männer	Frauen
Erholung	20	360	540	80	560	440
Bildung	80	540	360	20	440	560
rel. Häuf. für Erholung	0.2	0.4	0.6	0.8	0.56	0.44

Die Blythsche Argumentation: Die (verfälschte) Summe entsteht, weil in der Zaubermühle anscheinend viel mehr Bildungsurlauber sind als im Bräuhaus. Wenn man das Interesse der Frauen besonders bei den Zaubermühlenbesuchern testet, dann schneiden die Frauen natürlich (in der Neigung zum Erholungsurlaub) besonders schlecht ab. Werbung ist auf Frauen abzustimmen, da sich diese eher für Erholung interessieren.

Anti-Blythsche-Argumentation: Tatsächlich ist unter den Besuchern der Zaubermühle der Erholungsurlaub unbeliebt. Frauen gehören aber vorwiegend in diese Kategorie. Ihr geringfügig "besseres" Abschneiden in der Zaubermühle könnte entstehen, weil die Männer dort Bildungshungriger als der Durchschnitt sind. Analoge Argumentation ist im Bräuhaus möglich. Werbung für Erholungsurlaub richte sich also nach der Gesamtheit, falle also eher bei Männern auf fruchtbaren Boden.

Welcher Interpretation trauen Sie? Und wie würden Sie entscheiden, wenn die Wirtshäuser anders hießen (z.B. "Bei Harry" und "Bei Larry")?

Über Zuschriften mit weiteren Beispielen für das Auftreten des Simpsonparadoxons würde ich mich freuen. Ich möchte sie sammeln. Schreiben Sie mir doch, wenn ihnen dazu etwas einfällt.

Literatur

[1] Barth, F. und R. Haller: *Stochastik Leistungskurs*, 3. Auflage, Ehrenwirth Verlag, München 1986.

[2] Faletta, N. : *Fischer Logo 8702: Paradoxon*, Fischer Taschenbuch Verlag GmbH, Frankfurt 1988.

[3] Bickel, P. J. , E. A. Hamel und J. W. O'Connell: "Sex Bias in Graduate Admissions: Data from Berkeley", *Science,* Febuary 1975, S. 398-403.

[4] Blyth, C. : "On Simpson's Paradoxon and the Sure-things Principle and some Probability Paradoxes in Choice from Among Random Alternatives", *Journal of the American Statistical Association*, June 1972, S. 364-373.

Adresse des Autors: Erhard Künzel, Am Mühlbach 7, 8121 Eberfing.