Ein alternder Playboy, die Medien und eine fragwürdige Statistik: eine kleine Anregung für den Unterricht

Raphael Diepgen, Ruhr-Universität Bochum

Stochastikunterricht auf der Schule, so der Tenor vieler Empfehlungen und Richtlinien, habe auch den möglichen Mißbrauch von Statistik kritisch zu beleuchten. Dazu empfiehlt sich die Beschäftigung mit entsprechenden „mathematikhaltigen" aktuellen Texten. Einen solchen fand ich Anfang letzten Jahres in der - allgemein als seriös geltenden - Wochenzeitung DIE ZEIT (Nr. 2 vom 03.01.1997, S. 52), nämlich einen Beitrag mit dem Titel „Die Sterne der Liebe. Eine Statistik der Eheschließungen zeigt: So ohne ist Astrologie nicht" aus der Feder des „Diplom-Mathematikers" Gunter Sachs. Ja, richtig, es handelt sich bei diesem Autor tatsächlich um jenen berühmten Gunter Sachs, Erbe eines gigantischen Vermögens, Ex-Ehemann von Brigitte Bardot, Fotograf leichtbekleideter Damen und und und. Kurzum: Ein schillernder Autor, der besonderes SchülerInneninteresse verspricht, ebenso wie selbstverständlich auch die spannende Thematik Liebe und Astrologie. Der vielversprechenden unterrichtlichen Verwendung wegen sei der - unterhaltsam geschriebene - Text daher vollständig wiedergegeben:

> Von Sigmund Freud bis zu Hedwig Courths-Mahler haben viele diesem Phänomen auf die Spur zu kommen versucht: Was bringt Menschen dazu, sich auf Dauer an einen Partner zu binden? Eheschließung setzt schließlich einen starken Trieb voraus, stark genug, die Hemmung vor einem Verlust an Freiheit zu überwinden. Große Geister haben vor dem Akt der Heirat gewarnt. Schopenhauer mahnte: „In unserem monogamischen Weltteile heißt heiraten seine Rechte halbieren und seine Pflichten verdoppeln." Der Dramatiker Christian Grabbe vertrat die Auffassung: „Heiraten, das heißt Nachtigallen zu Stubenvögeln zu machen." Und Goethe, auf dessen Urteil man nicht verzichten sollte, meinte: „Leider haben überhaupt die Heiraten ... etwas Tölpelhaftes: sie verderben die zartesten Verhältnisse." Aber hörte Goethe auf Goethe? Nein, er heiratete.
>
> Auch wenn der Autor nach 27 glücklichen Ehejahren den großen Geistern nicht in allem beipflichten kann, hat er sich doch gefragt: Was macht uns zu Stubenvögeln? Selbstverständlich gibt es klassische - oft heftig umstrittene - Gründe. Die Liebe. Die Leidenschaft. Die Mitgift. Die Konvention. Die vorzeitig sich ankündigende Nachkommenschaft. Daneben gibt es einen Grund, der sich nicht nur der Ratio entzieht, sondern auch dem bewußten Gefühl: die unbewußte Anziehungskraft zwischen zwei Menschen. Manche sprechen von Ma-

gnetismus, andere von einer nicht näher erklärbaren Macht, wieder andere versteigen sich gar zur Behauptung, Ehen würden im Himmel geschlossen.

Wie auch immer: Es scheint ein Agens zu geben, das bei der Entscheidung für eine dauerhafte Partnerschaft eine wichtige Rolle spielt. Nach jahrtausendealter Meinung der Astrologen ist diese Anziehungskraft zwischen zwei Menschen zwar nicht ein direkt dem Himmel zu verdankendes Geschenk, doch geprägt von dem zum Zeitpunkt der Geburt errechneten Stand der Gestirne. In einer Zeit, die vom Glauben zum Wissen als oberster Maxime gewechselt hat, wird diese Meinung vielfach abgelehnt, bestenfalls belächelt.

In dem vom Autor in der Schweiz gegründeten IMWA-Institut (Institut zur empirischen und mathematischen Untersuchung des möglichen Wahrheitsgehaltes der Astrologie) haben wir den Versuch unternommen, die von der Astrologie postulierten Phänomene mit wissenschaftlichen Mitteln zu prüfen, also festzustellen, ob diese Phänomene sich bei näherer Betrachtung als Irrglauben entpuppen oder nach den strengen mathematischen Kriterien der Statistik als nachweisbar existent anerkannt werden müssen. Eine der ersten Fragen, denen wir uns zuwandten, war, ob die Anziehungskraft zwischen zwei Menschen nachprüfbar von deren Sternzeichen (Sonnenzeichen) beeinflußt sein könnte.

Die Untersuchung wurde möglich durch Unterstützung des Bundesamtes für Statistik in Bern, dessen Abteilung für Bevölkerungsentwicklung jede Eheschließung in der Schweiz seit 1987 mit den Geburtsdaten der Partner registriert. Freundlicherweise stellte das Bundesamt diese Daten über alle Eheschließungen zur Verfügung, die in den Jahren 1987 bis 1994 in der Schweiz geschlossen wurden: 717 526 Frauen und Männer konnten so ihrem Sternzeichen zugeordnet werden.

Wenn man unterstellt, daß die Zugehörigkeit zu einem bestimmten Sternzeichen keinerlei Einfluß auf die Partnerwahl hat, müßte sich die Zahl der Eheschließungen in etwa gleichmäßig auf die 144 möglichen Kombinationen verteilen. Wobei zu berücksichtigen ist, daß Widder, Stiere, Zwillinge und die anderen im heiratsfähigen Alter in erheblich unterschiedlicher Zahl existieren. Es ist bekannt, daß - jedenfalls auf der nördlichen Erdhälfte - im Frühjahr wesentlich mehr Kinder geboren werden als im Herbst. Da die Widder unter den Sternzeichen überproportional vertreten sind, sind natürlich auch erheblich mehr von Widdern eingegangene Ehen zu erwarten.

Die 1990 vorgenommene jüngste Volkszählung in der Schweiz ermöglichte die prozentuale Aufgliederung nach Sternzeichen. Wir haben die Daten des Bundesamtes für Statistik mit den über die Volkszählung erhaltenen Angaben über die Bezugspopulation ausgewertet und die Resultate Statistikern der Ludwig-Maximilians-Universität in

München zur Verfügung gestellt mit der Bitte, das Material auf Auffälligkeiten hin zu untersuchen, die bei der Partnerwahl auf einen Einfluß der Zugehörigkeit zum Sternzeichen schließen lassen könnten.

Es wurde dann mit Hilfe einer Tabelle zunächst statistisch errechnet, wie viele Eheschließungen pro möglicher Sternzeichenkombination zu erwarten waren. Die *geschlossenen* Ehen wurden mit den in der Tabelle zu *erwartenden* Ehen verglichen. Die Differenz gab Aufschluß darüber, ob signifikante Unterschiede vorlagen. So errechneten die Statistiker zum Beispiel, daß die Verbindung Stier-Frau / Wassermann-Mann aus den Anteilen an der Gesamtzahl der Heiratenden 2705mal zu erwarten wäre.

Da sich das Leben nicht so exakt nach den Daten richtet, wie es Statistiker gern hätten, ist auch bei großen Datenmengen mit Abweichungen zu rechnen. Nach wissenschaftlichen Kriterien kann festgelegt werden, welche Abweichungen noch als normal anzusehen sind, welche als signifikant oder sogar hochsignifikant zu gelten haben. Um bei dem Beispiel mit der Kombination von Stier-Frau und Wassermann-Mann zu bleiben: Wäre es in den acht Jahren nicht zu den erwarteten 2705 Ehen dieser Art gekommen, sondern etwa zu 2740 oder aber nur 2685, so wäre das gerade noch im Rahmen des Zufalls geblieben. Aber es fanden in der Schweiz nur 2544 Stier-Frauen und Wassermann-Männer zueinander. Und dieses Ergebnis nennen Statistiker hochsignifikant. Ein solches Minus gilt als ein derart starker Ausreißer aus der statistischen Norm, daß ein zusätzlicher Faktor eine Rolle gespielt haben muß.

Wir versuchten zu prüfen, ob tatsächlich die Zugehörigkeit zu einem bestimmten Sternzeichen dieses gesuchte Agens, diese unerklärbare Kraft, ist. Dazu zerlegten wir ein Jahr in 52 Wochen und setzten es nach dem Zufallsprinzip neu zusammen, unterteilten es gewissermaßen in zwölf fiktive Sternzeichen. Die Frage war, ob sich, nachdem wir dieselben Berechnungen wie für die echten Sternzeichen ausgeführt hatten, auch hier spürbare Abweichungen oder gar Signifikanzen einstellen würden. Also haben wir die gleiche Untersuchung mit diesem neuen, künstlichen Sternzeichenjahr durchführen lassen. Interessanterweise kam es hier zu keiner Signifikanz und lediglich zu Abweichungen im statistischen Normalbereich.

Wir standen vor dem eindrucksvollen Ergebnis: Die Berechnung von rund 358 000 Eheschließungen mit nicht verschobenen Sternzeichen ergab sechzehn Hochsignifikanzen oder Signifikanzen. Und auf der anderen Seite: Die Berechnung von rund 358 000 Eheschließungen mit verschobenen, also künstlichen Sternzeichen ergab Abweichungen, wie sie von Statistikern als normal angesehen werden.

Diese Erkenntnisse - mit einer extrem langen Datenkette von über 700 000 Ehepartnern - lassen den Schluß zu, daß zwischen den Sternen beziehungsweise dem zyklischen Verlauf unseres Sonnensystems und

den Eheschließungen Schweizer Bürger zwischen 1987 und 1994 ein Zusammenhang besteht.

Aus den Berechnungen geht hervor: Die Sternzeichenkombinationen, die - hochsignifikant oder signifikant - größere Zurückhaltung bei der Wahl des Ehepartners bewirken und somit zu weniger Heiraten führten, als zu erwarten gewesen wäre, sind: Wassermann-Mann / Stier-Frau, Löwe-Mann / Wassermann-Frau, Wassermann-Mann / Skorpion-Frau, Stier-Mann / Löwe-Frau, Krebs-Mann / Widder-Frau.

Die Kombinationen, die - hochsignifikant oder signifikant - zusätzliche Anziehungskraft auf die Wahl des Ehepartners ausüben und somit zu mehr Heiraten führten, als zu erwarten gewesen wäre, sind: Wassermann-Mann / Wassermann-Frau, Steinbock-Mann / Steinbock-Frau, Widder-Mann / Widder-Frau, Zwilling-Mann / Zwilling-Frau, Schütze-Mann / Widder-Frau, Jungfrau-Mann / Jungfrau-Frau, Stier-Mann / Waage-Frau, Schütze-Mann / Schütze-Frau, Löwe-Mann / Widder-Frau, Fische-Mann / Skorpion-Frau, Skorpion-Mann / Fische.

Auffallend ungern binden sich Steinbock-Männer, Stier-Frauen und Waage-Männer. Eine signifikante Ehefreudigkeit beweisen dagegen Fische-Männer, Krebs-Frauen und Löwe-Frauen. Die Zahlen über die Eheschließungen 1995, die wir zur Zeit auswerten, erhärten dieses Ergebnis. Und sie bestätigen unsere Untersuchungen im Zusammenhang mit zwölf anderen Komplexen, zu denen unter anderem Suizide, Strafhandlungen, Berufswahl, Verkehrsdelikte gehören. Zahlreiche Hochsignifikanzen und Signifikanzen auch hier.

Warum es aufgrund der Sternzeichenzugehörigkeit zu einzelnen, mit dem Zufall nicht mehr zu erklärenden Abweichungen von der zu erwartenden Norm kommt, ist Spekulation über die Natur der Astrologie.

Uns interessiert allein die Frage, ob Astrologie lediglich ein Mythos ist, der sich seltsamerweise über Jahrtausende und in fast allen Kulturen der Menschheit bis heute erhalten hat, oder ob es sich um ein - bisher nicht zu erklärendes - Phänomen handelt, dessen Vorhandensein nachweisbar ist. Wir halten diesen Nachweis für erbracht.

Soweit der vollständige Text von Gunter Sachs. Klar, daß er mit dieser publikumswirksamen Thematik in verschiedensten Medien präsent und insbesondere gerngesehener Gast in Talkshows war, in denen er mit intellektueller Halbbrille den halbherzigen Versuch inszenierte, wenigstens „ein bißchen" von der „komplizierten" mathematischen Statistik zu erläutern, bis er lächelnd angesichts der offensichtlichen Vergeblichkeit dieser Bemühung aufgeben und sich wieder den Fragen der Moderation nach seinem bewegten und interessanten Vorleben widmen konnte. Und selbstverständlich liegen nach dieser PR-Vorbereitung inzwischen Gunter Sachs´ Erkenntnisse

über die Wirkung der Sterne auf dies und das als heftig beworbenes Buch vor. Hier winkt ein gutes Geschäft.

Der Text läßt sich im Unterricht in vieler Hinsicht analysieren, auch noch unterhalb der im eigentlichen Sinne statistischen Ebene. Selbst wenn man die inferenzstatistische Schlußweise von Gunter Sachs akzeptierte, so wären seine Befunde natürlich auch auf ganz andere Faktoren als den Einfluß der Sterne zurückzuführen, insbesondere auf den hier naheliegenden Mechanismus der selbsterfüllenden Prophezeiung: Wenn auch nur ein geringer Anteil der Menschen an Astrologie glaubt und sich bei der Partnerwahl entsprechend verhält, sind - insbesondere angesichts der riesigen Stichprobe - die „signifikanten", vom Ausmaß zumeist wahrscheinlich geringfügigen Abweichungen Gunter Sachs zu erwarten. Das Konzept der self-fulfilling-prophecy sollte jedem Schüler einer allgemeinbildenden Schule vertraut sein - eher aus dem gesellschaftswissenschaftlichen denn mathematisch-naturwissenschaftlichen Unterricht -, ebenso wie aus dem Stochastikunterricht die Erkenntnis, daß bei sehr großen Stichproben auch minimale Abweichungen von der Nullhypothese mit hoher Wahrscheinlichkeit signifikant werden. Beides läßt sich in Auseinandersetzung mit diesem Text noch einmal vergegenwärtigen (vgl. Kasten 5). Und schließlich kämen als erklärende Faktoren für Gunter Sachs Befunde grundsätzlich alle Variablen in Frage, die mit den Sternzeichen - also der jahreszeitlichen Lage der Geburtstage - korrelieren.

Mathematisch interessanter ist aber die Analyse der inferenzstatistischen Strategie von Gunter Sachs, die angesichts ihrer rhetorisch geschickten Umschreibung für viele Schüler auf den ersten Blick durchaus überzeugend erscheinen mag: Augenscheinlich hat Gunter Sachs für alle möglichen 144 Sternzeichenkombinationen jeweils in einem „Signifikanztest" geprüft, ob die tatsächliche Häufigkeit der betreffenden Kombination (erfaßt in der entsprechenden Eheschließungsstatistik des Berner Bundesamtes der Jahre 1987 bis 1994) ihrer unter der Annahme der Unabhängigkeit des Heiratsverhaltens vom Sternzeichen zu erwartenden Häufigkeit (bestimmt aus der Schweizer Volkszählungsstatsitik für das Jahr 1990) entspricht. Hier fragt sich zunächst, inwieweit es für diese Prüfung überhaupt inferenzstatistischer Verfahren bedarf: Denn es ist nicht ganz einsichtig, in welchem Sinne die Totalerhebung der heiratswilligen Schweizer Bürger von 1987 bis 1994 eine Zufallsstichprobe aus einer Population sein soll, wie in der Inferenzstatistik grundsätzlich vorausgesetzt. Für den von Gunter Sachs gezogenen Schluß, „daß zwischen den Sternen beziehungsweise dem zyklischen Verlauf unseres Sonnensystems und den Eheschließungen Schweizer Bürger zwischen 1987 und 1994 ein Zusammenhang besteht", bedarf es angesichts dieser Totalerhebung jedenfalls keiner Inferenzstatistik. Anders formuliert: Solange sich Hypothesen auf relative Häufigkeiten und nicht auf durch diese zu schätzende Wahrscheinlichkeiten beziehen, ist Inferenzstatistik gänzlich überflüssig. Dennoch werden heute als „wissenschaftlicher Standard" ritualisierte inferenzstatistische Verfahren angewandt auch dann, wenn ein Schluß von einer Zufallsstichprobe auf eine Population im Ernst überhaupt nicht beabsichtigt zu sein scheint, wenn es also gar nicht gilt, die Wahrscheinlichkeiten möglicher Fehlschlüsse zu kontrollieren.

Gestehen wir aber Gunter Sachs zu, die heiratswilligen Schweizer Bürger der Jahre 1987 bis 1994 seien irgendwie als eine Zufallsstichprobe aus einer interessierenden Population zu interpretieren, Inferenzstatistik sei also zu Recht angewandt. Dann stoßen wir sofort auf folgendes Problem: Gunter Sachs überprüft die *eine* Nullhypothese der Unabhängigkeit des Heiratsverhaltens von den Sternzeichen in 144 einzelnen Signifikanztests. „Signifikante" oder „hochsignifikante" - differenziert vermutlich nach den Signifikanzniveaus von 5 % und 1 % - Ergebnisse in diesen Einzeltests interpretiert Gunter Sachs als Beleg für die Abhängigkeit des Heiratsverhaltens von den Sternzeichen; anders kann man ihn kaum verstehen. Daß aber von 144 durchgeführten Signifikanztests einer oder einige signifikant werden, dies dürfte auch dann noch eine nennenswerte Wahrscheinlichkeit haben, wenn für alle einzelnen Tests die Nullhypothese gilt. Schließlich würde man hier schon aufgrund der fünfprozentigen Alpha-Fehlerwahrscheinlichkeit durchschnittlich allein 5 % von 144, also 7 bis 8 signifikante Befunde erwarten. Hier liegt der Verdacht nahe: Gunter Sachs nutzt die klassische Alpha-Fehler-Inflation, indem er sich gleichsam 144 mal die Chance gibt, die *eine* inhaltliche Nullhypothese der Unabhängigkeit der Heiraten von den Sternzeichen aufgrund eines „signifikanten" Befundes zu verwerfen (vgl. Kasten 3). Begleitet wird dies von einer unscharfen Sprechweise, in der Signifikanz nicht im Sinne von Neyman und Pearson als eine Entscheidung gegen eine Nullhypothese (in einem auf lange Sicht konzipierten Entscheidungsverfahren) erscheint (vgl. Kasten 2), ja noch nicht einmal im rudimentären Sinne von Fisher als Überschreitungswahrscheinlichkeit (vgl. Kasten 1), sondern in einem vagen und kaum mathematisierbaren Sinne einer Abweichung von einem statistischen „Normalbereich".

Die Schüler können hier am konkreten Beispiel ein Bewußtsein für folgende Dinge entwickeln: „Signifikanztests" sind ziemlich problematisch, wenn nicht *vorab spezifische* Nullhypothesen formuliert werden. Es ist problematisch, eine einzige inhaltliche Nullhypothese dadurch zu „testen", daß man sie in eine Vielzahl von statistischen Nullhypothesen aufspaltet und dann ein zufallskritisches Verwerfen einer oder einiger weniger dieser Nullhypothesen als zufallskritisches Verwerfen der einen inhaltlichen Nullhypothese interpretiert. Eine vom ganzen Entscheidungs- und Handlungskontext isolierte geringe Überschreitungswahrscheinlichkeit oder „Signifikanz" besagt kaum etwas. Diese Erkenntnis kommt im klassischen Stochastikunterricht leider häufig zu kurz, weil er die Ebene eines einzelnen Tests zumeist nie verläßt.

Natürlich könnte man ausgehend von dieser Betrachtung die Idee der Alpha-Fehler-Adjustierung bei multiplem Testen entwickeln (vgl. Kasten 4). Und falls die Schüler aus dem vorangegangenen Stochastikunterricht Chi-Quadrat-Techniken kennen sollten, werden sie sicherlich sofort verwundert und mit Recht danach fragen, warum denn Gunter Sachs die Nullhypothese von der Unabhängigkeit der Heiraten von den Sternzeichen nicht in Form eines einzigen Chi-Quadrat-Testes auf der Basis der entsprechenden 12x12-Kontingenztafel geprüft hat; genau für solche Fragestellungen ist der Chi-Quadrat-Test schließlich doch überhaupt entwickelt worden. Skeptische Vermutung: Dieser Test wäre vielleicht nicht signifikant geworden, und das Geschäft mit den „wissenschaftlichen" Belegen für die Astrologie wäre damit gestorben (vgl. Kasten 6). Naheliegend wäre hier lebendige Demonstration durch ein kleines Schülerprojekt: Jeder Schüler befragt „zufällig" 100 Paare nach den Sternzeichen der beiden Partner; die Daten werden gesammelt und dann zweifach ausgewertet: Einerseits in Form eines einzigen Chi-Quadrat-Testes auf Basis der erhaltenen 12x12-Kontingenztafel, andererseits entsprechend der Strategie von Gunter Sachs in 144 Einzeltests.

Hier stößt man noch auf folgendes Problem: Die erwarteten Häufigkeiten für die 144 Sternzeichenkombinationen errechnet Gunter Sachs nicht etwa - wie beim klassischen Chi-Quadrat-Unabhängigkeitstest - aus den entsprechenden Randhäufigkeiten der 12x12-Kontingenztafel, sondern augenscheinlich aus den Sternzeichenhäufigkeiten in der Schweizer Gesamtbevölkerung gemäß Volkszählung von 1990. Wenn es etwa 1990 in der männlichen Schweizer Bevölkerung 10 % Jungfrauen und in der weiblichen Schweizer Bevölkerung 6 % Jungfrauen gegeben haben sollte, hätte Gunter Sachs für die Kombination Jungfrau-Mann / Jungfrau-Frau 10 % mal 6 % gleich 0,6 % der untersuchten 358.763 Paare erwartet, also 2.152 Paare. Wenn nun aber tatsächlich "signifikant" mehr Paare Jungfrau-Mann / Jungfrau-Frau beobachtet worden sein sollen, so muß dies nicht unbedingt daran liegen, daß Jungfrau-Männer Jungfrau-Frauen anderen Frauen vorziehen - und / oder umgekehrt. Es könnte, vom trivialen Zufall einmal abgesehen, auch daran liegen, daß Jungfrau-Männer und / oder Jungfrau-Frauen mehr als andere nach der Ehe streben - egal mit wem -, daß sich also beispielsweise unter den untersuchten Heiratenden 11 % Jungfrau-Männer und / oder 7 % Jungfrau-Frauen fanden. Die Nullhypothese der 144 Tests von Gunter Sachs, also die Annahme für die Berechnung der erwarteten Häufigkeiten, ist nämlich eine doppelte: Erstens sind in der Subpopulation der Heiratenden die Sternzeichen der beiden Partner stochastisch unabhängig voneinander, *und* zweitens entspricht die Verteilung der Sternzeichen in den Subpopulationen der heiratenden Männer und der heiratenden Frauen der Verteilung in den Gesamtpopulationen der Männer und der Frauen. Das Verwerfen dieser zweiteiligen Nullhypothese läßt grundsätzlich offen, welcher Teil unplausibel geworden ist. Dies dürfte Gunter Sachs freilich nicht sonderlich stören: Denn er würde sicherlich das Verwerfen beider Teile auf die Sterne zurückführen; auch die fiktiv skizzierte überdurchschnittliche Heiratsfreudigkeit von Jungfrauen wäre für ihn also Wirken der Sterne. Insofern ist es inkonsequent, daß Gunter Sachs den zweiten Teil dieser schon in 144 Tests überprüften Nullhypothese augenscheinlich noch einmal separat überprüft hat, nämlich in Form von 2 (Geschlechter) mal 12 (Sternzeichen) gleich weiteren 24 Signifikanztests jeweils zu der Frage, ob das betreffende Sternzeichen bei den Heiratenden genauso stark vertreten ist wie in der Gesamtbevölkerung betreffenden Geschlechts, mündend in der Feststellung: "Auffallend ungern binden sich Steinbock-Männer, Stier-Frauen und Waage-Männer. Eine signifikante Ehefreudigkeit beweisen dagegen Fische-Männer, Krebs-Frauen und Löwe-Frauen." Auch hier geht´s augenscheinlich nur darum, mit der noch einmal gesteigerten Anzahl durchgeführter Tests die Chance auf einige weitere signifikante Befunde noch einmal zu erhöhen. (Auch hier: Die jeweils 12 Tests pro Geschlecht wären natürlich jeweils durch einen Chi-Quadrat-Test zu ersetzen gewesen.) Und ebenso wie oben gilt: Sollten tatsächlich Unterschiede in der Sternzeichenverteilung zwischen der Population der Heiratenden in der Schweiz 1987 bis 1994 - weit überwiegen dürften

dort die wenigen Geburtsjahrgänge 1960 bis 1970 - und der Schweizer Gesamtbevölkerung von 1990 bestehen, so müssen diese nicht an den Sternen liegen; sie können ihre Ursache auch in vielen damit korrelierenden Phänomenen haben, beispielsweise in über die Generationen hinweg veränderten Vorlieben für bestimmte "Zeugungszeiten" (etwa: Winterurlaubsnächte statt Sommerurlaubsnächte). Dies gilt insbesondere angesichts der riesigen Stichprobengröße von rund 360.000 für jeden dieser 24 zusätzlichen Tests, die auch kleine Unterschiede mit hoher Wahrscheinlichkeit signifikant werden läßt.

Lehrreich und verblüffend ist auch die Diskussion der von Gunter Sachs durchgeführten „Tests" auf der Basis entsprechend der Nullhypothese randomisierter Sternzeichen. Die Suche nach dem Sinn dieser einmaligen Zufallssimulation führt so oder so ins Leere: Mißtraut Gunter Sachs etwa der mathematischen Logik der von ihm benutzten 144 Tests, also den für diese Tests analytisch abgeleiteten Fehlerwahrscheinlichkeiten, und will er sich dieser daher empirisch durch Simulation vergewissern? Oder will er dem mathematisch ungebildeten Publikum durch Simulation weismachen, es seien keine signifikanten Ergebnisse zu erwarten, wären die Sternzeichen der Ehepartner voneinander unabhängig? Wie auch immer: In jedem Falle hätte er - computergestützt - eine große Vielzahl von Simulationen durchführen müssen, um die nullhypothetischen Fehlerwahrscheinlichkeiten durch relative Häufigkeiten schätzen zu können. Die einmalige Simulation von Gunter Sachs ist hier so sinnvoll wie das einmalige Werfen eines Würfels zur Bestimmung der Wahrscheinlichkeit einer Augenzahl. Daß also von seinen einmal durchgeführten 144 Tests auf der Basis der nach Zufall ein einziges Mal zugeordneten fiktiven Sternzeichen keiner signifikant wurde, besagt schlicht gar nichts - und schon gar nicht das, was sich Gunter Sachs davon augenscheinlich erhofft. Ironischerweise könnte es vielmehr erhebliche Zweifel an der Ehrlichkeit von Gunter Sachs wecken: Denn daß von 144 Signifikanztests bei exakter - nämlich durch Randomisierung sichergestellter - Geltung aller Nullhypothesen trotz zugestandener Alpha-Fehlerwahrscheinlichkeit von fünf Prozent kein einziger signifikant wird, ist ziemlich unwahrscheinlich (vgl. Kasten 3). Und noch unwahrscheinlicher erscheint das Ausbleiben eines signifikanten Ergebnisses, wenn man bedenkt, daß diese Simulation für beide Geschlechter eine Gleichverteilung der Sternzeichen erzeugt haben müßte, eine solche Gleichverteilung aber in der für die Berechnung der erwarteten Häufigkeiten herangezogenen Schweizer Gesamtpopulation - wie von Gunter Sachs eingangs ausdrücklich erwähnt - gar nicht vorliegt.

Fazit: Wenn die Argumentation von Gunter Sachs irgend etwas zeigt, dann sicherlich nicht die Abhängigkeit der Lebensschicksale von den Sternzeichen; eher schon die Abhängigkeit der Lebensmöglichkeiten vom ererbten Vermögen: Denn als „normaler" Diplom-Mathematiker hätte Gunter Sachs seine Argumentation wohl kaum publizieren können, schon gar nicht in der großen Öffentlichkeit.

Die Schüler lernen an diesem Beispiel, daß es in der öffentlichen Diskussion und den großen Medien durchaus möglich ist, ohne jedes Risiko „mathematikhaltige" Beiträge zu veröffentlichen, die sich nach kritischer Rekonstruktion ihres mathematischen Kerns als ziemlich grober Unsinn erweisen. Und daß es möglich ist, für diesen Unsinn das Gütesiegel einer angesehenen Universität zu erhalten. Daß also die Rolle von Mathematik und Mathematikern im öffentlichen Diskurs eine aufklärerische sein kann, aber durchaus nicht unbedingt sein muß.

Nachtrag: Natürlich hätte Gunter Sachs versuchen können, seine inferenzstatistische Argumentation zu retten, indem er zumindest ex post im Fisherschen Sinne die nullhypothetische Überschreitungswahrscheinlichkeit für seine Empirie abgeschätzt hätte, also die Wahrscheinlichkeit, im Falle der Unabhängigkeit von Heiraten und Sternzeichen mindestens 16 signifikante Ergebnisse in den Einzeltests zu erhalten (vgl. Kasten 7). Er tat es aber nicht. Und selbst wenn er es erfolgreich getan hätte, bliebe im Sinne von Neyman und Pearson die Frage, für welchen "course of action" damit die Entscheidung gefallen wäre. Ich fürchte, nur für ein schlichtes Achselzucken.

Kasten 1: Mit **Fisher** wird gemeinhin - unbeschadet der Frage historischer Richtigkeit und sehr vereinfachend formuliert - die Idee verbunden, eine „altes" Wissen repräsentierende Nullhypothese wegen Unplausibilität aufzugeben, wenn in einem entsprechenden Zufallsexperiment ein Datum eingetreten ist, für das unter der Nullhypothese die Überschreitungswahrscheinlichkeit hinreichend klein ist. Überschreitungswahrscheinlichkeit meint die nullhypothetische Wahrscheinlichkeit dafür, daß das beobachtete oder ein noch extremeres, also noch mehr von dem unter der Nullhypothese zu Erwartenden abweichendes Datum eintritt. Aus dieser Idee entwickelte sich die in empirischen Wissenschaften weitverbreitete Praxis des Signifikanztestens.

Kasten 2: **Neyman** und **Pearson** haben den Fisherschen Signifikanztest rekonstruiert als ein vor dem Zufallsexperiment fixiertes regelhaftes Entscheidungsverfahren zur datengestützten Entscheidung zwischen zwei konkurrierenden Hypothesen, typischerweise einer Null- und einer Alternativhypothese, mit folgenden Eigenschaften: Die nullhypothetische Wahrscheinlichkeit für eine Fehlentscheidung der ersten Art, nämlich irrtümlich zuungunsten einer wahren Nullhypothese, ist durch eine vorzuwählende Grenze Alpha beschränkt - daher auch die Bezeichnung Alpha-Fehler -, und die üblicherweise nur als Funktion beschreibbare alternativhypothetische Wahrscheinlichkeit für eine Fehlentscheidung zweiter Art, nämlich irrtümlich zugunsten einer falschen Nullhypothese, genügt gewissen Minimalisierungsforderungen. Die Entscheidung zwischen den beiden konkurrierenden Hypothesen wird in diesem für die Mathematische Statistik grundlegenden Konzept letztlich als Entscheidung zwischen „two courses of action" gedacht. Eine Fehlentscheidung zweiter Art ist desto unwahrscheinlicher, je weiter die wahren Verhältnisse von der Nullhypothese entfernt sind und je größer die untersuchte Zufallsstichprobe ist. Bei sehr großen Stichproben - wie etwa in der zitierten Untersuchung von Gunter Sachs - werden auch sehr kleine Abweichungen von der Nullhypothese mit hoher Wahrscheinlichkeit signifikant, münden also in eine Entscheidung gegen die Nullhypothese.

Kasten 3: Das Problem der **Alpha-Fehler-Inflation** oder **Alpha-Fehler-Kumulierung** tritt dann auf, wenn eine „globale" Nullhypothese in vielen Einzeltests überprüft wird und signifikante Ergebnisse in einem oder mehreren Einzeltests als Entscheidung gegen die globale Nullhypothese interpretiert werden. Ist etwa eine globale Nullhypothese die Konjunktion von *m* einzelnen Nullhypothesen und wird jede dieser einzelnen Nullhypothesen jeweils zum Niveau α getestet, so beträgt die Wahrscheinlichkeit, daß mindestens einer dieser einzelnen Signifikanztests signifikant wird, obwohl die globale Nullhypothese - und damit auch jede einzelne Nullhypothese - gilt, genau

$$1-(1-\alpha)^m,$$

sofern die einzelnen Signifikanztests stochastisch unabhängig voneinander sind. Die globale Nullhypothese von Gunter Sachs etwa - Unabhängigkeit der Eheschließungen vom Sternzeichen der Ehepartner - ist die logische Konjunktion von 144 einzelnen Nullhypothesen, die jeweils die Wahrscheinlichkeit spezifizieren, daß ein Ehepaar in die betreffende Zelle der 12x12-Kontingenztafel gehört. Die Wahrscheinlichkeit, bei Geltung der globalen Nullhypothese wenigstens in einem der 144 Tests dieser Einzelhypothesen ein signifikantes Ergebnis auf dem 5 %-Niveau zu erhalten, beträgt - Unabhängigkeit der einzelnen Tests vorausgesetzt - demnach

$$1-(1-0{,}05)^{144} = 1-0{,}0006 = 0{,}9994\,.$$

Selbst wenn man die stochastische Abhängigkeit der von Gunter Sachs durchgeführten 144 Tests vergröbernd so modelliert, daß jedes signifikante Ergebnis „automatisch" ein weiteres signifikantes Ergebnis (in entgegengesetzter Richtung) nach sich zieht, selbst wenn man also unterstellt, daß eigentlich nur 72 unabhängige Tests durchgeführt wurden, ergibt sich als Wahrscheinlichkeit zumindest eines signifikanten Ergebnisses trotz Geltung der globalen Nullhypothese immer noch

$$1-(1-0{,}05)^{72} = 1-0{,}0249 = 0{,}9751\,.$$

Die überaus geringen Gegenwahrscheinlichkeiten dazu - 0,0006 und 0,0249 - wären übrigens die Wahrscheinlichkeiten dafür, daß die von Gunter Sachs durchgeführte einmalige Simulation - wie von ihm behauptet - kein einziges signifikantes Ergebnis erbracht hätte. Zweifel sind hier wohl erlaubt.

Kasten 4: Als Gegenmittel gegen die unerwünschte Alpha-Fehler-Kumulierung bei multiplem Testen sind Strategien der Alpha-Fehler-Korrektur oder Alpha-Fehler-Adjustierung vorgeschlagen worden. Grundidee dabei ist, das Signifikanzniveau α' der Einzeltests so zu verschärfen, daß die Wahrscheinlichkeit, die globale Nullhypothese im Falle ihrer Geltung irrtümlich zurückzuweisen, weil mindestens ein Einzeltest signifikant wird, durch das globale Signifikanzniveau α limitiert bleibt. Soll etwa eine globale Nullhypothese zum Niveau α dadurch getestet werden, daß m unabhängige Einzeltests jeweils zum Niveau α' durchgeführt werden, und soll dabei die globale Nullhypothese zurückgewiesen werden genau dann, wenn wenigstens ein Einzeltest ein signifikantes Ergebnis erbringt, so sollte für die Einzeltests das Niveau $\alpha'=1-(1-\alpha)1/m$ gewählt werden. Denn unter Geltung der globalen und damit auch aller einzelnen Nullhypothesen wird bei diesem Niveau α' mindestens ein Einzeltest signifikant mit der Wahrscheinlichkeit $1-(1-\alpha')m=1-(1-(1-\alpha)1/m))m=\alpha$, also genau der für den Globaltest zugestandenen Fehlerwahrscheinlichkeit erster Art. Da $1-(1-\alpha)1/m$ für wachsendes m sich recht schnell dem einfacheren Term α/m annähert, benutzt man in der Praxis häufig letzteren Wert, bekannt unter der Bezeichnung Bonferoni-Korrektur. Gunter Sachs hätte mit dieser Korrektur jeden einzelnen Binomialtest auf dem Niveau 0,05/144=0,000347 durchführen müssen; hier hätte dann jeweils die beobachtete von der erwarteten Häufigkeit der entsprechenden Sternzeichenkombination um mindestens 174 abweichen müssen, um signifikant zu werden. Davon ist nirgends die Rede.

Sind - wie häufig in der Praxis - die Einzeltests nicht unabhängig voneinander, so führt eine solche Korrektur tendenziell zu einer Reduktion der globalen Fehlerwahrscheinlichkeit erster Art unter die vorgewählte Grenze α; der Globaltest wird also konservativer als geplant.

Kasten 5: Durchgeführt wurde von Gunter Sachs vermutlich für alle 144 möglichen Sternzeichenkombinationen jeweils ein - mittels Normalverteilung approximierter - zweiseitiger Binomialtest jeweils der Nullhypothese H0: p=p0 über die Wahrscheinlichkeit p, daß ein untersuchtes Ehepaar die betreffende Sternzeichenkombination aufweist. p0 ergab sich dabei jeweils unter der Unabhängigkeitsannahme aus der Bevölkerungsstatistik. Bei - tatsächlich nicht vorliegender, hier aber der Einfachheit halber unterstellter - Gleichverteilung der Sternzeichen in der Population ergäbe sich für alle Sternzeichenkombinationen jeweils

$$p_0 = \frac{1}{12}\cdot\frac{1}{12} = \frac{1}{144} = 0{,}006944444$$

Bei diesem p0 und einem Signifikanzniveau von 5% ergibt sich bei dem Gunter Sachs verfügbaren gigantischen Stichprobenumfang von n= 358.763 Paaren im Binomialtest eine Entscheidung gegen die Nullhypothese genau dann, wenn höchstens 2.393 oder mindestens 2.589 Paare entsprechender Sternzeichenkombination beobachtet werden, falls also die Abweichung von den unter der Nullhypothese zu erwartenden Zahl von

$$358.763\cdot\frac{1}{144} = 2.491{,}4 \approx 2.491$$

Paaren entsprechender Sternzeichenkombination mindestens 98 Paare beträgt. Dieser Binomialtest hat aufgrund der großen Stichprobe große Teststärke: Liegt etwa die tatsächliche Wahrscheinlichkeit p der Sternzeichenkombination nur um 5 % unter dem nullhypothetischen Wert p0, gilt also tatsächlich

$$p = 0{,}95 \cdot p_0 = 0{,}95 \cdot 0{,}006944444 = 0{,}006597222 \approx \frac{1}{152},$$

so werden in der Stichprobe mit einer Wahrscheinlichkeit von rund 70 % höchstens 2.393 Paare mit der entsprechenden Sternzeichenkombination auftreten, wird also mit rund 70 %iger Wahrscheinlichkeit der Binomialtest ein signifikantes Ergebnis zeigen. Liegt das wahre p tatsächlich um 10 % unter dem nullhypothetischen p0, gilt also tatsächlich

$$p = 0{,}90 \cdot p_0 = 0{,}90 \cdot 0{,}006944444 = 0{,}006250000 = \frac{1}{160},$$

so beträgt die Wahrscheinlichkeit eines signifikanten Ergebnisses im Binomialtest mehr als 99,9 %. Gleiches gilt vice versa aufgrund von Symmetrie auch für Abweichungen nach oben. Fazit: In jedem der 144 von Gunter Sachs durchgeführten Signifikanztests ist ein signifikantes Ergebnis wahrscheinlich auch dann, wenn die Abweichung von der Nullhypothese unwesentlich und kaum von praktischer Relevanz ist.
Für diese Überlegungen benötigen die Schüler lediglich Fertigkeiten im Umgang mit der Binomialverteilung bei großem n; sie erleben hier noch einmal den Unterschied zwischen statistischer Signifikanz und praktischer Relevanz.

Kasten 6: Folgt aus den Angaben von Gunter Sachs, daß auch der eigentlich angemessene Chi-Quadrat-Test signifikant geworden wäre? Eine kleine Überschlagsrechnung: Berichtet wird von 16 signifikanten Einzelergebnissen in den 144 Tests. Nach den Feststellungen in Kasten 5 waren dafür Abweichungen von den - im Mittel - unter der Nullhypothese erwarteten rund 2.500 Paaren von mindestens rund 100 notwendig. Angenommen, neben diesen 16 signifikanten Abweichungen vom Ausmaß 100 - in quadrierter Form also 10.000 - waren die übrigen Ergebnisse im wesentlichen nullhypothesenkonform: Dann ergibt sich die Prüfgröße $\chi 2$ ungefähr als Summe aus 16 Brüchen mit dem Zähler 10.000 und dem Nenner 2.500, also als 16 mal 4 gleich 64. Dieser Wert erreicht nicht annähernd den kritischen Wert auf dem 5 %-Niveau, der bei etwa 170 liegt.

Kasten 7: Wie groß ist die nullhypothetische Überschreitungswahrscheinlichkeit für Gunter Sachs' Empirie von 16 signifikanten Ergebnissen in den 144 Einzeltests? Wären die Einzeltests voneinander unabhängig, so wäre die Anzahl der auf dem 5 %-Niveau signifikanten Ergebnisse unter der globalen Nullhypothese binomial-, genauer B(144; 0,05)-verteilt und somit die nullhypothetische Wahrscheinlichkeit für mindestens 16 signifikante Ergebnisse kleiner als 1 %, wie ein Blick in die Binomialtabelle (oder vielmehr die approximierende Normalverteilungstabelle) lehrt. Unabhängig wären diese Tests aber nur dann, wenn für jeden Test erneut eine Zufallsstichprobe von 358.763 Paaren gezogen worden wäre. Tatsächlich wurden aber alle 144 Tests

mit derselben Stichprobe durchgeführt. Da somit die Summe der unter der globalen Nullhypothese erwarteten ebenso wie der beobachteten Häufigkeiten über alle Sternzeichenkombinationen den Stichprobenumfang 358.763 ergeben muß, muß jede Abweichung von der erwarteten Häufigkeit in einer Zelle der 12x12-Kontingenztafel an anderer Stelle wieder ausgeglichen werden. Eine signifikante Abweichung in einer Zelle erhöht also die Wahrscheinlichkeit für das Auftreten in Gegenrichtung signifikanter Abweichungen in anderen Zellen. Die 144 Tests sind also tatsächlich auf eine kaum exakt modellierbare Weise stochastisch abhängig
Wenn man die stochastische Abhängigkeit der von Gunter Sachs durchgeführten 144 Tests so modelliert, daß jedes signifikante Ergebnis „automatisch" ein weiteres signifikantes Ergebnis (in entgegengesetzter Richtung) nach sich zieht, wenn man also unterstellt, daß eigentlich nur 72 unabhängige Tests durchgeführt wurden, ergibt sich als Wahrscheinlichkeit für mindestens 16 halbe gleich 8 signifikante Ergebnisse in diesen 72 Tests trotz Geltung der globalen Nullhypothese 3 %.

Autor:
Raphael Diepgen
Ruhr-Universität
D-44780 Bochum