Leitidee L5 – Daten und Zufall

Stefan Krauss, Georg Bruckmaier, Christine Schmeisser
Fakultät für Mathematik, Didaktik der Mathematik
Universität Regensburg

Fortbildungsveranstaltung für MathematiklehrerInnen
zum Teilgebiet Stochastik
Universität Regensburg
17.3.2011

Einführende Bemerkungen

14.30 Uhr:	Begrüßung und Vorstellung des Teams
14.45 Uhr:	„Prozente, Anteile, Wahrscheinlichkeiten“ (Vortrag)
	(Folie 3 – 55)
16.00 Uhr:	Pause (Imbiss und Informationsstände zur Stochastik,
	zur COACTIV-Studie und des Buchner-Verlags)
16.30 Uhr:	Angebot 1: „Wie lügt man mit Statistik?“ (Workshop)
	Angebot 2: „Was bedeutet eigentlich signifikant?“ (Vortrag)
	(Folie 56 – 105)
17.45 Uhr:	Ausklang und Möglichkeit zur weiteren Diskussion

Besuchen Sie uns auch unter: http://www-tz.uni-regensburg.de/mathe/

„Prozente, Anteile, Wahrscheinlichkeiten“ – Einige Vorschläge zur Umsetzung der Leitidee L5 „Daten und Zufall“ in der Sekundarstufe I

Stefan Krauss

Wozu Wahrscheinlichkeitsrechnung und Statistik in der Schule?

→ Wir sind heutzutage einem „Trommelfeuer“ aus Daten, Statistiken, Kurven und Trends ausgesetzt.

→ In einer durchschnittlichen Zeitung finden sich mehr Statistiken als Goethe und Schiller in ihrem ganzen Leben gesehen haben.

→ Das Wort „Prozent“ mittlerweile eines der häufigsten Substantive in deutschen Tageszeitungen!

Walter Krämer (1998): So lügt man mit Statistik. Campus Verlag

Aus der Zeitschrift Cicero: Magazin für politische Kultur vom 1.1.2007

„ […] Unsere Gesellschaft muss stärker lernen, Risiken zu bewerten, ganz generell gesprochen. Das Leben mit der Chance und dem Risiko ist ein wichtiges gesellschaftliches Problem. Ich finde es in einer komplexer werdenden Welt auch wichtig, Kinder bereits frühzeitig an solche Abwägungen heranzuführen, die sie später immer wieder vornehmen müssen […]. Im Kindergarten und in der Schule können Kinder spielerisch lernen, was Wahrscheinlichkeit und Risiko bedeuten. Nehmen Sie die morgendliche Diskussion nach Hören des Wetterberichts, ob man nun die Regenjacke mitnimmt oder nicht. Denn die Regenjacke zu schleppen, wenn die Sonne scheint, ist das Unangenehmste, was einem nach der Schule passieren kann. Aber bei 30 Prozent Regenwahrscheinlichkeit keinen Schutz zu haben und nass zu werden, wäre auch ungemütlich. Darüber zu diskutieren, dass man für Schutz höheren Aufwand betreiben und abwägen muss, ob dieser sich lohnt, halte ich für wichtig.“

Auszug aus einem Interview mit Bundeskanzlerin Dr. Angela Merkel zum Thema: „Was bedeutet Ihnen die Natur?“

Wozu also Wahrscheinlichkeitsrechnung und Statistik in der Schule?

→ Zurechtfinden in der Informationsgesellschaft („Daten“)

→ Abschätzen und Bewerten von Chancen und Risiken („Zufall“)

Ziel des folgenden Vortrags (14.45 Uhr):

Wie kann Kompetenz vor allem in diesen beiden Aspekten in der Schule unterstützt werden?

Ziel der beiden nachfolgenden Veranstaltungen (16.30 Uhr):

Spezifizierung der beiden Aspekte auf

- den Signifikanzbegriff (Verwirrung, Missbrauch und weitere Tücken)
- Sensibilisierung gegenüber „statistischen Manipulationen“

Die folgenden beiden Folien sind entnommen aus:

Eichler, A. (2011). Daten und Zufall - eine realitätsorientierte (Leit-)Idee für beide Sekundarstufen. 16. Dresdener Kolloquium zur Mathematik und ihrer Didaktik. Dresden: TU Dresden.

Sehr empfehlenswert zur Umsetzung der Leitidee L5 ist weiterhin:

http://www.leitideedatenundzufall.de/

Zu diesem Buch wünschen die Autoren Rückmeldung unter:

eichlervogel@leitideedatenundzufall.de

Fragen an die reale Realität?

Produktion:

„Eine Firma für elektronische Geräte stellt Transistoren her; sie weiß, dass im Durchschnitt 2% davon defekt sind. Wie groß ist die Wahrscheinlichkeit dafür, dass von den 20 Transistoren genau 3 defekt sind?“

Produktion:

Eine Tankstelle hat gute und schlechte Kunden; sie weiß, dass im Durchschnitt 2% der Kunden, ohne zu bezahlen, davon fahren. Wie groß ist die Wahrscheinlichkeit dafür, dass von den 20 Kunden eines Tagen genau 3 nicht bezahlen?“

Produktion:

Ein Schnellimbiss bezieht von einem Hersteller Hamburger; er weiß, dass im Durchschnitt 2% der Hamburger zwei statt einer Gurke als Belag haben. Wie groß ist die Wahrscheinlichkeit dafür, dass von den 20 verkauften Hamburgern in einer Stunde genau 3 mit einer Gurke belegt waren?“

Fragen an die reale Realität?

Produktion:

Bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla im Durchschnitt 2% bla bla bla bla bla bla bla bla bla bla bla bla. Bla bla bla bla bla bla. Wie groß ist die Wahrscheinlichkeit dafür, dass von den 20 bla bla bla bla bla bla genau 3 bla bla bla bla bla bla?“

In Schulbüchern leider oft:

Text, Sachsituation und Daten sind irrelevant („eingekleidete Aufgaben“), nur die Zahlen sind relevant (Modell: Binomialverteilung)

Die Frage im heutigen Vortrag lautet:

Welche Wahrscheinlichkeiten, Risiken und Chancen werden in unserer Informationsgesellschaft ***tatsächlich*** kommuniziert?

Und wie (z.B. in Fernsehen, Zeitungen und Radio)?

Regenwahrscheinlichkeit

Eine Umfrage unter US-amerikanischen Radiohörern ergab folgende verschiedene Interpretationen für die Meldung

„30% Regenwahrscheinlichkeit“

- Es wird mit 30% Wahrscheinlichkeit im *gesamten* Sendegebiet regnen
- Es wird mit 30% Wahrscheinlichkeit *irgendwo* im Sendegebiet regnen
- Es wird in 30% der *Fläche* des Sendegebietes regnen, wann weiß nur nicht wo
- Es wird in 30% der *Zeit* regnen, man weiß nur nicht wann

Am seltensten (die intendierte Interpretation)

- In 30% der *Tage mit vergleichbaren Wetterbedingungen* regnet es

→ Menschen haben bereits Verständnisschwierigkeiten bei „scheinbar einfachen“ Aussagen!

Gut, aber das mit der Regenwahrscheinlichkeit ist schwierig!

1) Der Moderator hat nicht richtig kommuniziert, was darunter zu verstehen ist!
2) Außerdem: In der Aussage kommt der Begriff „Wahrscheinlichkeit“ vor, das macht alles schwierig!

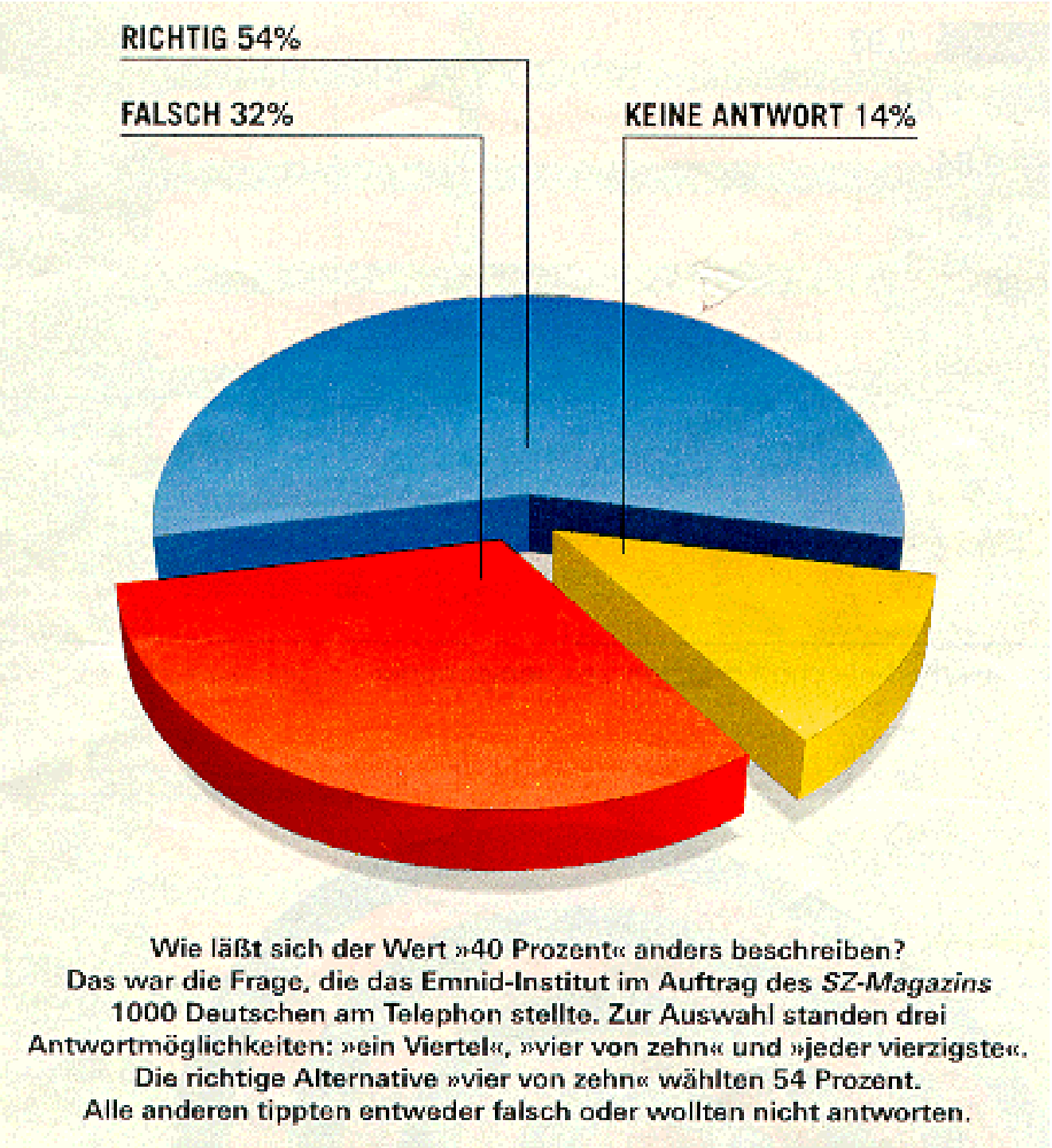
→ Jedem Menschen ist klar, was „30%“ bedeutet!

→ Ja?

- *Emnidumfrage*:

„Was bedeutet 40%?“

- *ein Viertel?*
- *vier von zehn?*
- *jeder vierzigste?*



Süddeutsche Zeitung
Magazin (1.1.2006)

Darstellungen von „Unsicherheit“ bzw. „Wahrscheinlichkeit“

Nummerische Darstellungen	*Beispiel*
Prozente	25%
Dezimalbrüche	0,25
Gewöhnliche Brüche	¼
Absolute Häufigkeiten	1 von 4
„Jeder wievielte“	jeder vierte
Chancenverhältnisse	1 zu 3

Bereits das Verständnis (und noch mehr) die Umrechnungen dieser verschiedenen Darstellungen kann ein großes Problem darstellen!

Jeder vierte will unsterblich sein

HAMBURG (kna) – Einer Umfrage zufolge wollen 44 Prozent der Deutschen nicht älter als 80 Jahre alt werden. Höchstens 100 Jahre alt wollen 18 Prozent werden, wie eine gestern veröffentlichte Befragung für die Zeitung „Die Woche“ ergeben hat. Vier Prozent hätten angegeben, sie wollten unsterblich werden.

Die äußerst beliebte
„jeder x-te entspricht x%“
-Täuschung

Wie viele Deutsche wollen nun unsterblich sein?

Jeder vierte (25%) oder 4%?

Mainzer Allgemeine Zeitung, 7. 8. 1997

Aus der *Norderneyer Badezeitung*:
„Fuhr vor einigen Jahren noch jeder zehnte Autofahrer zu schnell, so ist es mittlerweile heute ‚nur noch' jeder fünfte. Doch auch fünf Prozent sind zu viele, und so wird weiterhin kontrolliert, und die Schnellfahrer haben zu zahlen."

„Doppelfehler":

„Jeder fünfte"
besser als
„jeder zehnte"?

Wie viele Deutsche sind zu schnell?

5% oder 20%?

Wochenpost (1995) (HWH) (1993)

Was ist hier falsch?

Oder:

Warum berichten mehr Leute von vollen als von leeren Zügen?

Noch schlimmer: „Einfache“ Rechnungen (z.B. „Verdopplung“ …)

Anstieg der Rauschgiftopfer gegenüber dem Vorjahr alarmierend

Anzahl der Drogentoten hat sich 1990 fast verdoppelt

WIESBADEN (dpa) Die Zahl der Rauschgifttoten in der Bundesrepublik ist 1990 alarmierend gestiegen und gegenüber dem vergangenen Jahr um fast 50 Prozent angewachsen.

Wie das Bundeskriminalamt (BKA) in Wiesbaden mitteilte, wurden bis Donnerstag 1365 Menschen Opfer ihrer Drogensucht.

Bis zum 27. Dezember 1989 waren der Wiesbadener Behörde 950 Rauschgifttote bekanntgeworden. Die neueste Zahl der Drogenopfer schließt erstmals die fünf neuen Bundesländer ein.

2.1990

Können Schüler selbst feststellen, welche Angabe stimmt?

„Verdoppelt“ oder um „50% gestiegen“?

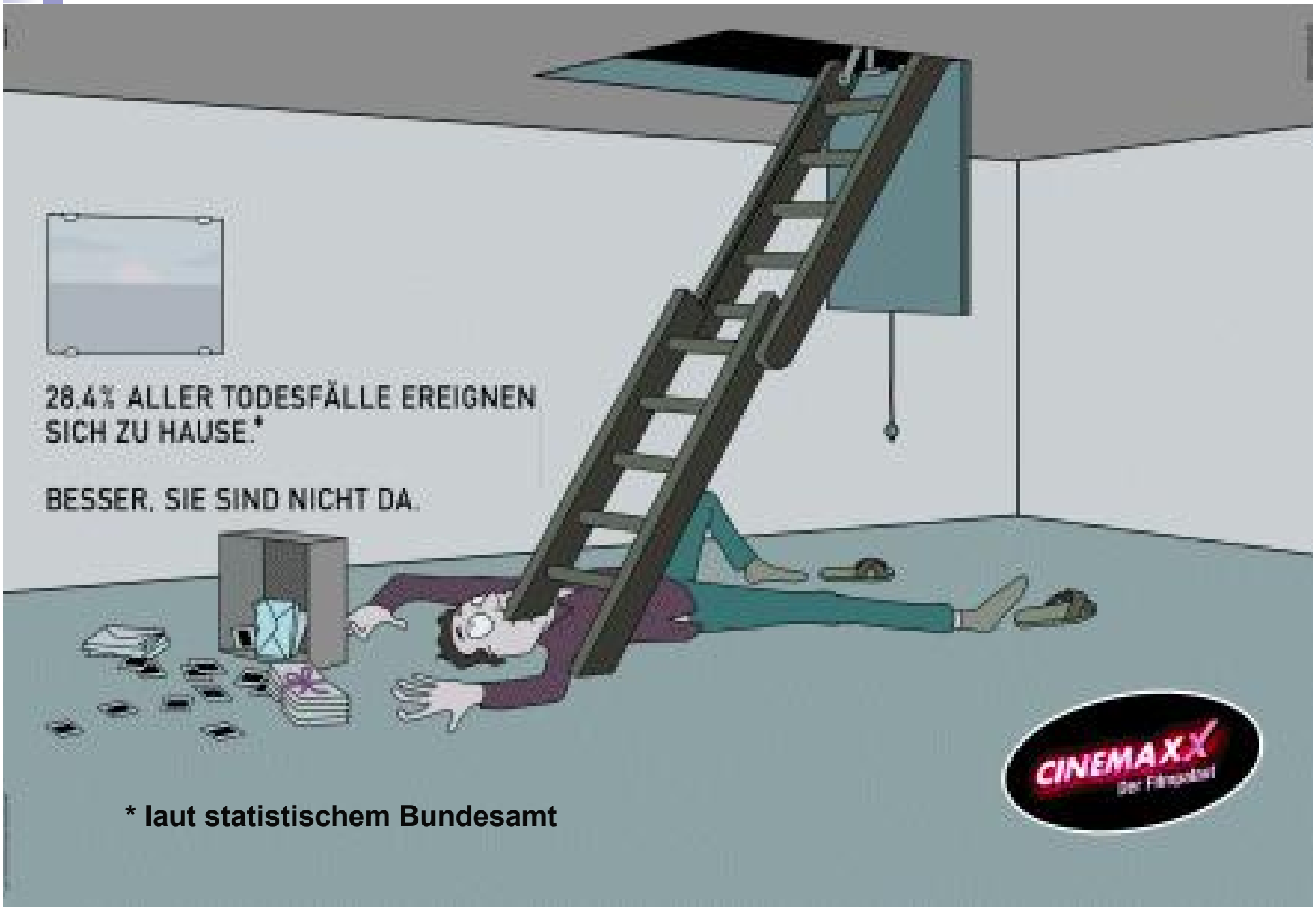
Energiesparen

280 Prozent Strom können leicht gespart werden – beim Kochen, wenn der Deckel nicht vergessen wird. Das ist nur ein Beispiel: So empfiehlt sich etwa bei Speisen mit Garzeiten von mehr als 20 Minuten ein Schnellkochtopf. Damit lassen sich bis zu 50 Prozent Energie und 75 Prozent Zeit einsparen. Grundsätzlich verbrauchen Töpfe mit gewölbtem Boden 50 Prozent mehr Energie als solche mit einem ebenen Boden.

Hannoversches Wochenblatt, zitiert nach *„Der Spiegel"*, Nr. 51/1995 (UL)

Doch selbst wenn wir ganz genau wissen, was „Wahrscheinlichkeit“ und „Prozent“ bedeuten, kann uns die *Interpretation* einer einfachen Wahrscheinlichkeits- bzw. Prozentangabe (und etwaige Schlussfolgerungen daraus) immer noch vor Schwierigkeiten stellen …

Schlussfolgerungen aus Prozenten?



Wo sollte man sich nun aufhalten?

28,4% ALLER TODESFÄLLE EREIGNEN SICH ZU HAUSE.

BESSER, SIE SIND NICHT DA.

Aber: Wenn sich 28,4% aller Todesfälle zuhause ereignen, ereignen sich 71,6% aller Todesfälle außerhaus!

Sollte man also besser zuhause bleiben?

Die Cinemaxx-Werbung ist deshalb wirkungsvoll, da die meisten von uns den ersten Prozentsatz auf ca. 10% schätzen würden. Verglichen dazu ist 28,4% „erschreckend hoch“!

Unsere Interpretationen von berichteten Prozenten (z.B. Risiken) richten sich also nicht immer nach einem 50%-Kriterium, sondern oftmals nach dem Vergleich mit einer subjektiv empfundenen eigenen Risikoeinschätzung.

Kahneman & Tversky

Wie gut sind Menschen bei (scheinbar) einfachen Schlussfolgerungen?

Das Wahrscheinlichkeitsdenken der Menschen (d.h., Einschätzen von Risiken; Urteilen unter Unsicherheit) wurde vor allem von den Kognitionspsychologen Amos Tversky und Daniel Kahneman *empirisch* untersucht.

Dabei konnte eine Fülle von Aufgaben gefunden werden, bei denen die menschliche Intuition *systematisch* und *dramatisch* von den normativ richtigen Antworten der Wahrscheinlichkeitsrechnung und Logik abwich.

Dieses Forschungsprogramm hatte seitdem (und hat immer noch) großen Einfluss auf Disziplinen wie Medizin, Ökonomie, oder Rechtswissenschaften.

Einige einfache Beispiele aus diesem Forschungsprogramm …

„Lindaaufgabe“

Linda ist 32, sie hat in Philosophie promoviert und ist ausgesprochen intelligent. Sie ist sozial sehr engagiert und war früher in der Anti-Atomkraft Bewegung aktiv. Was ist wahrscheinlicher?

a) Linda ist Bankangestellte

b) Linda ist Bankangestellte und in der feministischen Bewegung aktiv

a) Linda ist Bankangestellte

b) Linda ist Bankangestellte und in der feministischen Bewegung aktiv

Die meisten Menschen entscheiden sich für „b“, aber es gilt allgemein:

$P(A \text{ „und“ } B) = P(A \cap B)$

$P(A \cap B) = P(A) \cdot P(B \mid A)$, und da P immer zwischen 0 und 1:

$P(A \cap B) <$ Wahrscheinlichkeit eines jeden Einzelereignisses

- Fehler im Wesentlichen unabhängig von Ausbildung
- Versuchspersonen, die sich für „b“ entscheiden, berichten außerdem von sehr hoher Sicherheit ihrer Einschätzung

Ein etwas schwierigeres Beispiel

Mittlerweile ein kognitionspsychologischer „Klassiker“: Die Mammografieaufgabe

(Aus dem Buch von Kahneman, Slovic und Tversky, 1982)

Mit dem Ziel der Früherkennung von Brustkrebs werden Frauen angehalten, ab einem bestimmten Alter regelmäßig eine Mammografie durchführen zu lassen, selbst wenn keine Symptome vorliegen. Für symptomfreie Frauen im Alter zwischen 40 und 50 Jahren, die im Rahmen einer Reihenuntersuchung eine Mammografie durchführen lassen, liegen folgende Informationen vor:

- Die Wahrscheinlichkeit, dass eine dieser Frauen Brustkrebs hat, beträgt 1%.

- Wenn die Krankheit vorliegt, wird sie durch die Mammografie mit einer Wahrscheinlichkeit von 80% erkannt („positiver Mammografie-Befund“).

- Jedoch auch gesunde Frauen erhalten mit einer Wahrscheinlichkeit von 9,6% fälschlicherweise einen positiven Mammografie-Befund.

- Eine Frau dieser Altersgruppe erhält nun einen positiven Mammografie-Befund. Wie groß ist die Wahrscheinlichkeit, dass sie tatsächlich an Brustkrebs erkrankt ist?

- _____%

Ergebnisse empirischer Untersuchungen

Übliche Schätzungen liegen um die 80%

Auch hier liegt wieder – im Vergleich zur richtigen Lösung von 7,8% – eine *deutliche Abweichung* der menschlichen Intuition von der Wahrscheinlichkeitsrechnung vor

- Kahneman et al. (1982):
  Sogar 95 von 100 Ärzten schätzten unter den gegebenen Umständen die Wahrscheinlichkeit für das Vorliegen der Krankheit auf zwischen 70% und 80%

→ Schätzungen liegen also am anderen Ende des Wahrscheinlichkeitsspektrums!

Mit folgender Formalisierung der Informationen

B : Brustkrebs

$\overline{B}$: nicht Brustkrebs

M+ : positiver Mammografiebefund

liefert der Satz von Bayes

$$p(B \mid M+) = \frac{p(M+ \mid B) \cdot p(B)}{p(M+ \mid B) \cdot p(B) + p(M+ \mid \overline{B}) \cdot p(\overline{B})}$$

$$= \frac{0{,}8 \cdot 0{,}01}{0{,}8 \cdot 0{,}01 + 0{,}096 \cdot 0{,}99}$$

$$= 0{,}078$$

$$\approx 8\%$$

Kann man vielleicht eine andere Darstellung für die nummerische Information wählen?
In welcher Darstellung sind Menschen „am besten“?

Nummerische Darstellungen	*Beispiel*
Prozente	25%
Dezimalbrüche	0,25
Gewöhnliche Brüche	¼
Absolute Häufigkeiten	1 von 4
„Jeder wievielte“	jeder vierte
Chancenverhältnisse	1 : 3

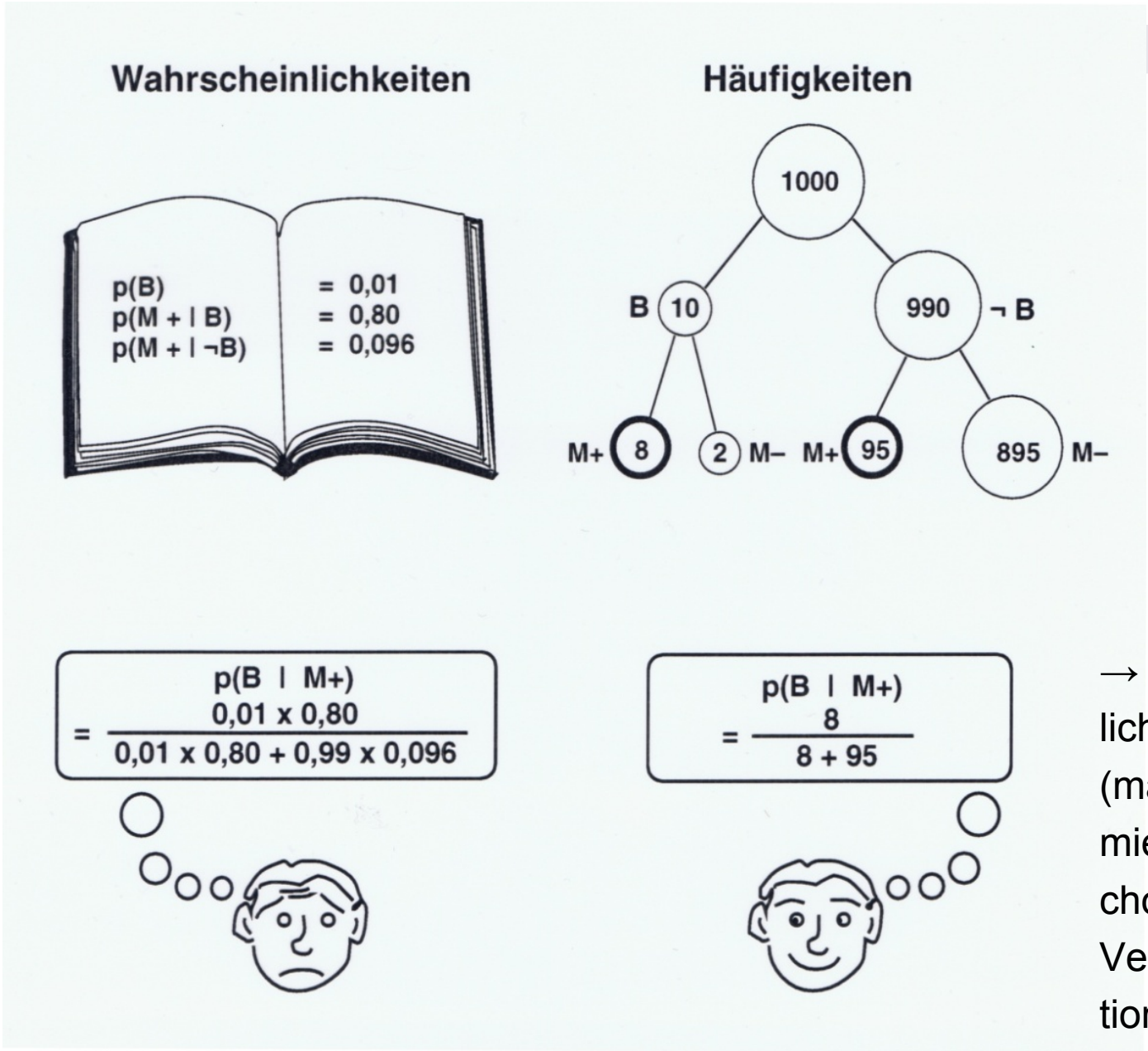
Kann man vielleicht eine andere Darstellung für die nummerische Information wählen?
In welcher Darstellung sind Menschen „am besten“?

Nummerische Darstellungen	*Beispiel*
Prozente	25%
Dezimalbrüche	0,25
Gewöhnliche Brüche	¼
Absolute Häufigkeiten	**1 von 4**
„Jeder wievielte“	jeder vierte
Chancenverhältnisse	1 : 3

Mammografieaufgabe

Gigerenzer und Hoffrage (1995) setzten dies experimentell um:

- 10 von 1000 dieser Frauen haben Brustkrebs **(statt 1%)**
- 8 von 10 Frauen, die Brustkrebs haben, erhalten einen positiven Mammographie-Befund **(statt 80%)**
- 95 der 990 Frauen, die keinen Brustkrebs haben, erhalten dennoch einen positiven Mammografie-Befund **(statt 9,6%)**
- Wieviele der Frauen, die einen positiven Mammografie-Befund erhalten haben, sind tatsächlich an Brustkrebs erkrankt?
- _____von_____ **(statt eine bedingte Wahrscheinlichkeit)**



Die meisten Versuchspersonen finden jetzt die richtige Antwort:

8 von 103 (≈ 7,8%)

→ Durch die Wahrscheinlichkeitstheorie erhält man (mathematisch) die Normierung und verliert (psychologisch) die natürliche Vernetzung der Information und das Referenzset

Gigerenzer & Hoffrage (1995): „Natürliche Häufigkeiten“

Vierfeldertafel

	Positiver Testbefund	Negativer Testbefund	Summe
Krebs	8	2	10
Kein Krebs	95	895	990
Summe	103	897	1000

Fragen z.B.:

- Wie viel Prozent der Frauen haben einen negativen Testbefund?
- Welcher Anteil der Frauen mit positivem Testbefund hat tatsächlich Krebs?
- Was ist die Wahrscheinlichkeit, dass eine Frau mit Krebs einen negativen Testbefund erhält? (→ sehr schwer!)

Mit „A“ bzw. „B“ kann z.B. bedeuten:

- „krank“ bzw. „gesund“
- „positiver Testbefund“ bzw. „negativer Testbefund“

gibt es für jede mögliche Frage „drei Erscheinungsformen“:

1) „Wie viele der A sind / haben B?“
   (absolute Häufigkeitsfrage)
2) „Welcher Anteil der A sind / haben B?“
   (relative Häufigkeitsfrage)
3) „Was ist die Wahrscheinlichkeit für A unter der Bedingung B?“ (Frage nach bedingten Wahrscheinlichkeiten)

Sehr viele Aussagen in Medien sind von der Art 1 bzw. 2 !!!

Wie viele „Anteile“ gibt es überhaupt (ohne Schnitt bzw. Vereinigung)?

Die Wahrscheinlichkeitsschreibweise hilft zum Abzählen:

„A priori“ Wahrscheinlichkeiten:

p(A), p(nicht A), p(B), p(nicht B) (Vier „Anteile von allen“)

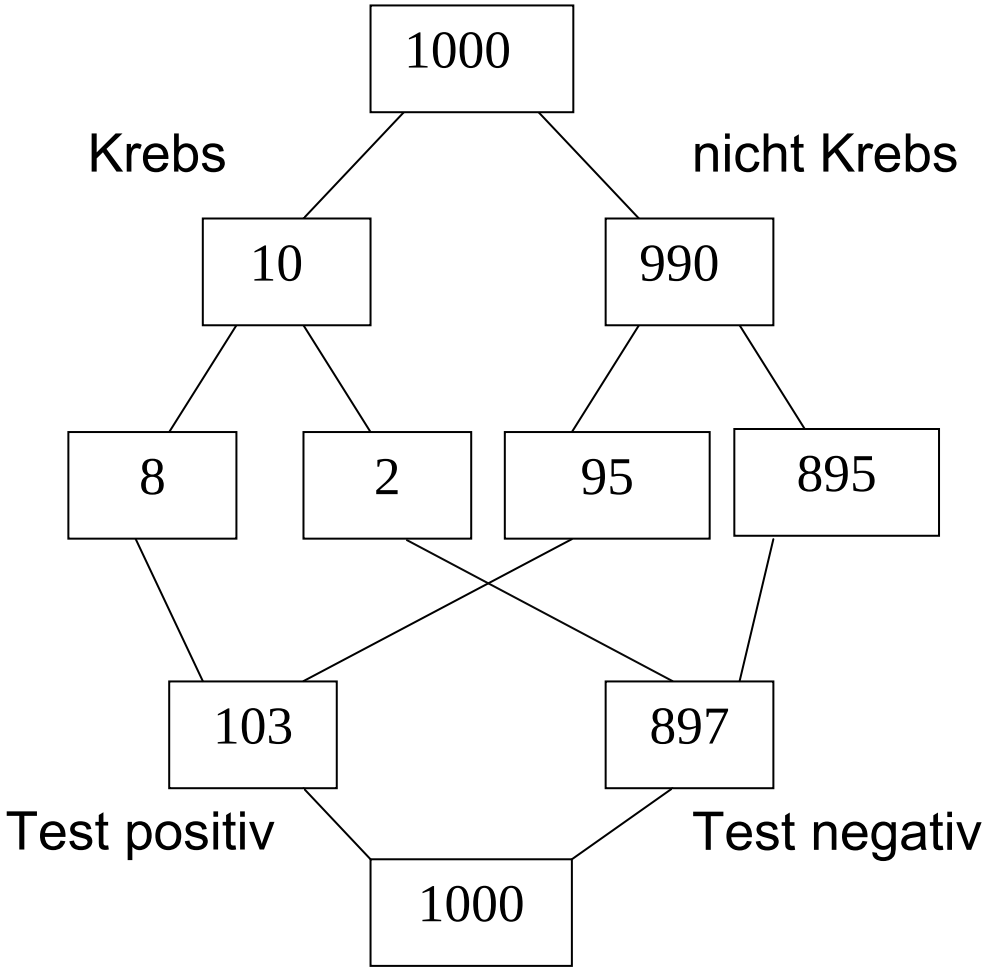
Acht „Bedingte“ Wahrscheinlichkeiten:

p(A / B), p(A / nicht B), p(nicht A / B), p(nicht A / nicht B)

p(B / A), p(B / nicht A), p(nicht B / A), p(nicht B / nicht A)

Formulieren Sie alle 12 möglichen Häufigkeiten / Anteile / bedingte Wahrscheinlichkeiten auf alle drei Arten!

Didaktisch empfehlenswert: Doppelbaum mit absoluten Häufigkeiten



- Hierarchische Struktur der Daten wird sichtbar
- „Leserichtung“ deutlich
- „Referenzset“ deutlich (“auf was bezieht sich eine Aussage?“)
- Relative Häufigkeiten lassen sich bei Bedarf (wie üblich) an den Ästen ergänzen
- *Alle* relevanten Fragen (z.B. nach allen Anteilen) lassen sich damit beantworten (auch ohne bedingte Wahrscheinlichkeiten!!!)

„Die Regenwahrscheinlichkeit ist 30%“

→ Von was? Auf welches Referenzset bezieht sich die Aussage?
Auf eine bestimmte Fläche? Auf eine bestimmte Zeit?

Dieses Problem tritt bei allen *normierten* Darstellungsarten auf:

„Regenwahrscheinlichkeit p = 0,3“
„Regenwahrscheinlichkeit 3/10“
„Regenwahrscheinlichkeit 30%“

Von was?

Bei absoluten Häufigkeiten ist die Bezugsgröße (Referenzset) automatisch enthalten und die Operationalisierung wird gleich mitgeliefert:

- In 30 von 100 Tagen …
- In 30 von 100 Hektar …
- In 30 von 100 Minuten …

Formulierung „in 30 von 100 regnet es“ ist unmöglich!

Und noch einmal …

Die „Lindaaufgabe“

Linda ist 32, sie hat in Philosophie promoviert und ist ausgesprochen intelligent. Sie ist sozial sehr engagiert und war früher in der Anti-Atomkraft Bewegung aktiv. Was ist wahrscheinlicher?

a) Linda ist Bankangestellte

b) Linda ist Bankangestellte und in der feministischen Bewegung aktiv

→ In Aufgabe ist gar keine Wahrscheinlichkeit gegeben?

→ Wie könnte „Häufigkeitsversion“ aussehen?

„Lindaaufgabe“ (Häufigkeiten)

Stellen Sie sich 200 Frauen vor, auf die Lindas Beschreibung passt:

Wie viele davon sind:

a) Bankangestellte

b) Bankangestellte und in der feministischen Bewegung aktiv

Über die Hälfte der Versuchspersonen sagen jetzt:

Mehr Personen in Gruppe a! (Gigerenzer; Fiedler, 1988)

O.J. Simpson und Häufigkeiten

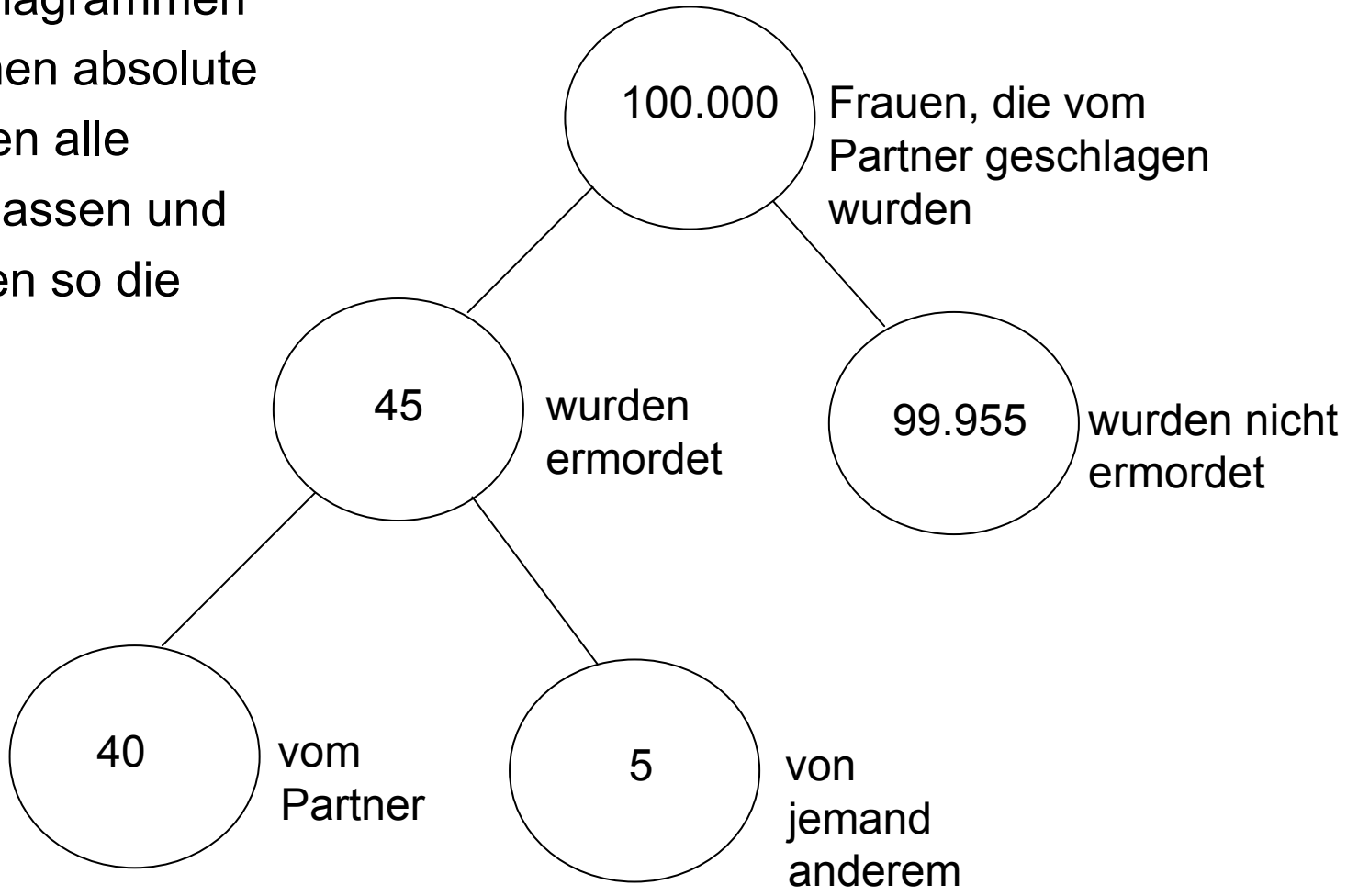
Alan Dershowitz, der Verteidiger von O.J. Simpson, argumentierte vor Gericht, dass die Tatsache, dass sein Mandant Nicole Brown Simpson geschlagen hat, kein aussagekräftiger Hinweis darauf sei, dass sein Mandant auch der Mörder sein muss.

Vielmehr belege eine Statistik, dass von 100.000 Frauen, die von ihrem Partner geschlagen wurden, „nur“ 40 auch später von ihrem Partner ermordet wurden. Deshalb, so Dershowitz, sei die Wahrscheinlichkeit, dass sein Mandant der Mörder von Nicole Brown Simpson sei, lediglich 1 / 2.500 (= 40 / 100.000)

Die Information war bereits in Häufigkeiten gegeben (das hatte der Richter Lance Ito verfügt!). Doch das ist die falsche Referenzklasse, denn Nicole Brown ist bereits tot. Die selbe Statistik sagt, dass von 45 *getöteten* Frauen, die vom Partner geschlagen wurden, 40 vom Partner ermordet wurden.

Referenzsets

Vor allem in Verbindung mit Baumdiagrammen verdeutlichen absolute Häufigkeiten alle Referenzklassen und unterstützen so die Kognition!



„Take-home-message“

Didaktischer Tipp bei Schwierigkeiten mit Wahrscheinlichkeiten:

Versuchen Sie es mit Häufigkeiten und / oder Häufigkeitsbäumen !!!

Damit wird klar, was Wahrscheinlichkeits- bzw. Prozentangaben *für eine konkrete Stichprobe bedeuten*.

Dazu beginnt man mit einer imaginären Stichprobe von 1000 oder 10000 Personen und teilt diese Gruppe sukzessive gemäß den Prozentangaben der Aufgabe auf. Endet man unten nicht mit ganzen Zahlen, lässt sich das durch entsprechende Vergrößerung der Stichprobe leicht erreichen.

Nach einer gewissen Übungszeit gelingt dies auch Schülern leicht und sie sehen die *Bedeutung von Prozentangaben*!

Und die berühmteste „Stochastik-Kopfnuss“?

Monty Hall

Moderator der US-Fernsehshow
“Let’s Make a Deal”

Ein Leserbrief an das Parade Magazine (1991)

Stellen Sie sich vor, Sie nehmen an einer Spielshow teil, bei der Sie eine von drei verschlossenen Türen auswählen sollen. Hinter einer Tür wartet der Preis, ein Auto, hinter den anderen beiden stehen Ziegen. Sie zeigen auf eine Tür, sagen wir Nummer 1. Sie bleibt vorerst geschlossen. Der Moderator weiß, hinter welcher Tür sich das Auto befindet. Mit den Worten “Ich zeige Ihnen mal was” öffnet er eine andere Tür, zum Beispiel Nummer 3, und eine Ziege schaut ins Publikum.

Er fragt: “Bleiben Sie bei Tür Nummer 1, oder wechseln Sie zu Tür Nummer 2?” Wie sollten Sie sich als Kandidat entscheiden?

Marylin vos Savant

- Autorin der Kolumne “Ask Marylin“ im *Parade Magazine*

- Leser durften alles fragen, was sie wollten und Marilyn bemühte sich um eine Antwort

- Marilyn vos Savant war zu dieser Zeit im Guinness Buch der Rekorde als der „intelligenteste Mensch“ aufgeführt (IQ = 228)

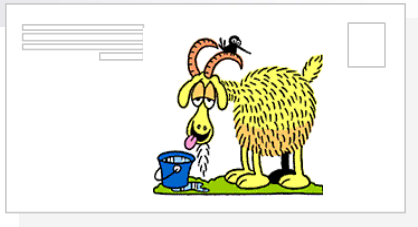
Sie antwortet:
„Ja, Sie sollten wechseln. Tür 1 hat eine 1/3 Chance auf den Gewinn, aber Tür 2 hat eine 2/3 Chance”

Leserbriefe

Das ist ja wohl ein Riesenschnitzer! Da Sie offensichtlich das Grundprinzip nicht sehen, erkläre ich das mal: Nachdem der Moderator eine Ziegentür geöffnet hat, ist die Chance auf das Auto 1 zu 2. Ob man jetzt die Tür wechselt oder nicht, die Chance auf das Auto bleibt gleich. Es gibt genug mathematisches Analphabetentum in diesem Land, wir brauchen nicht den höchsten IQ der Welt um noch mehr davon zu verbreiten! Schande!

Scott Smith, Ph.D.
University of Florida

Leserbriefe



Vielleicht lösen Frauen Mathematikprobleme ja anders als Männer?

Don Edwards,
Sunriver, Oregon

Sie sind die Ziege!

Glenn Calkins
Western State College

und Marilyn …

“Uff! Wenn diese Kontroverse andauert, passt bald nicht mal mehr der Postbote in’s Postamt. Ich erhalte Tausende von Briefen und nahezu alle insistieren, dass ich falsch liege, darunter leitende Direktoren und Statistiker aus der angewandten Forschung […].

Aber mathematische Wahrheiten werden nicht durch Abstimmungen entschieden.”

Marilyn

Insgesamt erhielt Marilyn über 10.000 Briefe.

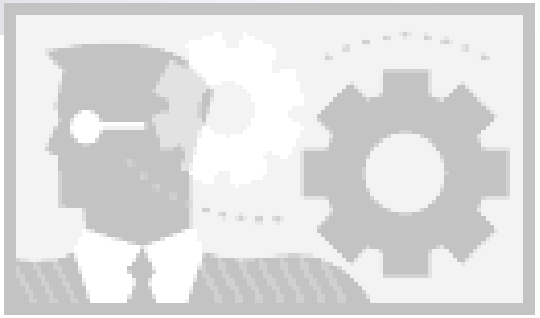
Manche Leserbriefschreiber waren sich so sicher, dass sie sogar Wetten über $20,000 anboten.

Am 21. Juli 1991 schaffte es das Ziegenproblem (“Monty Hall Dilemma”) sogar auf die Titelseite der New York Times.

Statt auf die Wetten einzugehen, beschloss Marilyn vos Savant die explodierende Debatte in einem Buch zu veröffentlichen:

The Power of Logical Thinking (1997)

„Die Königin der kognitiven Täuschungen“

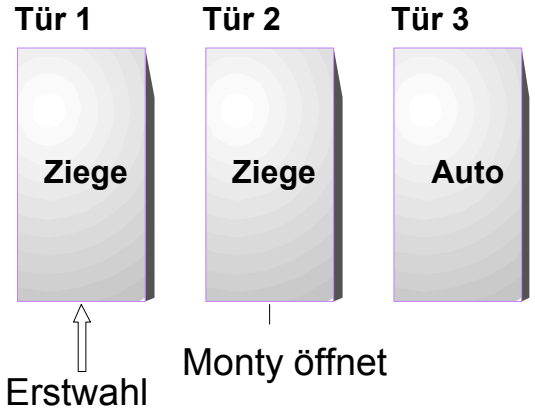


Vielleicht mit Häufigkeiten?

Das Phänomen ist vor allem deshalb so interessant, da es beliebig reproduzierbar und scheinbar auch immun gegen höhere Bildung ist […] sogar nobelpreisgekrönte Physiker geben systematisch die falsche Antwort […] und insistieren auf ihrem Irrtum, und beschimpfen sogar öffentlich diejenigen, die die richtige Lösung vertreten (Massimo Piattelli-Palmarini)

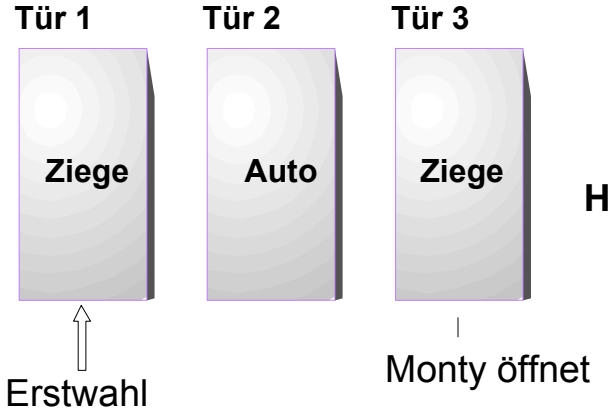
Das Ziegenproblem ist der endgültige Beweis dafür, dass unsere Gehirne nicht richtig verdrahtet sind, um Wahrscheinlichkeitsaufgaben zu lösen (Persi Diaconis)

Konstellation 1:



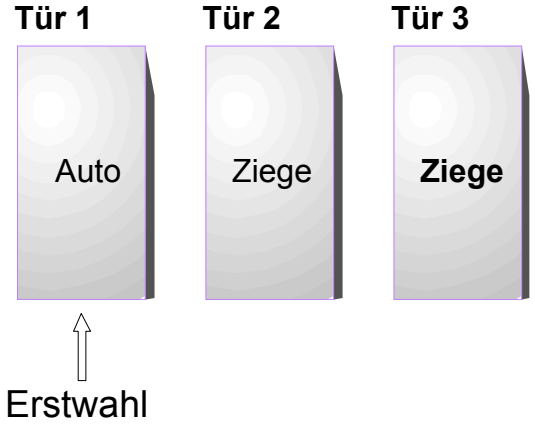
Hier gewinnt der Kandidat durch Wechseln

Konstellation 2:



Hier gewinnt der Kandidat durch Wechseln

Konstellation 3:



Hier gewinnt der Kandidat durch Bleiben, egal was Monty Hall macht

Stelle die Zwischenfrage (Häufigkeitsfrage!):

„In wie vielen der drei möglichen Auto-Ziege-Konstellationen würde der Kandidat durch Bleiben gewinnen und in wie vielen würde er durch Wechseln gewinnen?“

Dann erst Frage:

„Was sollte er also tun?“

→ 60% der Versuchspersonen wechseln, sogar 7. Klässler können die richtige Antwort geben und begründen!

Literatur zum Ziegenproblem:

Atmaca, S. & Krauss, S. (2001). Der Einfluss der Aufgabenformulierung auf stochastische Performanz – Das “Drei-Türen-Problem”. In: Stochastik in der Schule, 21, 3, 14-21.

Krauss, S. & Wang, X.T. (2003). The Psychology of the Monty Hall Problem. Discovering Psychological Mechanisms in Solving a Tenacious Brain Teaser. *Journal of Experimental Psychology: General, 132,* 3-22.

Literatur zum Häufigkeitskonzept allgemein:

Krauss, S. (2003). Wie man das Verständnis von Wahrscheinlichkeiten verbessern kann: Das „Häufigkeitskonzept". *Stochastik in der Schule, 23, 1,* 2-9.

Martignon, L., Atmaca, S. & Krauss, S. (2001). Wie kann man Wahlergebnisse und AIDS-Risiken intuitiv darstellen? Ein Kommentar zu den Beiträgen von Hildebrand und Quermann. *Stochastik in der Schule, 21, 1*, 11-12.

Krauss, S. & Hertwig, R. (2000). Muss DNA-Evidenz schwer verständlich sein? Der Ausweg aus einem Kommunikationsproblem*, Monatsschrift für Kriminologie und Strafrechtsreform, 3, 155-162.*

Krauss, S. (2001). Wahrscheinlichkeit und Intuition – 2 Seiten einer Medaille? In: Borovcnik, M., Engel, J. & Wickmann, D. (Hrsg.), *Anregungen zum Stochastik-unterricht: Die NCTM-Standards 2000, Klassische und Bayessche Sichtweise im Vergleich*. Hildesheim: Franzbecker, 139-146.

Wassner, C., Krauss, S. & Martignon, L. (2002). Muss der Satz von Bayes schwer verständlich sein? *Praxis der Mathematik, Heft1/44*, 12-16.

In der Pause ausgelegte Literatur, in der sich u.a. auch weitere reelle Beispiele (z.B. aus der Zeitung) finden lassen:

Borovcnik, M., Engel, J., Wickmann, D.: Anregungen zum Stochastikunterricht, Die NCTM-Standards 2000, Klassische und Bayessche Sichtweise im Vergleich, Hildesheim/Berlin 2001.
Dewdney, A. K.: 200 Prozent von nichts, Basel/Boston/Berlin 1994.
Dubben, H.-H., Beck-Bornholdt, H.-P.: Der Hund, der Eier legt, Hamburg 2010.
Dörner, D.: Die Logik des Mißlingens, Hamburg 1989.
Eichler, A., Vogel, M.: Leitidee Daten und Zufall, Wiesbaden 2009.
Fischer, G.: Stochastik einmal anders, Wiesbaden 2005.
Gigerenzer, G.: Das Einmaleins der Skepsis, Berlin 2002.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L.: Das Reich des Zufalls, Heidelberg/Berlin 1999.
Hauser, B., Humpert, W.: Signifikant? Einführung in statistische Methoden für Lehrkräfte, Zug 2009.
Herget, W. (Hrsg.): Wege in die Stochastik (mathematiklehren, Sammelband), Seelze 2008.
Hughes, P., Brecht, G.: Die Scheinwelt des Paradoxons, Braunschweig 1978.
Knöpfel, H., Löwe, M.: Stochastik – Struktur im Zufall, München/Wien 2007.
Krämer, W.: Denkste! Trugschlüsse aus der Welt des Zufalls und der Zahlen, Frankfurt am Main/New York 1996.
Krämer, W.: So lügt man mit Statistik, Frankfurt am Main/New York 1998.
Krämer, W.: Statistik verstehen, Frankfurt am Main 2010.
Von Randow, G.: Das Ziegenproblem, Hamburg 2009.
Székely, G. J.: Paradoxa, Klassische und neue Überraschungen aus Wahrscheinlichkeits-rechnung und mathematischer Statistik, Thun/Frankfurt am Main 1990.

„Was bedeutet eigentlich ‚signifikant‘? –
Wissenswertes zum Thema Hypothesentesten“

Stefan Krauss

Überblick

1) Zur Rolle der Signifikanztests in der Schule und in der empirischen Sozialforschung

2) *Was bedeutet eigentlich „signifikant“?*
Empirische Befunde zu Fehlvorstellungen zum Signifikanzkonzept

3) *Woher kommen diese Fehlvorstellungen?*
Historische und psychologische Wurzeln

4) *Was kann man gegen die Fehlvorstellungen tun?*
Didaktische Konzepte für ein besseres Verständnis des „Signifikanztestens“

Brauchen wir Signifikanztests in der Schule?

Deschauer (1999) schreibt dazu:

„Ich denke doch, dass man in den frühen achtziger Jahren die Chancen eines anwendungsorientierten Mathematikunterrichts zu optimistisch eingeschätzt und die Stochastik-Leistungskurse etwas zu bombastisch ausgebaut hat. Welcher fachliche Aufwand muss betrieben werden, damit man zur beurteilenden Statistik gelangt! Welcher Leistungskurs-Absolvent ohne nachfolgendes Mathematikstudium kommt einmal in die Situation, Hypothesen testen zu müssen!“

Nun, z.B. Studentinnen und Studenten der …

Medizin, Wirtschaftswissenschaften, Politikwissenschaften, Soziologie, Publizistik, Kommunikationswissenschaften, Erziehungswissenschaften, Pharmazie, Psychologie, Linguistik, Chemie, Biologie, Geologie, Geographie, Pädagogik, Sport, usw, …

→ Statistik ist das meist unterrichtete Fach an deutschen Universitäten

An vielen Universitäten (v.a. im angloamerikanischen Sprachraum) ist oft neben einem Studium der Mathematik auch ein eigenes, separates Studium der Statistik möglich.

→ **Das häufigste Wort dabei ist „Signifikant“**

Signifikanz als Gütesiegel seriöser Wissenschaft

Signifikanztests in psychologischen Zeitschriften

- 1934 – 1940: 17 Artikel mit Signifikanztests
- 1940 – 1955: „Inferenzrevolution“ (Gigerenzer et al., 1989)
- ab 1955: ca. 80% aller wissenschaftlicher psychologischer Arbeiten verwenden Signifikanztests
- heute: knapp 100%

Signifikanz als Gütesiegel seriöser Wissenschaft

Aus dem *Publication Manual* (1974) der American Psychological Association (APA):

> *Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such. Treat the result section like an income tax return. Take what's coming to you, but no more* (p.19)

Diese Passage wurde in der dritten Ausgabe (1983) wieder gestrichen

1) Zur Rolle der Signifikanztests in der Schule und in der empirischen Sozialforschung

2) *Was bedeutet eigentlich „signifikant“?*
Empirische Befunde zu Fehlvorstellungen zum Signifikanzkonzept

3) *Woher kommen diese Fehlvorstellungen?*
Historische und psychologische Wurzeln

4) *Was kann man gegen die Fehlvorstellungen tun?*
Didaktische Konzepte für ein besseres Verständnis des „Signifikanztestens“

Geschichtliche Anekdote

John Arbuthnott (1710, Arzt von Queen Ann) war wahrscheinlich der erste, der einen „Nullhypothesentest“ durchführte:

- Da Männer zu diesen Zeiten gefährlicher lebten als Frauen, müsse Gott mit einer höheren Geburtenrate von Jungen dafür sorgen, dass die Voraussetzungen allgegenwärtiger Monogamie weiterhin gegeben sind
- Er stellte fest, dass in 82 Jahren Geburtenaufzeichnungen in London jedes mal mehr Jungen als Mädchen geboren waren.
- Um seine Hypothese zu untermauern, berechnete er die Wahrscheinlichkeit, ein solches Ergebnis „zufällig“ zu erhalten, mit $(1/2)^{82}$
- Da diese Zahl astronomisch klein ist, ist nach Arbuthnott nicht nur die Existenz Gottes bewiesen, sondern auch dessen Vorliebe für Monogamie

Was bedeutet die Aussage „auf dem 5%-Niveau signifikant“?

Schülerantworten:

- Auf dem 5%-Niveau können nur 5% überzufällig auftreten. Der Rest bewegt sich im Bereich des Zufalls.
- Die Richtigkeit des Ergebnisses ist mit 95% Wahrscheinlichkeit bewiesen.
- 5%-Signifikanzniveau heißt, dass in 5 von 100 Fällen ist das Ergebnis zufällig entstanden.
- Es bedeutet, dass er um 5% über der Zufalls-Prozentualität liegt.
- Das bedeutet, die Wahrscheinlichkeit, dass das Ergebnis ein Irrtum ist, beträgt 5%.
- 5% Signifikanz bedeutet: mit 95% Wahrscheinlichkeit wird sich bei Versuchswiederholung das gleiche Versuchsergebnis einstellen.

„Was bedeutet die Aussage, ein statistischer Kennwert sei auf dem 5%-Niveau signifikant?“

Studentenantworten:

- Ein Signifikanzniveau von 5% bedeutet, dass mit einer Wahrscheinlichkeit von 95% das Ergebnis nicht durch Zufall zustande gekommen ist, sondern eine repräsentative Wiedergabe der Verhältnisse in der Grundgesamtheit durch die Stichprobenmitglieder erzielt wurde.
- Es handelt sich annähernd um eine Normalverteilung. Es wurde eine Anzahl von weniger als 30 Versuchspersonen untersucht. Der α-Fehler beträgt .05. 95% der Aussage lässt sich aus dem Treatment erklären. H_0 ist bestätigt.

Empirische Untersuchung: *Was bedeutet eigentlich „signifikant“?*

Stellen Sie sich vor, Sie wenden einen einfachen t-Test für unabhängige Stichproben an, um einen Mittelwertsunterschied zwischen einer Experimental- und einer Kontrollgruppe zu untersuchen. Der Unterschied zwischen den Gruppen ist auf dem 1%-Niveau signifikant (genauer: t = 2,7, df = 18 Freiheitsgrade, p = 0,01). Bitte markieren Sie jede der folgenden Aussagen als „richtig“ oder „falsch“. „Falsch“ bedeutet, dass die Aussage nicht streng logisch aus den o.g. Prämissen folgt. Es können auch mehrere oder gar keine richtigen Aussagen dabei sein!

Was bedeutet eigentlich „signifikant“?

1) Es ist eindeutig bewiesen, dass die Nullhypothese (dass zwischen den Populationsmittelwerten kein Unterschied besteht) falsch ist. *richtig* □ *falsch* □
2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden. *richtig* □ *falsch* □
3) Es ist eindeutig bewiesen, dass Ihre Alternativhypothese (dass es einen Unterschied zwischen den Populationsmittelwerten gibt) wahr ist. *richtig* □ *falsch* □
4) Man kann nun die Wahrscheinlichkeit ableiten, dass die Alternativ-hypothese richtig ist. *richtig* □ *falsch* □
5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann weiß man jetzt die Wahrscheinlichkeit, dass diese Entscheidung falsch sein könnte. *richtig* □ *falsch* □
6) Der experimentelle Befund ist reliabel in dem Sinne, dass man in 99% der Fälle ein signifikantes Ergebnis bekäme, wenn man das Experiment sehr oft wiederholen würde. *richtig* □ *falsch* □

Was bedeutet eigentlich „signifikant“?

1) Es ist eindeutig bewiesen, dass die Nullhypothese (dass zwischen den Populationsmittelwerten kein Unterschied besteht) falsch ist. *richtig* □ *falsch* ☒
2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden. *richtig* □ *falsch* □
3) Es ist eindeutig bewiesen, dass Ihre Alternativhypothese (dass es einen Unterschied zwischen den Populationsmittelwerten gibt) wahr ist. *richtig* □ *falsch* □
4) Man kann nun die Wahrscheinlichkeit ableiten, dass die Alternativ-hypothese richtig ist. *richtig* □ *falsch* □
5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann weiß man jetzt die Wahrscheinlichkeit, dass diese Entscheidung falsch sein könnte. *richtig* □ *falsch* □
6) Der experimentelle Befund ist reliabel in dem Sinne, dass man in 99% der Fälle ein signifikantes Ergebnis bekäme, wenn man das Experiment sehr oft wiederholen würde. *richtig* □ *falsch* □

Was bedeutet eigentlich „signifikant“?

1) Es ist eindeutig bewiesen, dass die Nullhypothese (dass zwischen den Populationsmittelwerten kein Unterschied besteht) falsch ist. *richtig* □ *falsch* ☒ X
2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden. *richtig* □ *falsch* □
3) Es ist eindeutig bewiesen, dass Ihre Alternativhypothese (dass es einen Unterschied zwischen den Populationsmittelwerten gibt) wahr ist. *richtig* □ *falsch* ☒ X
4) Man kann nun die Wahrscheinlichkeit ableiten, dass die Alternativ-hypothese richtig ist. *richtig* □ *falsch* □
5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann weiß man jetzt die Wahrscheinlichkeit, dass diese Entscheidung falsch sein könnte. *richtig* □ *falsch* □
6) Der experimentelle Befund ist reliabel in dem Sinne, dass man in 99% der Fälle ein signifikantes Ergebnis bekäme, wenn man das Experiment sehr oft wiederholen würde. *richtig* □ *falsch* □

Was bedeutet eigentlich „signifikant“?

1) Es ist eindeutig bewiesen, dass die Nullhypothese (dass zwischen den Populationsmittelwerten kein Unterschied besteht) falsch ist. *richtig* □ *falsch* ☒
2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden. *richtig* □ *falsch* ☒
3) Es ist eindeutig bewiesen, dass Ihre Alternativhypothese (dass es einen Unterschied zwischen den Populationsmittelwerten gibt) wahr ist. *richtig* □ *falsch* ☒
4) Man kann nun die Wahrscheinlichkeit ableiten, dass die Alternativ-hypothese richtig ist. *richtig* □ *falsch* □
5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann weiß man jetzt die Wahrscheinlichkeit, dass diese Entscheidung falsch sein könnte. *richtig* □ *falsch* □
6) Der experimentelle Befund ist reliabel in dem Sinne, dass man in 99% der Fälle ein signifikantes Ergebnis bekäme, wenn man das Experiment sehr oft wiederholen würde. *richtig* □ *falsch* □

Was bedeutet eigentlich „signifikant“?

1) Es ist eindeutig bewiesen, dass die Nullhypothese (dass zwischen den Populationsmittelwerten kein Unterschied besteht) falsch ist. *richtig* ☐ *falsch* ☒
2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden. *richtig* ☐ *falsch* ☒
3) Es ist eindeutig bewiesen, dass Ihre Alternativhypothese (dass es einen Unterschied zwischen den Populationsmittelwerten gibt) wahr ist. *richtig* ☐ *falsch* ☒
4) Man kann nun die Wahrscheinlichkeit ableiten, dass die Alternativhypothese richtig ist. *richtig* ☐ *falsch* ☒
5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann weiß man jetzt die Wahrscheinlichkeit, dass diese Entscheidung falsch sein könnte. *richtig* ☐ *falsch* ☐
6) Der experimentelle Befund ist reliabel in dem Sinne, dass man in 99% der Fälle ein signifikantes Ergebnis bekäme, wenn man das Experiment sehr oft wiederholen würde. *richtig* ☐ *falsch* ☐

Bei einem Nullhypothesentest wird in den Tabellen ein Wahrscheinlichkeitswert, der sogenannte p-Wert, abgelesen

Dieser bedeutet bei einem Nullhypothesentest:

$p = p(D/H_0)$: *Die Wahrscheinlichkeit der gefundenen Daten (oder noch extremerer Daten), unter der Annahme. die Nullhypothese stimmt*

$p(H_0/D)$, $p(H_1/D)$ *oder gar* $p(H_0)$ *bzw.* $p(H_1)$ *können mit Signifikanztests nicht bestimmt werden*

„Bayesianisches Wunschdenken“

What‘s wrong with significance testing? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe in that it does! What we want to know is „Given these data, what is the probability that H_0 is true?“ But as most of us know, what it tells us is „Given that H_0 is true, what is the probability of these (or more extreme) data?“

Cohen, J. (1994). *American Psychologist, 49*, p. 997

Was bedeutet eigentlich „signifikant“?

1) Es ist eindeutig bewiesen, dass die Nullhypothese (dass zwischen den Populationsmittelwerten kein Unterschied besteht) falsch ist. *richtig* □ *falsch* ☒
2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden. *richtig* □ *falsch* ☒
3) Es ist eindeutig bewiesen, dass Ihre Alternativhypothese (dass es einen Unterschied zwischen den Populationsmittelwerten gibt) wahr ist. *richtig* □ *falsch* ☒
4) Man kann nun die Wahrscheinlichkeit ableiten, dass die Alternativhypothese richtig ist. *richtig* □ *falsch* ☒
5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann weiß man jetzt die Wahrscheinlichkeit, dass diese Entscheidung falsch sein könnte. *richtig* □ *falsch* □
6) Der experimentelle Befund ist reliabel in dem Sinne, dass man in 99% der Fälle ein signifikantes Ergebnis bekäme, wenn man das Experiment sehr oft wiederholen würde. *richtig* □ *falsch* □

Was bedeutet eigentlich „signifikant“?

1) Es ist eindeutig bewiesen, dass die Nullhypothese (dass zwischen den Populationsmittelwerten kein Unterschied besteht) falsch ist. *richtig* □ *falsch* ☒
2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden. *richtig* □ *falsch* ☒
3) Es ist eindeutig bewiesen, dass Ihre Alternativhypothese (dass es einen Unterschied zwischen den Populationsmittelwerten gibt) wahr ist. *richtig* □ *falsch* ☒
4) Man kann nun die Wahrscheinlichkeit ableiten, dass die Alternativ-hypothese richtig ist. *richtig* □ *falsch* ☒
5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann weiß man jetzt die Wahrscheinlichkeit, dass diese Entscheidung falsch sein könnte. *richtig* □ *falsch* ☒
6) Der experimentelle Befund ist reliabel in dem Sinne, dass man in 99% der Fälle ein signifikantes Ergebnis bekäme, wenn man das Experiment sehr oft wiederholen würde. *richtig* □ *falsch* □

Fehlvorstellung 6 findet sich auch im Editorial des

Journals of Experimental Psychology (Melton, 1962):

The level of significance measures the confidence that the results of the experiment would be repeatable under the conditions described (p. 553)

„Replication fallacy“

Wichtig! *Alle Fehlvorstellungen gehen in die selbe Richtung:* Sie alle weisen einem signifikanten Testergebnis *größere* Bedeutung zu als es tatsächlich hat

Was bedeutet eigentlich „signifikant“?

1) Es ist eindeutig bewiesen, dass die Nullhypothese (dass zwischen den Populationsmittelwerten kein Unterschied besteht) falsch ist. *richtig* □ *falsch* ☒
2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden. *richtig* □ *falsch* ☒
3) Es ist eindeutig bewiesen, dass Ihre Alternativhypothese (dass es einen Unterschied zwischen den Populationsmittelwerten gibt) wahr ist. *richtig* □ *falsch* ☒
4) Man kann nun die Wahrscheinlichkeit ableiten, dass die Alternativhypothese richtig ist. *richtig* □ *falsch* ☒
5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann weiß man jetzt die Wahrscheinlichkeit, dass diese Entscheidung falsch sein könnte. *richtig* □ *falsch* ☒
6) Der experimentelle Befund ist reliabel in dem Sinne, dass man in 99% der Fälle ein signifikantes Ergebnis bekäme, wenn man das Experiment sehr oft wiederholen würde. *Fehlt: gegeben Ho ist richtig* *richtig* □ *falsch* ☒

„Unter der Annahme, Ho ist richtig“ fehlt bei allen Aussagen!!!

Ergebnisse Oakes (1986); Haller & Krauss (2002)

% angekreuzt	Deutschland 2000 Psychologische Fachbereiche deutscher Universitäten			USA 1986 (Oakes)
Äußerungen (gekürzt)	Statistik Dozenten	Psychologie Dozenten	Psychologie Studenten	Psychologen
1) H_0 ist widerlegt				1%
2) Wsk der H_0				36%
3) H_1 ist bewiesen				6%
4) Wsk der H_1				66%
5) Wsk des Fehlers 1. Art				86%
6) Wsk einer Replikation				60%
% mindestens einen Fehler				97%
# durchschnittlich angekreuzt				
N				

Ergebnisse Oakes (1986); Haller & Krauss (2002)

% angekreuzt	Deutschland 2000 Psychologische Fachbereiche deutscher Universitäten			USA 1986 (Oakes)
Äußerungen (gekürzt)	Statistik Dozenten	Psychologie Dozenten	Psychologie Studenten	Psychologen
1) H_0 ist widerlegt			34%	1%
2) Wsk der H_0			32%	36%
3) H_1 ist bewiesen			20%	6%
4) Wsk der H_1			59%	66%
5) Wsk des Fehlers 1. Art			68%	86%
6) Wsk einer Replikation			41%	60%
% mindestens einen Fehler			100%	97%
# durchschnittlich angekreuzt			2,5	
N			*44*	

Ergebnisse Oakes (1986); Haller & Krauss (2002)

% angekreuzt	Deutschland 2000 Psychologische Fachbereiche deutscher Universitäten			USA 1986 (Oakes)
Äußerungen (gekürzt)	Statistik Dozenten	Psychologie Dozenten	Psychologie Studenten	Psychologen
1) H_0 ist widerlegt		15%	34%	1%
2) Wsk der H_0		26%	32%	36%
3) H_1 ist bewiesen		13%	20%	6%
4) Wsk der H_1		33%	59%	66%
5) Wsk des Fehlers 1. Art		67%	68%	86%
6) Wsk einer Replikation		49%	41%	60%
% mindestens einen Fehler		90%	100%	97%
# durchschnittlich angekreuzt		2,0	2,5	
N		*39*	*44*	

Ergebnisse Oakes (1986); Haller & Krauss (2002)

% angekreuzt	Deutschland 2000 Psychologische Fachbereiche deutscher Universitäten			USA 1986 (Oakes)
Äußerungen (gekürzt)	Statistik Dozenten	Psychologie Dozenten	Psychologie Studenten	Psychologen
1) H_0 ist widerlegt	10%	15%	34%	1%
2) Wsk der H_0	17%	26%	32%	36%
3) H_1 ist bewiesen	10%	13%	20%	6%
4) Wsk der H_1	33%	33%	59%	66%
5) Wsk des Fehlers 1. Art	73%	67%	68%	86%
6) Wsk einer Replikation	37%	49%	41%	60%
% mindestens einen Fehler	80%	90%	100%	97%
# durchschnittlich angekreuzt	1,9	2,0	2,5	
N	*30*	*39*	*44*	

1) Zur Rolle der Signifikanztests in der Schule und in der empirischen Sozialforschung

2) *Was bedeutet eigentlich „signifikant“?*
Empirische Befunde zu Fehlvorstellungen zum Signifikanzkonzept

3) *Woher kommen diese Fehlvorstellungen?*
Historische und psychologische Wurzeln

4) *Was kann man gegen die Fehlvorstellungen tun?*
Didaktische Konzepte für ein besseres Verständnis des „Signifikanztestens“

Aber in den Lehrbüchern steht es richtig!?

Foster Lloyd Brown, Introduction to Statistical Methods in Psychology

S. 522-523. In: G.A. Miller, R. Buckhout, Psychology: The Science of Mental Life

Probability that the null hypothesis could be correct and that only chance is involved	about 1 in 10,000
Probability that the null hypothesis is incorrect and should be rejected	about 9,999 in 10,000

Probability null hypothesis could be correct	about 1 in 20
Probability null hypothesis is incorrect	about 19 in 20

In seinem Lehrbuch „Introduction to Statistics for Psychology and Education“ zählt J.C. Nunally (1975) die folgenden acht Bedeutungen eines signifikanten Testergebnisses an, die alle zumindest unglücklich sind (S. 194-196):

- the improbability of observed results being due to error
- the probability that an observed difference is real
- if the probability is low, the null hypothesis is improbable
- the statistical confidence ... with odds of 95 out of 100 that the observed difference will hold up in investigations
- the degree to which experimental results are taken „seriously“
- the danger of accepting a statistical result as real when it is actually due only to error
- the degree of faith that can be placed in the reality of the finding
- the investigator can have 95 percent confidence that the sample mean actually differs from the population mean

In seinem Lehrbuch „Introduction to Statistics for Psychology and Education“ zählt J.C. Nunally (1975) die folgenden acht Bedeutungen eines signifikanten Testergebnisses an, die alle zumindest unglücklich sind (S. 194-196):

- the improbability of observed results being due to error
- the probability that an observed difference is real
- if the probability is low, the null hypothesis is improbable
- the statistical confidence ... with odds of 95 out of 100 that the observed difference will hold up in investigations
- the degree to which experimental results are taken „seriously“
- the danger of accepting a statistical result as real when it is actually due only to error
- the degree of faith that can be placed in the reality of the finding
- the investigator can have 95 percent confidence that the sample mean actually differs from the population mean

Am Ende versichert der Autor seinen Lesern: „All of these are different ways to say the same thing“

Woher kommen diese Fehlkonzeptionen?

Abgewandelt aus Gigerenzer et al. (1989) und Gigerenzer, Krauss & Vitouch (2002):

- In den Lehrbüchern wird eine „Hybridtheorie“ unterrichtet. Keiner der Väter des Signifikanztestens, weder Fisher noch Neyman und Pearson, würden dieser Theorie zustimmen (nächste Folie).

- Die tatsächlich auftauchenden Widersprüche werden zu Gunsten eines Kochrezepts totgeschwiegen, das in der empirischen Sozialforschung ritualisiert durchgeführt wird („Nullritual“, vgl. übernächste Folie).

Die Väter der Signifikanztests: Was ist das „Signifikanzniveau“?

„Früher“ Fisher (1935)	**„Später“ Fisher (1956)**	**Neyman und Pearson**
Lege Signifikanz-niveau vor der Durchführung des Tests im Sinne einer Konvention (z.B. 5%) fest. Das Signifikanz-niveau ist eine Eigenschaft des Tests.	Berechne nach der Durch-führung des Tests das exakte Signifikanzniveau aus den Daten (p-Werte). Das Signifikanzniveau ist somit eine Eigenschaft der Daten. Durch den p-Wert wird Information kommuni-ziert. Es wird keine willkürlich festgelegte Konvention mehr benötigt.	Spezifiziere H_0 und H_1. α und β sind die relativen Häufigkeiten eines Fehlers erster bzw. zweiter Art. Lege α und β aufgrund einer Kosten-Nutzen-Abschätzung der möglichen Fehlentschei-dungen vor der Durchführung des Tests fest. α und β sind Eigenschaften des Tests. Der Test liefert lediglich eine Entscheidungsregel. Signifikanzniveau ist „unsinnig“

Das „Nullritual“:

1) *Stelle ein Nullhypothese auf (z.B. „es gibt keinen Mittelwertsunterschied“ oder „die Korrelation ist Null“). Spezifiziere keine Alternativhypothese.*
2) *Wenn der p-Wert kleiner 1% ist, tue so, als hättest du vorher ein Signifikanzniveau von 1% festgelegt und berichte das sehr signifikante Ergebnis*

   Wenn der p-Wert zwischen 1% und 5% ist, tue so, als hättest du vorher ein Signifikanzniveau von 5% festgelegt und berichte das signifikante Ergebnis

   Wenn der p-Wert über 5% ist: Pech gehabt!
3) *Stelle das Ergebnis so dar, als hättest du etwas über die Wahrscheinlichkeit deiner Hypothese oder über die Replizierbarkeit herausgefunden*
4) *Mache das immer so*

Das „Nullritual“: Ist die Bezeichnung „Ritual“ gerechtfertigt?

Nach Dulaney & Fiske (1994) haben soziale Rituale folgende Eigenschaften:

1) *Wiederholung von Handlungen*
2) *Fokussierung auf spezielle Zahlen oder Farben*
3) *Angst vor Sanktionen gegenüber Regelverletzungen*
4) *Wunschdenken statt kritischem Denken*

Der (unbewusste?) Konflikt im Kopf des Forschers (nach Gigerenzer, 1993)

Ich (Fisher)

Papers müssen publiziert und Reviewer befriedigt werden.

Also Ritual durchführen.

Über-Ich (Neyman-Pearson)

Eigentlich sollte ich α und β vorher festlegen und am besten auch mehrere Hypothesen genau spezifizieren. Folge: Schuldgefühle

Es (Bayes)

Was ich WIRKLICH will, ist die Wahrscheinlichkeit meiner Hypothese

Folge: Wunschdenken

1) Zur Rolle der Signifikanztests in der Schule und in der empirischen Sozialforschung

2) *Was bedeutet eigentlich „signifikant“?*
Empirische Befunde zu Fehlvorstellungen zum Signifikanzkonzept

3) *Woher kommen diese Fehlvorstellungen?*
Historische und psychologische Wurzeln

4) *Was kann man gegen die Fehlvorstellungen tun?*
Didaktische Konzepte für ein besseres Verständnis des „Signifikanztestens“

Massnahme 1: Kontrastierung

Drei Paradigmen des Hypothesentestens nebeneinander stellen

- $p(D/H_0)$ wird bei Nullhypothesentests berechnet (Fisher)
- $p(D/H_0)$ und $p(D/H_1)$ werden bei Neyman und Pearson betrachtet
- $p(H/D)$ wird bei Bayesschen Testverfahren, und nur dort, ermittelt

→ *Das geht nur mit einer Einführung von p-Werten!*

Im bayerischen G9 war das übliche Verfahren:

zunächst...	danach…
... werden die bedingte Wahrscheinlichkeit und die Formel von Bayes (im LK) eingeführt, und zwar mit den allgemeinen Ereignissen A und B, ohne die Einführung des Begriffs der Hypothese H.	... werden davon losgelöst Signifikanztests und der Begriff der Hypothese H eingeführt, und zwar unabhängig vom Begriff der bedingten Wahrscheinlichkeit; p-Werte (und somit die *Bedeutung* von Signifikanztests) werden nicht eingeführt)

Im G8:

Satz des Bayes nur noch „versteckt“

Hypothesentests nur noch einseitig, keine p-Werte

Vorschlag zur Vermeidung von Fehlvorstellungen:

vgl. Krauss, S. & Wassner, C. (2001). Wie man das Testen von Hypothesen einführen sollte. Stochastik in der Schule, 21, 1, 29-34.

zunächst...	danach…
... sollten die bedingte Wahrscheinlichkeit und die Formel von Bayes inklusive des Begriffs der Hypothese H eingeführt werden; Die Bayes-Formel ist eine zentrale stochastische Formel, sie kann u.a. zum Hypothesentesten verwendet werden	... sollten Signifikanztests als inverses Konzept dazu behandelt werden, wobei Signifikanz auch als bedingte Wahrscheinlichkeit dargestellt und mit der Formel von Bayes verglichen werden soll; dazu benötigt man p-Werte

Leider entfernen wir uns in Bayern davon immer mehr …
(der Vorschlag wurde empirisch bereits erfolgreich getestet)

Vorschlag zur Vermeidung von Fehlvorstellungen

In Formeln …

Ersetze A und B in Bayesformel durch D und H und mache dadurch deutlich, dass sich mit der Bayesformel Hypothesen testen lassen:

$$p(H/D) = \frac{p(D/H) \cdot p(H)}{p(D/H) \cdot p(H) + p(D/\neg H) \cdot p(\neg H)}$$

Drücke auch die zugrunde liegende Idee des Signifikanztestens mit bedingten Wahrscheinlichkeiten aus: $p(D/H)$ („p-Wert“)

(Weitere) Alternativen aufzeigen …

„It is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail“ A. H. Maslow (1966)

- Effektgrößen, Konfidenzintervalle, Diagramme wie Scatterplots oder simple Balkendiagramme (mit error bars), …
- Automatisches Testen stoppen (z.B. bei sehr großem oder gar keinem Mittelwertsunterschied)

Massnahme 2: Grenzen des Signifikanzbegriffs aufzeigen

	Raucher	Nicht-raucher	**Summe**
krank	30	25	**55**
geheilt	20	25	**45**
Summe	**50**	**50**	**100**

$x^2 = 1{,}01 < 2{,}71$ nicht signifikant

	Raucher	Nicht-raucher	**Summe**
krank	150	125	**275**
geheilt	100	125	**225**
Summe	**250**	**250**	**500**

$x^2 = 5{,}05 > 3{,}84$ signifikant auf dem 5%-Niveau

	Raucher	Nicht-raucher	**Summe**
krank	300	250	**550**
geheilt	200	250	**450**
Summe	**500**	**500**	**1000**

$x^2 = 10{,}10 > 6{,}63$ signifikant auf dem 1%-Niveau

	Raucher	Nicht-raucher	**Summe**
krank	3000	2500	**5500**
geheilt	2000	2500	**4500**
Summe	**5000**	**5000**	**10000**

$x^2 = 101{,}01 > 10{,}83$ hochsignifikant auf dem 0,1%-Niveau

Massnahme 3: Teach the conflicts

- Fisher vs. Neyman-Pearson
- Beide vs. Bayes
- Berichte auch vom Streit und der (nicht enden wollenden) Debatte in den Sozialwissenschaften

→ Auch das kann beitragen zum Abbau der üblichen *Überschätzung* der Bedeutung eines signifikanten Testergebnisses

Hays berichtet jedoch, dass in seinem Statistiklehrbuch sowohl ein Kapitel über Bayesstatistik als auch ein Absatz über den Streit zwischen Fisher und Neyman/Pearson auf Wunsch des Herausgebers wieder entfernt werden musste, da dies den Kochbuchcharakter zerstören würde

Teach the conflicts

The Case Against Statistical Significance Testing

RONALD P. CARVER
University of Missouri–Kansas City

In recent years the use of traditional statistical methods in educational research has increasingly come under attack. In this article, Ronald P. Carver exposes the fantasies often entertained by researchers about the meaning of statistical significance. The author recommends abandoning all statistical significance testing and suggests other ways of evaluating research results. Carver concludes that we should return to the scientific method of examining data and replicating results rather than relying on statistical significance testing to provide equivalent information.

Statistical significance testing has involved more fantasy than fact. The emphasis on statistical significance over scientific significance in educational research represents a corrupt form of the scientific method. Educational research would be better off if it stopped testing its results for statistical significance.

The case against statistical significance testing has been developed by many critics (see Morrison & Henkel, 1970b). For example, after a detailed analysis Bakan (1966) concluded that "the test of statistical significance in psychological research may be taken as an instance of a kind of essential mindlessness in the conduct of research" (p. 436); and as early as 1963 Clark made the following comment after comparing various statistical viewpoints: "The null hypothesis of no difference has been judged to be no longer a sound or fruitful basis for statistical investigation. . . . Significance tests do not provide the information that scientists need, and, furthermore, they are not the most effective method for analyzing and summarizing data" (pp. 466, 469). Shulman (1970) admonished that "the time has arrived for educational researchers to divest themselves of the yoke of statistical hypoth-

The preparation of this paper was supported in part by the U. S. Office of Naval Research, Contract No. N00014-75-C-0958. The constructive criticisms of Martin Levit, James Hoffman, Daniel Tira, Joseph Wolff, Richard Cahoon, Ron Karraker, Bill Ghiselli, Bill Lewis, Jack Lutz, Roger Carlson, Robert Leibert, Marilyn Eanet, David Rindskopf, and students of the author are gratefully acknowledged.

Harvard Educational Review, Vol. 48, No. 3, August 1978.

0017-8055/78/0200-0378$01.79/0

ournal of Consulting and Clinical Psychology
:ogy, Vol. 62, No. 1, 75-82

0022-006X/94/$3.00

Teach the conflicts

METHODOLOGICAL DEVELOPMENTS

Misuse of Statistical Tests in Three Decades of Psychotherapy Research

Reuven Dar, Ronald C. Serlin, and Haim Omer

This article reviews the misuse of statistical tests in psychotherapy research studies published in the *Journal of Consulting and Clinical Psychology* in the years 1967–1968, 1977–1978, and 1987–1988. It focuses on 3 major problems in statistical practice: inappropriate uses of null hypothesis tests and *p* values, neglect of effect size, and inflation of Type I error rate. The impressive frequency of these problems is documented, and changes in statistical practices over the past 3 decades are interpreted in light of trends in psychotherapy research. The article concludes with practical suggestions for rational application of statistical tests.

For half a century, articles criticizing the methodology of traditional null hypothesis testing (where the null hypothesis posits quality of means, zero correlation, etc.) have appeared regularly in journals of the social sciences, especially with regard to research in psychology. The most severe criticisms questioned the basic rationale of this methodology (e.g., Berkson, 1942; Carver, 1978; Rozeboom, 1960), some going as far as to blame for the slow progress in psychology (Dar, 1987; Meehl, 1978). Most of these critics advocated abolishing null hypothesis tests altogether, but these recommendations have had little impact: Null hypothesis tests seem to be an institutionalized tradition (Gigerenzer, 1987), unlikely to forgo its central role in psychological research methodology.

Others critics, rather than urging to do away with null hypothesis tests, have suggested various adaptations devised to make hypothesis testing methodology more rational, such as making statistical tests tougher (Serlin & Lapsley, 1985), eliminating their decisional role (Folger, 1989), and paying more attention to considerations of power (Cohen, 1969). Strikingly, however, these proposals also seem to have had little impact on associated methodology has been so refractory to criticism? Two factors may have contributed to this puzzling phenomenon. First, it is possible that researchers and editors fail to connect abstract statistical arguments with what they are actually doing. One goal of this article, therefore, is to sample the actual statistical testing practices in psychological research, thereby making the discussion of common errors and the proposal of rational alternatives more tangible and relevant.

Second, we believe that statistical testing practices are intimately linked to general attitudes toward theory and research. Recently, Omer and Dar (1992) documented a shift from theoretical to pragmatic interests during the past 3 decades of psychotherapy research. Thus, theoretical questions about the mechanisms underlying therapeutic change have been discarded in favor of practical questions concerning the effectiveness of specific treatments and diagnostic methods. On the one hand, this shift was manifested by a rise in the standards of clinical validity, but on the other hand, it was manifested by a decline in the role of theoretical rationales and predictions.

Teach the conflicts

Significance Tests Die Hard

The Amazing Persistence of a Probabilistic Misconception

Ruma Falk and Charles W. Greenbaum
THE HEBREW UNIVERSITY OF JERUSALEM

ABSTRACT. We present a critique showing the flawed logical structure of statistical significance tests. We then attempt to analyze why, in spite of this faulty reasoning, the use of significance tests persists. We identify the illusion of probabilistic proof by contradiction as a central stumbling block, because it is based on a misleading generalization of reasoning from logic to inference under uncertainty. We present new data from a student sample and examples from the psychological literature showing the strength and prevalence of this illusion. We identify some intrinsic cognitive mechanisms (similarity to *modus tollens* reasoning; verbal ambiguity in describing the meaning of significance tests; and the need to rule out chance findings) and extrinsic social pressures which help to maintain the illusion. We conclude by mentioning some alternative methods for presenting and analyzing psychological data, none of which can be considered the ultimate method.

The social sciences in general, and psychology in particular, rely ubiquitously on proving the existence of effects through the standard significance

Teach the conflicts

Why Psychology Will Never be a Real Science Until We Change the Way We Analyze Data

Geoffrey R. Loftus
University of Washington

Much of my adult life has been spent trying to understand the meaning of various data sets—both my own and other peoples' data sets—and that's what I want to talk about today.

As we all know, the fundamental technique for interpreting data in the social sciences is based on hypothesis testing, the basic unit of which works as follows. First you ask some simple question (e.g., is some H_0 true?). Next, you churn your data through some algorithmic process, which produces a simple answer to your question. By tradition (not logic) the usual answer is either "no the null hypothesis is probably false" or "we can't tell whether the null hypothesis is true or false." (Sometimes the answer, either implicitly or explicitly, is "yes the null hypothesis is probably true.") In any event, based on a series of such binary decisions, one tries to make sense of a data set, no matter how complex the data set may be.

In case it isn't clear already, I'll admit my bias right up front. Practically from the time of my very first statistics course, I've been dissatisfied with hypothesis testing as the primary means by which we try to figure things out. I believe that hypothesis testing has channeled our field into a series of methodological cul-de-sacs, and it's been my observation over the years (particularly in the past two years as

I'm by no means the first to issue these kind of complaints. Periodically an article will appear in the *Psychological Bulletin* or *Psychological Review* or the *American Psychologist*, decrying the enormous reliance we place on hypothesis testing (see, for example, Bakan, 1966; Gigerenzer, Swijink, Porter, Daston, Beatty, & Kruger, 1991; Grant, 1962; Loftus, 1991, 1993; 1995; Nunnally, 1960). But these complaints never seem to attract much attention (much less impel action). They are carefully crafted and put forth for consideration, only to just kind of dissolve away in the vast acid bath of our existing methodological orthodoxy.

Today I have two goals. The first is to articulate some of the *reasons* why I'm so disenchanted with hypothesis testing. I don't think that any of these reasons will come as a complete shock to anyone, but I believe it's still useful to consider them in concert. My second goal is to suggest some alternative techniques for extracting more insight and understanding from a data set.

If experience is any guideline I imagine that my remarks will be somewhat controversial (some of my *Memory & Cognition* authors certainly seem to think they are). I'm not so naïve as to think that I can personally force the field to turn on a dime. I do hope, however, that I can stir things up a little.

Teach the conflicts

The major shortcoming of this book is that no attempt has been made to provide a conceptual framework that would orient the reader to the overarching concepts of resiliency. An overview would have brought the larger picture into better focus. Nor, alternatively, is there a concluding, integrative chapter. As it stands, the best general overview of the resiliency literature is provided by Hauser et al. in their chapter, "Family Aspects of Vulnerability and Resilience in Adolescence: A Theoretical Perspective." Their conceptualization of diabetes within a risk and resiliency model also provides an example of how these concepts can be effectively applied to research, with important implications for treatment.

Clinicians and general readers will find this volume informative and enjoyable to read. Those unfamiliar with the area would glean more from this book if they first read a recent review of the concepts of and research on resilience in children (e.g., Garmezy, 1985; Masten, 1989; Masten et al., in press; Rutter, 1990; Werner, 1990).

References

Garmezy, N. (1985). Stress-resistant children: The search for protective factors. In J. E. Stevenson (Ed.), *Recent research in developmental psychopathology. Journal of Child Psychology and Psychiatry* (Book Suppl. 4, pp. 213–233). Elmsford, NY: Pergamon Press.

Masten, A. S. (1989). Resilience in development: Implications of the study of successful adaptation for developmental psychopathology. In D. Cicchetti (Ed.), *The emergence of a discipline: Rochester Symposium on Developmental Psychopathology* (Vol. 1, pp. 261–294). Hillsdale, NJ: Erlbaum.

On the Tyranny of Hypothesis Testing in the Social Sciences

Gerd Gigerenzer, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Krüger
The Empire of Chance: How Probability Changed Science and Everyday Life
Cambridge, England: Cambridge University Press, 1989. 340 pp.
ISBN 0-521-33115-3. $44.50

Review by
Geoffrey R. Loftus

Gerd Gigerenzer, professor of psychology at the University of Salzburg (Austria), is coeditor, with L. Krüger, L. J. Daston, and M. Heidelberger, of The Probabilistic Revolution, Vol. 1: Ideas in History. ■ *Zeno Swijtink is assistant professor of history and philosophy of science at Indiana University Bloomington.* ■ *Theodore Porter is associate professor of history at the University of Virginia (Charlottesville) and author of* The Rise of Statistical Thinking. ■ *Lorraine Daston is professor of the history of science at the University of Göttingen (Germany) and author of* Classical Probability in the Enlightenment. ■ *John Beatty is associate professor of the history of science and technology at the University of Minnesota (Minneapolis).* ■ *Lorenz Krüger is professor of philosophy at the University of Göttingen (Germany).* ■ *Geoffrey R. Loftus, professor of psychology at the University of Washington (Seattle), is coauthor, with E. F. Loftus, of* Essence of Statistics (2nd ed.).

The Empire of Chance is about the history and current use of probability theory and statistics. The book provides a broad treatment of these topics; one could, accordingly, read or review it from a variety of perspectives. Because this review is for psychologists, I will organize it around the book's insights into a question that I believe is at the heart of much malaise in psychological research: How has the virtually [illegible] technique of hypothesis testing come to

edicts issued by Arthur Melton (1962) upon assuming editorship of the august *Journal of Experimental Psychology:*

> Melton's message was, in short, that manuscripts that did not reject the null hypothesis were almost never published, and that results significant only at the 0.05 level were barely acceptable, whereas those significant at the 0.01 level deserved a place in the journal. Psychology students could no longer avoid statistics, and the experimenter who hoped to publish could no longer avoid a test of sig-

Teach the conflicts

The Earth Is Round ($p < .05$)

Jacob Cohen

After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including its near-universal misinterpretation of p *as the probability that* H_0 *is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects* H_0 *one thereby affirms the theory that led to the test. Exploratory data analysis and the use of graphic methods, a steady improvement in and a movement toward standardization in measurement, an emphasis on estimating effect sizes using confidence intervals, and the informed use of available statistical methods is suggested. For generalization, psychologists must finally rely, as has been done in all the older sciences, on replication.*

I make no pretense of the originality of my remarks in this article. One of the few things we, as psychologists, have learned from over a century of scientific study is that at age three score and 10, originality is not to be expected. David Bakan said back in 1966 that his claim that "a great deal of mischief has been associated" with the test of significance "is hardly original," that it is "what 'everybody knows,'" and that "to say it 'out loud' is . . . to assume the role of the child who pointed out that the emperor was really outfitted in his underwear" (p. 423). If it was hardly original in 1966, it can hardly be original now. Yet this naked emperor has been shamelessly running around for a long time.

sure how to test H_0, chi-square with Yates's (1951) correction or the Fisher exact test, and wonders whether he has enough power. Would you believe it? And would you believe that if he tried to publish this result without a significance test, one or more reviewers might complain? It could happen.

Almost a quarter of a century ago, a couple of sociologists, D. E. Morrison and R. E. Henkel (1970), edited a book entitled *The Significance Test Controversy*. Among the contributors were Bill Rozeboom (1960), Paul Meehl (1967), David Bakan (1966), and David Lykken (1968). Without exception, they damned NHST. For example, Meehl described NHST as "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" (p. 265). They were, however, by no means the first to do so. Joseph Berkson attacked NHST in 1938, even before it sank its deep roots in psychology. Lancelot Hogben's book-length critique appeared in 1957. When I read it then, I was appalled by its rank apostasy. I was at that time well trained in the current Fisherian dogma and had not yet heard of Neyman-Pearson (try to find a reference to them in the statistics texts of that day—McNemar, Edwards, Guilford, Walker). Indeed, I had already had some dizzying success as a purveyor of plain and fancy NHST to my fellow clinicians in the Veterans Administration.

What's wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What

Literatur zum Thema Signifikanztests:

Krauss, S. & Wassner, C. (2001). Wie man das Testen von Hypothesen einführen sollte. *Stochastik in der Schule, 21, 1*, 29-34.

Gigerenzer, G. & Krauss, S. (2001). Statistisches Denken oder statistische Rituale? Was sollte man unterrichten? In: Borovcnik, M., Engel, J. & Wickmann, D. (Hrsg.), *Anregungen zum Stochastikunterricht: Die NCTM-Standards 2000, Klassische und Bayessche Sichtweise im Vergleich*. Hildesheim: Franzbecker, 53-62.

Viele der englischen Beispiele des Vortrags finden sich in:

Gigerenzer, G., Krauss, S. & Vitouch, O. (2004). *The Null Ritual. What You Always Wanted to Know About Null Hypothesis Testing, but Were Afraid to Ask*. In: Kaplan, D.: The SAGE Handbook of Quantitative Methodology for the Social Sciences, 389-406.

Vielen herzlichen Dank für Ihre Aufmerksamkeit!