Ein statistisches Experiment mit Schülern auf Bevorzugung von Erfrischungsgetränken

von **Joachim Kunert**, **Frank Lehmkuhl** und **Anja Schleppe**,
Dortmund.

Zusammenfassung: Während einer Projektwoche an einem Dortmunder Gymnasium beschäftigte sich eine Projektgruppe von 18 Schülern aller Altersstufen mit der Untersuchung von Geschmacksunterschieden. Das Projektziel bestand darin, das Interesse der Schüler am Gebiet der Stochastik zu wecken. Auf eine detaillierte Herleitung der verwendeten Testverfahren wurde wegen des unterschiedlichen Alters der Schüler verzichtet. Es ist natürlich möglich, die hier nicht angeführten Details im Mathematik-Unterricht der oberen Klassenstufen herzuleiten. In diesem Zusammenhang möchten wir auf Standardlehrbücher wie das von Hartung (1985) verweisen. Unabhängig vom Alter der Schüler hat jeder einzelne jedoch eine wichtige Erkenntnis aus dem Projekt gewonnen: Das Geschmacksempfinden ist beeinflußbar und subjektiv, dennoch kann es mit Hilfe statistischer Methoden gemessen werden.

1. Einleitung

Im April 1995 fand am Goethe-Gymnasium der Stadt Dortmund eine Projektwoche statt. Eine der Projektgruppen trug den Titel: "Cola-Getränke in Europa".

Ziel des Projektes war es, den Schülern zu vermitteln, daß Geschmack objektiv meßbar ist, wenn man dazu geeignete statistische Methoden verwendet. Der Schwerpunkt lag also auf der richtigen Anwendung der statistischen Methodik und weniger auf der inhaltlichen Seite. Daher waren wir bereit, einige Einschränkungen in der technischen Durchführung der Geschmackstests in Kauf zu nehmen: Die Schüler aus der Projektgruppe sollten nämlich selbst als Versuchsleiter aktiv werden, um so die Probleme und Schwierigkeiten, die bei sensorischen Tests auftreten, direkt zu erfahren.

Wir führten zu dieser Zeit am Fachbereich Statistik ein Anfängerpraktikum für Studenten der Statistik im Grundstudium durch. Auch diese Studenten kamen (jeder für jeweils einen Vormittag) an die Schule, um als Versuchsleiter aktiv zu werden und um als Prüfperson teilzunehmen. Während der Projektwoche waren auch Schüler von Partnerschulen aus England, Frankreich, Italien und aus Weimar zu Gast. Die Schüler der Partnerschulen wurden gezielt angesprochen und gebeten, als Prüfperson mitzuwirken. Daher hatten wir 6 Gruppen von Versuchspersonen - 5 Schülergruppen aus verschiedenen Teilen Europas und eine Gruppe von Studenten. Dies gab uns die Möglichkeit, die Frage zu untersuchen, ob es zwischen den einzelnen Gruppen Unterschiede darin gibt, welche Cola-Sorten sie bevorzugen.

Stochastik in der Schule **15**(1995), Nr. 3, S. 23-42

Insgesamt wurde das Projekt sehr groß: Wir hatten über 100 Prüfpersonen. Dennoch kann man einen Teil des Projektes sehr wohl im normalen Unterricht durchführen.

Dies gilt insbesondere für den sog. Dreieckstest, den wir neben den sog. Bevorzugungstests durchführten. Mit Hilfe des Dreieckstests kann man feststellen, ob es Geschmacksunterschiede zwischen zwei Produkten gibt. Wir untersuchten damit die Frage, ob die Verpackung einen Einfluß auf den Geschmack von Coca Cola hat. Neben der inhaltlichen und didaktischen Begründung (auf die wir gleich eingehen werden) gab es auch einen organisatorischen Grund, zusätzlich einen Dreieckstest zu machen. Die Schüler des Goethe-Gymnasiums sahen, daß wir Cola-Tests durchführten, und sehr viele wollten daran teilnehmen. Damit wir bei den Bevorzugungstests nicht ein zu großes Übergewicht der Gruppe der Dortmunder Schüler hatten und dennoch niemanden abweisen mußten, ließen wir alle interessierten Schüler aus Dortmund beim Dreieckstest mitmachen.

Wir begannen die Projektwoche mit einer Einführung für die 18 Schüler der Projektgruppe. Als erstes versuchten wir, ihnen zu vermitteln, daß ein Geschmackstest als Blindversuch durchgeführt werden muß. Dazu erzählten wir von Problemen, die es bei der Verwendung von PET-Flaschen gibt. So weist z.B. Aström (1993) darauf hin, daß es bei Pfandflaschen aus PET zu Nachgeschmack von vorher darin befindlichen Getränken kommen kann. Wir erzählten auch von dem Gerücht, daß eine Cola in einer einmal geöffneten PET-Flasche am nächsten Tag nicht mehr genießbar sei. Danach gaben wir jedem Teilnehmer zwei Gläser, in eines füllte er Coca Cola aus einer Glasflasche, in das andere Coca Cola aus einer PET-Flasche. Sensibilisiert durch die vorher gegebene Information, stellten fast alle Schüler einen deutlichen Geschmacksunterschied zwischen beiden Gläsern fest. Ganz anders war ihr Geschmackseindruck aber im nachfolgenden Blindversuch. Jetzt erhielt jeder drei Gläser. Davon waren jeweils zwei aus der einen Flasche und eines aus der anderen Flasche gefüllt (ohne daß die Prüfperson wußte, was worin war). Die Prüfperson sollte versuchen herauszufinden, welches der drei Gläser das "besondere" Getränk enthielt. Da die Gläser dabei in der Regel als Dreieck aufgestellt werden, nennt man solch eine Prozedur "Dreieckstest". Hier war sich keiner mehr so sicher, und die meisten identifizierten das falsche Glas.

Nach dieser für die Teilnehmer überraschenden Erkenntnis wurde dann ein Experiment durchgeführt, das verdeutlichen sollte, wieso die Anzahl richtiger Antworten im Dreieckstest einen Beweis für einen Geschmacksunterschied liefern kann. Wir gaben jedem Schüler aus der Projektgruppe einen Würfel und ließen ihn 20 mal werfen. Die Ergebnisse der 20 Würfe sollte er aufschreiben. Gibt es keinen Geschmacksunterschied, so ist die Wahrscheinlichkeit, daß man im Dreieckstest das richtige Glas identifiziert, gleich 1/3, d.h. so groß wie die Wahrscheinlichkeit, eine 1 oder 2 zu würfeln. Die Anzahl der richtigen Ergebnisse bei 20 Prüfpersonen ist also genauso verteilt wie die Anzahl der Einsen und Zweien

bei zwanzigmaligem Würfeln. Wir konnten den Schülern eine Wette anbieten: Keine der Reihen von Würfelergebnissen wird 14 mal oder öfter eine Eins oder Zwei enthalten. Die Wahrscheinlichkeit, bei zwanzigmaligem Würfeln auf 14 oder mehr Einsen oder Zweien zu kommen, ist nämlich ungefähr 1/1000. Das Risiko, daß eine der Würfelreihen das erreicht, ist auch bei einer größeren Gruppe gering. In unserem Fall trat dies auch tatsächlich nicht ein, und wir konnten begründen, daß wir bei einem Dreieckstest mit 20 Prüfpersonen bei 14 richtigen Antworten nicht mehr an Zufall glauben. (Wir tun das sogar schon bei 11 oder mehr richtigen Antworten nicht mehr, nur ist das Risiko eines Irrtums dann 5%).

Versehen mit dieser Erkenntnis, konnten die beiden eigentlichen Versuchsreihen in Angriff genommen werden. Der experimentelle Ablauf der Rangordnungsprüfung wird in Abschnitt 3 näher beschrieben. Im zweiten Abschnitt gehen wir auf die Details bei der Durchführung und Auswertung eines Dreieckstests ein.

Solch ein Dreieckstest scheint uns eine gute Möglichkeit für ein Statistikprojekt im normalen Unterricht zu sein. Wir möchten es zur Nachahmung empfehlen und geben daher genauere Einzelheiten, wie man solch ein Projekt durchführen kann.

2. Der Dreieckstest

2.1 Anleitung zur Durchführung eines Dreieckstests

Bei unserem Dreieckstest verglichen wir Coca Cola aus PET-Flaschen mit Coca Cola aus Glasflaschen. Andere Vergleiche zwischen zwei Produkten mit geringen Geschmacksunterschieden wären genauso möglich. Mit folgenden Arbeitsschritten kann man solch ein Experiment mit einer Schulklasse durchführen.

Die Schüler werden wie folgt aufgeteilt:

Die eine Hälfte der Schüler wird als Prüfperson eingesetzt und befindet sich daher außerhalb des Klassenraumes.

Die andere Hälfte, die als Versuchsleiter eingesetzt wird, bleibt im Klassenraum. Jeder Versuchsleiter bereitet dort unter Anleitung für jeweils eine Testperson das Experiment vor.

Danach wird gewechselt, so daß jeder einmal Versuchsleiter und einmal Prüfperson ist.

Der jeweilige Versuchsleiter geht wie folgt vor:

Zunächst wird der in Abb. 1 (siehe Anhang) dargestellte Arbeitsbogen ausgefüllt. Dabei wird der Inhalt der drei Gläser innerhalb der Dreiecksanordnung von links nach rechts ausgewürfelt und eingetragen:

Würfelergebnis 1,2,3 führt zur Abfüllung aus einer Glasflasche (kurz: Glas),

Würfelergebnis 4,5.6 zur Abfüllung aus einer Plastikflasche (kurz: Plastik).

Wurde bei den beiden ersten Gläsern beide Male Glas bzw. Plastik ausgewürfelt, so ergibt sich die dritte Position automatisch. Wurde dagegen in den beiden ersten Würfen Glas und Plastik ermittelt, so muß auch die dritte Position ausgewürfelt werden.

Anschließend wird für jedes Glas eine zweistellige Kennziffer zur Identifikation ermittelt und in den Bogen eingetragen. Auch dies geschieht mit einem Würfel. Wird auf einem Bogen für zwei Gläser die gleiche Kennziffer ausgewürfelt, so wird für das zweite Glas einfach neu gewürfelt.

Dann werden die Kennziffern in den zugehörigen Prüfbogen eingetragen (siehe Anhang, Abb. 2). Der jeweilige Prüfleiter wartet, bis er mit seiner Prüfung an der Reihe ist. Die Prüfungen werden nacheinander durchgeführt. Für eine Prüfung werden drei Gläser mit ihrer jeweiligen Kennziffer versehen und entsprechend befüllt. Die Gläser werden dann so angeordnet, daß sie aus der Sicht der Prüfperson dem Prüfbogen entsprechend positioniert sind. Der Arbeitsbogen wird so aufbewahrt, daß die Prüfperson ihn nicht einsehen kann. Die Prüfperson wird in den Klassenraum gerufen, und der Prüfbogen wird ihr zusammen mit den Gläsern vorgelegt. Neben den Gläsern sollte ein Teller mit Weißbrotscheiben zur Geschmacksneutralisierung zur Verfügung stehen.

Die Prüfperson geht wie folgt vor:

Sie testet die Getränke in den drei Gläsern in einer beliebigen Reihenfolge, wobei sie natürlich aus jedem Glas beliebig oft probieren kann. Dann kreuzt sie die Kennziffer des "besonderen" Getränks auf dem Prüfbogen an.

Der Versuchsleiter informiert die Prüfperson, ob ihre Entscheidung richtig war, und er macht einen entsprechenden Vermerk auf dem Prüfbogen.

Die Prüfperson verläßt den Klassenraum, eine neue Prüfperson wird zum Test gerufen.

Die gewonnenen Daten werden wie folgt ausgewertet:

Anhand der Prüfbögen wird die Anzahl der richtigen Antworten ermittelt.

Ist diese Anzahl größer oder gleich der nach der Binomialverteilung (siehe auch die Tabelle in DIN ISO 4120) erforderlichen Mindestanzahl korrekter Antworten, so kann auf einen statistisch signifikanten Geschmacksunterschied zwischen beiden Getränken geschlossen werden.

Tab. 1 zeigt für einige Anzahlen von Prüfpersonen die benötigten Mindestanzahlen korrekter Antworten, wobei eine Irrtumswahrscheinlichkeit von 5% eingeräumt wird.

Diese Irrtumswahrscheinlichkeit wird als Wahrscheinlichkeit für einen Fehler 1. Art bezeichnet. Man hält diese Wahrscheinlichkeit kleiner als eine Zahl α, das Niveau des Tests. In der Regel wird α gleich 5% gewählt. Es gibt auch einen Fehler 2. Art; dessen Wahrscheinlichkeit sucht man zu minimieren. Anschaulich lassen sich diese beiden Fehlerarten am Beispiel einer Gerichtsverhandlung darstellen: Ein Angeklagter kommt vor Gericht, weil aufgrund der Beweislage behauptet wird, er sei schuldig. Damit lautet die Nullhypothese, die gegebenenfalls widerlegt wird: "Der Angeklagte ist unschuldig". Zwei Ausgänge des Gerichtsverfahrens sind nun denkbar: Zum einen kann das Gericht korrekt entscheiden, indem es den tatsächlich Schuldigen verurteilt bzw. den tatsächlich Unschuldigen freispricht. Zum anderen kann das Gericht falsch entscheiden, indem es den Unschuldigen verurteilt (Fehler 1. Art) bzw. den Schuldigen freispricht (Fehler 2. Art). Um sicherzustellen, daß man im Zweifel für den Angeklagten entscheidet, kommt dem Fehler 1. Art eine besonders wichtige Bedeutung zu. Daher wird er kontrolliert.

Tab. 1: Mindestanzahl korrekter Antworten ($\alpha = 5\%$)

Anzahl der Prüfpersonen	Mindestanzahl korrekter Antworten bei einem Signifikanzniveau α von 5%
20	11
21	12
22	12
23	12
24	13
25	13
26	14
27	14
28	15
29	15
30	15
31	16
32	16
33	17
34	17
35	17

Nähere Informationen zur Testtheorie finden sich z.B. in dem Buch von Hartung (1985).

2.2 Zu beachtende Rahmenbedingungen

Folgende organisatorische Aspekte sollten bei der Durchführung des Experimentes berücksichtigt werden:

Das Weißbrot, das zum Neutralisieren des Geschmacks zur Verfügung gestellt wird, sollte nicht von den Versuchsleitern verzehrt werden.

Neben einer Spülmöglichkeit in der Nähe des Klassenzimmers muß auch eine Kühlmöglichkeit für die Getränke vorhanden sein. Dabei ist darauf zu achten, daß beide Getränke in demselben Kühlschrank auf gleicher Höhe positioniert und gleich lange gekühlt werden. Dies beeinflußt die Temperatur der Getränke und damit auch das Geschmacksempfinden.

Es muß darauf geachtet werden, daß die Anordnung der Gläser aus dem Blickwinkel der Prüfperson so, wie auf dem Prüfbogen angegeben, erfolgt. Es verfälscht die Ergebnisse, wenn der Versuchsleiter der Prüfperson gegenübersteht und die Anordnung gemäß Arbeitsbogen aus seiner eigenen Perspektive vornimmt und nicht aus derjenigen der jeweiligen Prüfperson. Erfahrungsgemäß schauen die Prüfpersonen nicht auf die Kennziffern der Gläser, sondern kreuzen auf dem Prüfbogen die Position an.

Es ist wichtig, daß die Anordnung der Getränke tatsächlich zufällig erfolgt und für jede Prüfperson neu ausgewürfelt wird.

Für eine Diplomarbeit am Fachbereich Statistik (Rusert 1994) wurden bei einem Lackhersteller Dreieckstests zur Farbwahrnehmung durchgeführt. Ein Teil der Tests wurde mit drei identischen Farbplatten gemacht. Dabei stellte sich heraus, daß dann fast alle Prüfpersonen sich für "links unten" als vermeintlich andere Farbe entschieden. Hätten wir also bei Farbplatten mit minimalen, sensorisch gar nicht wahrnehmbaren Unterschieden nur einmal gewürfelt und zufällig "links unten" als Position für die besondere Farbe erhalten, dann hätten wir mit hoher Wahrscheinlichkeit fast nur richtige Antworten bekommen und auf einen signifikanten Farbunterschied geschlossen!

2.3 Auswertung unseres Dreieckstests

Bei unserem Experiment wurden einen Tag vor Durchführung des Dreieckstests Glas- bzw. Plastikflaschen geöffnet und nach 5 min wieder verschlossen. Es sollte untersucht werden, ob sich am nächsten Tag ein Geschmacksunterschied zwischen den Inhalten der beiden Flaschen zeigt.

An dem Dreieckstest nahmen insgesamt 81 Prüfpersonen im Rahmen der Projektwoche teil. Davon gehörten 18 Schüler zu unserer Projektgruppe. Sie fungierten zunächst als Prüfperson und dann als Versuchsleiter. Von den 81 Prüfpersonen trafen 33 die richtige Entscheidung.

Damit haben wir zwar mehr als ein Drittel an richtigen Antworten. Dennoch reicht es nicht für den Nachweis eines Geschmacksunterschiedes: Für ein signifikantes Ergebnis hätten wir 35 richtige Antworten gebraucht.

Dies ist kein Nachweis dafür, daß es keinen Unterschied gibt. (Man könnte das so formulieren: Ein Freispruch aus Mangel an Beweisen ist kein Beweis der Unschuld). Man muß auch bedenken, daß wir Bedingungen hatten, unter denen subtile Unterschiede den Prüfpersonen entgehen können. Bei der Prüfung gab es sicherlich mehr Ablenkungen, als zu wünschen wären, und die Prüfpersonen waren nicht geübt, auf Geschmacksnuancen zu achten.

Dennoch läßt sich der Schluß ziehen, daß der Unterschied nicht so groß ist, wie wir eigentlich erwartet hatten. Wenn 33 von 81 Prüfpersonen richtig antworten, so sind das 6 mehr als die bei Geschmacksgleichheit zu erwartenden 27. Wir schätzen, daß der Anteil der Personen, die (unter den beim Versuch herrschenden Bedingungen) tatsächlich einen Unterschied bemerkt haben, gleich (33-27) / (81-27) = 1/9 ist. (Das ist immerhin nicht Null). Man beachte, daß dieser erwartungstreue Schätzer auch negativ werden kann; dies geschieht nämlich genau dann, wenn die Anzahl der richtig antwortenden Personen kleiner ist als die bei Geschmacksgleichheit zu erwartende Anzahl.

Nehmen wir unsere Prüfpersonen als repräsentative Stichprobe aller Schüler, so läßt sich aber zeigen, daß mit 90%-iger Wahrscheinlichkeit höchstens 25% der Schüler (unter den bei unserem Versuch herrschenden Bedingungen) einen Unterschied im Geschmack wahrnehmen würden.

Diese Aussage wird im folgenden hergeleitet. Dabei wird eine Sicherheitswahrscheinlichkeit von $1-\alpha=0.9$, d.h. ein Fehler 1. Art von 10% unterstellt. Wir testen nämlich jetzt zweiseitig im Gegensatz zum bisher einseitigen Fall und räumen daher einen doppelt so großen Fehler 1. Art ein; in jede Abweichungsrichtung beträgt er also 5%. Dies führt dazu, daß der Dreieckstest genau dann ablehnt, wenn das Konfidenzintervall für p die Null nicht überdeckt.

Bezeichne X die Anzahl korrekter Antworten, n den Stichprobenumfang und π den unbekannten Anteil korrekter Antworten. Wir wissen, daß X binomialverteilt ist mit Parametern n und π ; laut Zentralem Grenzwertsatz ist diese Binomialverteilung bei genügend großem Stichprobenumfang durch eine Normalverteilung mit Erwartungswert $n\pi$ und Varianz $n\pi(1-\pi)$ angenähert. Damit ergibt sich ein 90%-Konfidenzintervall für π durch Auflösen der Gleichung

$$P(|\frac{X-n\pi}{\sqrt{n\pi(1-\pi)}}|\le 1.64)=0.9$$

nach π . Dabei ist 1.64 gerade das 95%-Quantil der Standardnormalverteilung. Der Term $n\pi(1-\pi)$ wird nach oben abgeschätzt durch $n/4$. Man erhält dann durch entsprechendes Umformen die Ungleichung

$$P(|\frac{X-n\pi}{\sqrt{n}}|\le 0.82)\ge 0.9$$

und damit ein 90%-Konfidenzintervall für π der Form

$$\left[\frac{X-\sqrt{n}\cdot 0.82}{n};\frac{X+\sqrt{n}\cdot 0.82}{n}\right]=\left[\frac{33-9\cdot 0.82}{81};\frac{33+9\cdot 0.82}{81}\right]=[0.32;0.50] .$$

Also ist π mit mindestens 90%-iger Wahrscheinlichkeit kleiner oder gleich 0.5; zugleich gilt für π nach dem Satz von der totalen Wahrscheinlichkeit:

$$\pi=\frac{1}{3}(1-p)+p .$$

Der unbekannte Anteil korrekter Antworten π setzt sich nämlich zusammen aus dem Anteil der Personen, die zufälligerweise korrekt identifiziert haben (er ist gleich $\frac{1}{3}(1-p)$) und aus dem Anteil derjenigen, die tatsächlich einen Unterschied geschmeckt haben (er ist gleich p). Löst man die Gleichung nach p auf und berücksichtigt das Konfidenzintervall für π , so erhält man:

$$P(-\frac{1}{50}\le p\le\frac{1}{4})=0.9 .$$

Daraus folgt die Behauptung: $P(p\le\frac{1}{4})=0.9$.

Die theoretischen Details können z.B. in Hartung (1985) nachgelesen werden.

3. Die Rangordnungsprüfung

3.1 Vorbemerkungen zum Experiment

Wir verglichen hier fünf Sorten von Cola-Getränken miteinander. Mehr Sorten hätten wahrscheinlich zu sensorischer Ermüdung geführt. Mit fünf Sorten hatten wir aber auch eine hinreichende Komplexität der Ergebnisse, da es 120 mögliche Reihenfolgen für die Bevorzugung gab.

Die Vorgehensweise entspricht prinzipiell derjenigen bei der Durchführung des Dreieckstests. Die Anordnung der 5 zu testenden Cola-Sorten erfolgt wiederum zufällig. Dies sollte vermeiden, daß beim Probieren der Getränke in der vorgegebenen Reihenfolge von links nach rechts beispielsweise das zuerst probierte Getränk als das geschmacklich beste identifiziert wird, weil man da noch durstig war.

Die übrigen in Abschnitt 2.2 genannten Rahmenbedingungen sind hier analog zu beachten.

Abb. 3 (siehe Anhang) zeigt den vom Versuchsleiter auszufüllenden Arbeitsbogen, in den auch die Ergebnisse des Experimentes einzutragen sind. Abb. 4 (siehe Anhang) zeigt den zugehörigen Prüfbogen, der von der Prüfperson auszufüllen ist. Dabei wird den formulierten Anweisungen gefolgt.

3.2 Beschreibung des Experimentes

Es testeten insgesamt 109 Prüfpersonen in einem Blindversuch jeweils unter der Aufsicht eines Versuchsleiters. Neben den Schülern der Projektgruppe am Goethe-Gymnasium und den Studenten des Anfängerpraktikums am Fachbereich Statistik der Universität Dortmund gehörten auch Schüler aus fünf weiteren Gymnasien in Europa zur Gruppe der Prüfpersonen.

Die Produkte waren für jede Prüfperson in identischen Gläsern mit identischer Temperatur in einer zufälligen Reihenfolge angeordnet. Zur Identifikation bei der Auswertung waren die Einzelgläser mit zweistelligen Zufallszahlen markiert. Die Prüfpersonen wußten nicht, welche Produkte sie jeweils testeten. Jede Prüfperson sollte die Produkte nach ihrem Geschmack anordnen: Das Produkt, das ihr am meisten zusagte, erhielt 5 Punkte, das am wenigsten gemochte Produkt nur 1 Punkt.

Mit Hilfe dieser Daten lassen sich folgende interessierende Fragen beantworten:

Werden alle Produkte gleich gemocht, oder gibt es Unterschiede?

Wenn es Unterschiede gibt, zwischen welchen Produkten liegen sie?

Gibt es Unterschiede in der Beurteilung zwischen den Gruppen?

Die hierzu notwendigen Verfahren sowie die jeweiligen Analyseergebnisse werden im folgenden vorgestellt.

3.3 Ergebnisse und Interpretation

Jede Prüfperson sollte in dem Versuch eine Rangliste angeben, auch wenn alle Produkte ihr gleich gut (oder ungut) schmeckten. Wenn jemand also gar keine Unterschiede bemerkt hatte, so erhielten wir auch von dieser Person eine Rangliste. Wir wollen untersuchen, ob die erhaltenen Ranglisten Zufallsprodukte sind, oder ob man aus den Daten auf echte Unterschiede in der Bevorzugung der Produkte bei unseren Prüfern schließen kann.

Berechnet man aus allen 109 Urteilen die durchschnittliche Gesamtpunktzahl je Produkt, so erhielt

- River Cola die durchschnittliche Punktzahl 2.86,
- Classic Cola 3.23,
- Pepsi Cola 3.41,
- Cola-Hartinger / Wesergold 1.97 und
- Coca Cola 3.52.

Die durchschnittliche Punktzahl aus den einzelnen Gruppen zeigt Tab. 2.

Tab. 2: Urteile der Prüfpersonen

Gruppe von Prüfpersonen	Anzahl der Testteilnehmer in der Gruppe	Durchschnittliche Punktzahl für das Produkt				
		River Cola (Aldi)	Classic Cola (Lidl)	Pepsi Cola	Cola-Hartinger / Wesergold (Rewe)	Coca Cola
Engländer	8	3.4	2.6	2.4	3.4	3.3
Franzosen	5	3.0	3.2	4.2	2.0	2.6
Dortmunder	32	2.9	3.2	3.3	1.7	3.9
Italiener	21	3.2	3.0	3.5	1.9	3.4
Studenten	24	2.5	3.7	3.5	1.9	3.5
Weimarer	19	2.6	3.3	3.7	2.05	3.4

Man sieht, daß Coca Cola und Pepsi Cola von den Versuchspersonen im Schnitt etwas besser eingeordnet wurden als die übrigen Produkte, und man sieht, daß Cola-Hartinger / Wesergold von unseren Prüfpersonen doch deutlich schlechter eingeordnet wurde.

Dieses Ergebnis wird im folgenden durch entsprechende formale Analysen unterstützt, die nicht mehr mit den Schülern, sondern nur noch mit den Studenten durchgeführt wurden. Zunächst steht die Beantwortung von Frage 1 aus Abschnitt 3.2 im Vordergrund.

3.3.1 Der Friedman-Test zur Prüfung auf allgemeine Produktunterschiede

Der Friedman-Test überprüft die Nullhypothese: "Die vergebenen Rangreihenfolgen sind zufällig", deren Ablehnung den Nachweis signifikanter Produktunterschiede bedeutet. Dabei räumen wir eine Irrtumswahrscheinlichkeit von 5% ein. Die Überprüfung der Nullhypothese erfolgt mit Hilfe des Friedman-Wertes F, der wie folgt berechnet wird:

$$F = \frac{12}{nk(k+1)}(R_1^2 + R_2^2 + ... + R_k^2) - 3n(k+1),$$

wobei $k = 5$ die Anzahl der untersuchten Produkte,

n die Anzahl der Prüfpersonen und

$R_1, R_2 ... R_k$ die Rangsummen der k untersuchten Produkte bei n Prüfpersonen (also die Durchschnittspunktzahl $\times n$)

bezeichnet.

Der Friedman-Wert wird anschließend mit einem entsprechenden kritischen Wert verglichen, der vertafelt vorliegt (z.B. Büning; Trenkler 1994 oder ISO 8587). Bei 109 Prüfpersonen und 5 Produkten ist der kritische Wert gleich 8.96. Wir erhielten einen Friedman-Wert von 68.5 (was ja deutlich größer als 8.96 ist). Daraus kann abgeleitet werden, daß in der Bewertung der 5 Produkte echte, überzufällige Unterschiede bestehen.

Es interessiert die Frage, ob es auch in den Einzelgruppen signifikante Unterschiede zwischen den Produkten gibt. Als Friedman-Werte in den Einzelgruppen erhält man Tabelle 3.

Tab. 3: Berechnung der Friedman-Werte

Gruppe von Prüfpersonen	Friedman-Wert
Engländer	2.8
Franzosen	5.28
Dortmunder	32.9
Italiener	14.59
Studenten	24.07
Weimarer	13.85

Auch wenn man in einer einzelnen Gruppe eine eigene Analyse durchführt, kann man den Wert der Friedman-Statistik jeweils mit dem kritischen Wert 8.96 vergleichen. Lediglich im Fall der französischen Schüler beträgt der kritische Wert 11.52, da wir hier nur sehr wenige Prüfpersonen hatten. In den Einzelgruppen sind die Friedman-Statistiken somit im allgemeinen auch signifikant. Lediglich in der Gruppe der Engländer und der Franzosen sind keine signifikanten Unterschiede zwischen den Produkten festzustellen. Dies ist durch die geringe Zahl der Schüler in diesen Gruppen erklärbar. Allerdings fällt bei den Engländern auf, daß sich die Durchschnittsnoten nur sehr wenig unterscheiden (siehe Tab. 2).

3.3.2 Der Friedman-Test zur Prüfung auf paarweise Produktunterschiede

Dieser Test kann angewendet werden, wenn im vorhergehenden Schritt allgemeine Produktunterschiede nachgewiesen werden konnten. Mit Hilfe paarweiser Produktvergleiche wird dann versucht herauszufinden, zwischen welchen Produkten Unterschiede empfunden werden. Damit wird Frage 2 aus Abschnitt 3.2 beantwortet. Wir führen den Test wieder zunächst für die Gesamtheit aller Prüfpersonen durch und anschließend für diejenigen Einzelgruppen, bei denen der Friedman-Test signifikant war.

Überprüft wird die Nullhypothese: "Es besteht kein Unterschied in der Einschätzung der Produkte A und B", die beim Nachweis eines überzufällig großen Rangunterschiedes abgelehnt werden kann. Dabei räumt man wieder eine Irrtumswahrscheinlichkeit von 5% ein.

Die Überprüfung der Nullhypothese erfolgt mit Hilfe der betragsmäßigen Differenz der zugehörigen Rangsummen R_A und R_B, die dann mit einem kritischen Wert c der Gestalt

$$c = 1.96 \cdot \sqrt{\frac{nk(k+1)}{6}}$$

verglichen wird; hierbei bezeichnet wieder n die Anzahl der Prüfpersonen und $k = 5$ die Anzahl der untersuchten Produkte. Ist die beschriebene Differenz größer oder gleich dem kritischen Wert, so kann abgeleitet werden, daß sich die beiden Produkte A und B signifikant unterscheiden. Tab. 4 zeigt die betragsmäßigen Differenzen der jeweiligen Rangsummen und

Tab. 5 die zugehörigen gerundeten kritischen Werte.

Tab. 4: Paarvergleiche der Produkte je Gruppe von Prüfpersonen

Produktpaar A / B	$\lvert R_A - R_B \rvert$ je Gruppe				
	alle	Dortmunder	Italiener	Studenten	Weimarer
River/Classic	40	7	4	29	13
River/Pepsi	60	11	5	24	22
River/Hart.	97	39	29	14	10
River/Coca	72	31	3	26	15
Classic/Pepsi	20	4	9	5	9
Classic/Hart.	137	46	25	43	23
Classic/Coca	32	24	7	3	2
Pepsi/Hart.	157	50	34	38	32
Pepsi/Coca	12	20	2	2	7
Hart./Coca	169	70	32	40	25

Tab. 5: Kritische Werte je Gruppe von Prüfpersonen

	alle	Dortmunder	Italiener	Studenten	Weimarer
kritischer Wert	46	25	21	22	20

Die Ergebnisse des Paarvergleichs lassen sich anschaulich wie folgt darstellen: Die Produkte werden von links nach rechts so angeordnet, daß das jeweils beliebteste Produkt links steht und der Beliebtheitsgrad dann nach rechts abnimmt. Zwei Produkte mit nicht signifikantem Unterschied werden gemeinsam unterstrichen.

Es ergibt sich folgende Darstellung:

Coca Pepsi Classic River Hartinger	für alle Prüfpersonen
Coca Pepsi Classic River Hartinger	für die Schüler aus Dortmund
Classic Coca Pepsi River Hartinger	für die Studenten der Universität Dortmund
Pepsi Coca River Classic Hartinger	für die Schüler aus Italien
Pepsi Coca Classic River Hartinger	für die Schüler aus Weimar

Man sieht aus dieser Darstellung, daß sich bei der Gesamtheit der Prüfpersonen keine signifikanten Unterschiede zwischen Coca und Pepsi, Coca und Classic sowie Pepsi und Classic zeigten. Daher sind die drei Produkte gemeinsam unterstrichen. Zwischen Classic und River zeigten sich ebenso keine Unterschiede, aber zwischen Coca und River und zwischen Pepsi und River. Daher wurde der Strich unter den ersten drei Produkten nicht einfach auf River ausgedehnt, sondern ein neuer Strich zwischen Classic und River gezogen.

Insgesamt läßt sich damit festhalten, daß sich keine Unterschiede zwischen den Produkten Classic Cola, Pepsi Cola sowie Coca Cola nachweisen lassen. Dies gilt auch, wenn man die genannten Gruppen einzeln analysiert. Außerdem wurde das Produkt Cola-Hartinger / Wesergold durchweg als das schlechteste der getesteten Produkte klassifiziert. Eine Ausnahme bildeten hierbei lediglich die englischen Schüler, die jedoch aufgrund ihrer geringen Anzahl in den Paarvergleich nicht miteinbezogen werden konnten.

Interessant ist das schlechte Abschneiden von River Cola. Von geschulten Prüfpersonen wurde dieses Getränk sehr viel besser eingeordnet (Test 1991).

Nähere Angaben zum mathematischen Vorgehen finden sich in der ISO 8587 / Sensory Analysis - Methodology - Ranking. (Die entsprechende deutsche Norm DIN 10 963 / Sensorische Prüfverfahren - Rangordnungsprüfung wird z.Zt. gerade überarbeitet, sie soll an die ISO angepaßt werden).

3.3.3 Ein Permutationstest zur Prüfung auf regionale Unterschiede

Besonders interessant bei solch einem europaweiten Vergleich ist natürlich die Frage 3 aus Abschnitt 3.2, ob zwischen den einzelnen Prüfergruppen Unterschiede in der Bewertung der Produkte bestehen. Wir nehmen dazu unsere Studenten als Referenzgruppe und fragen, ob eine der fünf anderen Gruppen signifikant andere Einschätzungen geliefert hat.

Dazu führen wir Permutationstests durch, die klären sollen, ob die Studenten der Universität Dortmund systematisch andere Rangfolgen vergeben haben als eine

oder mehrere Schülergruppen. Hierzu wird für jede Schülergruppe ein Paarvergleich zwischen ihren Urteilen und den Urteilen der Studenten durchgeführt. Für jedes der 5 Gruppenpaare überprüft das Verfahren die Nullhypothese: "Es besteht kein Unterschied im Benotungsverhalten der beiden Gruppen". Insgesamt räumt man wieder eine Fehlerwahrscheinlichkeit von 5% ein. Da 5 Tests gleichzeitig durchgeführt werden, kann man bei jedem einzelnen Paarvergleich nur eine Fehlerwahrscheinlichkeit von je 1% einräumen. Der Permutationstest soll am Beispiel des Paarvergleichs mit den englischen Schülern verdeutlicht werden.

Als Maßzahl für den Vergleich dient der euklidische Abstand der Durchschnittsnoten. Er wird berechnet, indem die Differenzen der Durchschnittsnoten von Engländern und Studenten (siehe Tab. 2) je Produkt gebildet, quadriert und aufsummiert werden und aus der resultierenden Summe die Wurzel gezogen wird. Zwischen Engländern und Studenten hatte sich dabei der euklidische Abstand 2.33 ergeben.

Gäbe es keinen Unterschied im Benotungsverhalten zwischen den 8 Engländern und den 24 Studenten, so wären die Noten dieser Prüfpersonen unabhängig von ihrer Herkunft. Es hätten also mit gleicher Wahrscheinlichkeit 8 andere von den 32 erhobenen Rangordnungen aus beiden Gruppen von Engländern erstellt worden sein können. Der Permutationstest wählt unter den insgesamt 32 Urteilen 8 zufällig aus und bezeichnet sie als "Urteile der Engländer". Die übrigen 24 werden als "Urteile der Studenten" bezeichnet. Die Durchschnittsnoten der "Pseudo-Engländer" sowie der "Pseudo-Studenten" werden berechnet und miteinander verglichen.

Von den insgesamt 10 518 300 möglichen Auswahlen von 8 aus 32 Urteilen wurden 1000 mit Hilfe eines PC zufällig ausgewählt. Bei jeder Auswahl wurde der simulierte euklidische Abstand der Durchschnittsnoten berechnet. Die Nullhypothese kann dann mit einer Fehlerwahrscheinlichkeit von 1% abgelehnt werden, wenn von den insgesamt 1000 ermittelten simulierten euklidischen Abständen höchstens 10 mal der tatsächliche euklidische Abstand 2.33 überschritten wird. Da dieses Ereignis nur bei 6 der 1000 Simulationsläufe auftrat, kann auf einen signifikanten Unterschied im Benotungsverhalten zwischen Engländern und Studenten geschlossen werden.

Tab. 6 zeigt die euklidischen Abstände der wahren Urteile zwischen den Studenten und jeder anderen Gruppe, Tab. 7 gibt an, wie oft der jeweilige tatsächliche euklidische Abstand in 1000 Simulationsläufen jeweils überschritten worden ist.

Insgesamt stellt man nur zwischen Studenten und Engländern einen signifikanten Unterschied im Bewertungsverhalten fest. Dieser Unterschied soll abschließend durch eine graphische Darstellung der Urteile der englischen Schüler und der Dortmunder Studenten gezeigt werden.

Tab. 6: Euklidische Abstände zur Gruppe der Studenten

Gruppe von Prüfpersonen	euklidischer Abstand
Engländer	2.33
Franzosen	1.40
Dortmunder	0.82
Italiener	1.01
Weimarer	0.56

Tab. 7: Anzahl der Überschreitungen des tatsächlichen euklidischen Abstands

Gruppe von Prüfpersonen	Anzahl Überschreitungen
Engländer	6
Franzosen	424
Dortmunder	336
Italiener	254
Weimarer	802

3.3.4 Graphische Analyse

Die Durchschnittsränge der Produkte enthalten nicht die gesamte Information. So könnte ja beispielsweise eine durchschnittliche Note für ein Produkt auch dadurch zustandekommen, daß die Hälfte der Prüfpersonen ihm den Rang 1 (schlechtestes Produkt) und die andere Hälfte ihm den Rang 5 (bestes Produkt) gegeben hätte. Um solche Untergruppen gegebenenfalls erkennen zu können, ist eine graphische Darstellung von Nutzen (z.B. Greenhoff; Macfie 1991). Diese graphische Darstellung basiert auf der Hauptkomponentenanalyse und wird als Internal Preference Mapping bezeichnet (mir ist nicht bekannt, ob es einen deutschen Ausdruck dafür gibt). Die Bilder stellen die beste zweidimensionale Näherung der vergebenen Ränge dar. Für jedes Produkt ergibt sich ein Punkt, jede Prüfperson erhält einen Pfeil. Dabei können die Ränge dieser Prüfperson durch die Pfeile (angenähert) rekonstruiert werden. Wenn man in der Richtung des Pfeiles einer Prüfperson wandert, kommt man zuerst an dem Produkt vorbei, das von dieser Prüfperson den Rang 1 erhalten hat, dann an dem mit Rang 2, und so weiter.

Abb. 5 zeigt das Preference Mapping mit den Studenten der Universität Dortmund, Abb. 6 das Preference Mapping mit den Schülern aus England. Dabei

wurden die Hauptkomponenten und die Pfeile aus allen 109 Urteilen berechnet. Wir teilen die Gruppen aus Gründen der Übersichtlichkeit auf zwei Bilder auf. Vergleicht man die beiden Abbildungen miteinander, so bestätigt sich der bereits formal nachgewiesene Unterschied im Benotungsverhalten der beiden Gruppen: Die Pfeile der Studenten, die vorwiegend in der oberen rechten Hälfte der Graphik zu finden sind, deuten auf eine Bevorzugung der Produkte Classic Cola, Coca Cola sowie Pepsi Cola hin. Demgegenüber zeigt die Ausrichtung der Pfeile der englischen Schüler eine Bevorzugung der Produkte Cola-Hartinger / Wesergold, River Cola sowie Coca Cola an.

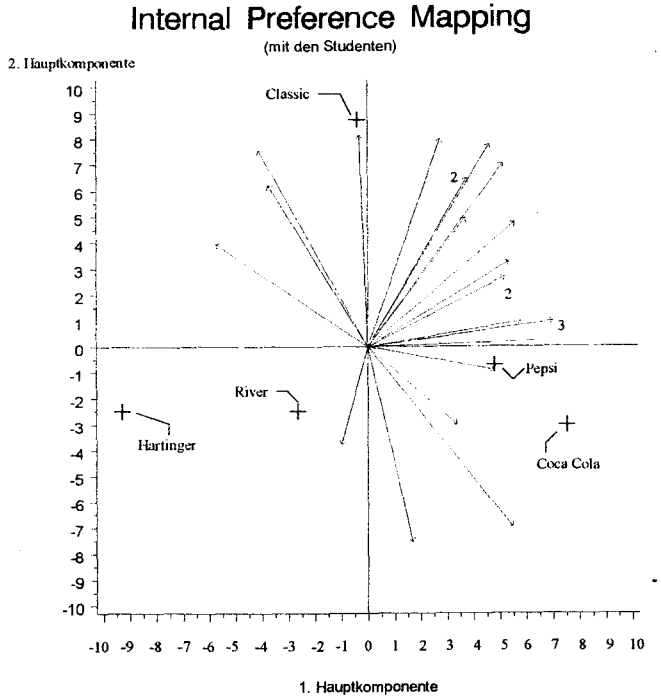

Abb. 5.

4. Abschließende Bemerkungen

Die Studie wollte nichts über die tatsächliche Qualität von Erfrischungsgetränken oder die tatsächlichen Bevorzugungen von Konsumenten aussagen. Dazu hatten wir zu viele methodische Mängel:

Die Studie fand im Rahmen einer Projektwoche an einem Gymnasium statt. Es sollte den Schülern Gelegenheit gegeben werden, statistische Methoden zu üben, die man für Untersuchungen des Geschmacks von Nahrungsmitteln braucht. Daher konnten die Schüler selbst als Versuchsleiter aktiv werden. Dies bedeutete aber, daß nicht alles ganz genauso ablief, wie es von Seiten der Kursleiter gedacht war.

Dennoch sind die Ergebnisse genau so, wie man sie wohl erwartet hätte:

Die beiden großen Marken wurden gegenüber den No-Name Produkten bevorzugt.

Eine eindeutige Bevorzugung einer der beiden großen Marken gab es nicht.

Es gab Unterschiede zwischen den einzelnen Ländern, insbesondere die Engländer hatten ihren eigenen Geschmack. Leider hatten wir nicht sehr viele Engländer (und Franzosen) in der Studie.

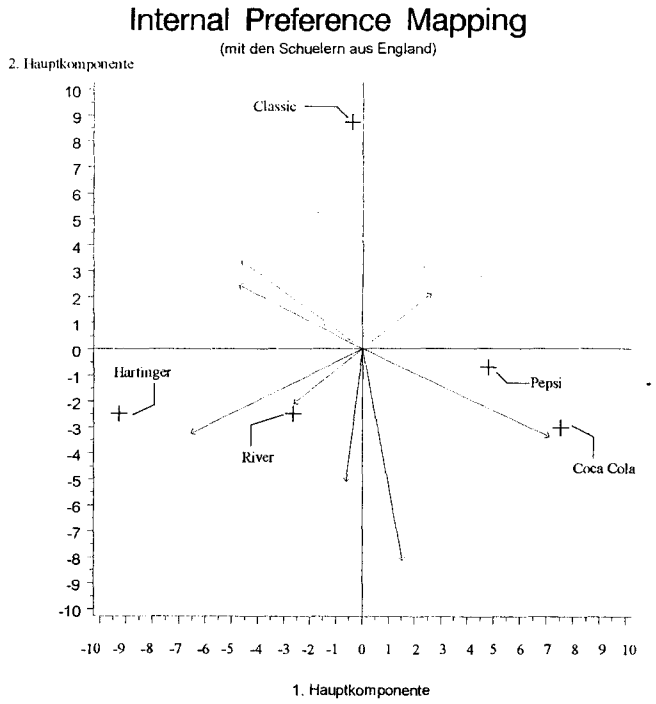

Abb. 6.

Danksagung

Die Studie wurde von der Firma Coca Cola durch Bereitstellung von Gläsern und der von ihr hergestellten Getränke unterstützt. Dafür danken wir insbesondere Herrn Grützner von der Firma Coca Cola sehr herzlich!

In Anbetracht des Ergebnisses der Studie möchte ich aber noch einmal ausdrücklich betonen, daß die Prüfpersonen zwar wußten, daß wir von Coca Cola unterstützt wurden, daß sie aber durch die Anlage des Versuches als Blindversuch nicht wissen konnten, welches Getränk es jeweils war, das sie bevorzugten!

Literatur

Aström, A. (1993). Sensory Analysis to Evaluate Flavour Transfer in Refillable Pet Bottles. FLAIR SENS meeting: Practical Models for Relationships between Data Sets, Vol. 2, No 3

Büning, H.; Trenkler,G. (1994). Nichtparametrische statistische Methoden. De Gruyter Verlag, Berlin, New York, (2. Auflage)

Greenhoff, K.; Macfie, H. J. (1991). Practical applications of preference mapping. Bericht im Tagungsband: Agro-Industrie et Methodes Statistiques, Nantes, 13./14. Juni 1991, S. 57-78

Hartung, J.; Elpelt, B.; Klösener, K.-H. (1985). Lehr- und Handbuch der angewandten Statistik. R. Oldenbourg Verlag, München, Wien (4. Auflage)

International Standard ISO 4120 (1983). Sensory Analysis - Methodology - Triangular Test

International Standard ISO 8587 (1988). Sensory Analysis - Methodology - Ranking

"Oft mehr Werbung als Geschmack". Artikel in der Zeitschrift Test 6/91, S. 602-605

Rusert, M. (1994). Statistische Methoden zur Ermittlung der Wahrnehmbarkeitsschwelle bei Farbunterschieden. Diplomarbeit am Fachbereich Statistik der Universität Dortmund

Joachim Kunert, Frank Lehmkuhl, Anja Schleppe

Fachbereich Statistik

Universität Dortmund

D-44221 Dortmund

Anhang

Abb. 1: Arbeitsbogen Dreieckstest

Arbeitsbogen Dreieckstest

Bogen Nr.

Mit diesem Arbeitsbogen sollen die Produkte sowie deren Anordnung und die zugehörigen Kennziffern ausgewürfelt werden.

O **(Produkt/Kennziffer)** O **(Produkt/Kennziffer)**

O **(Produkt/Kennziffer)**

Abb. 2: Prüfbogen Dreieckstest

Prüfbogen Dreieckstest

Bogen Nr.

Name der Prüfperson:

Schule:

Probiere die drei Getränke in einer beliebigen Reihenfolge. Bestimme die Probe, die sich Deiner Meinung nach von den beiden anderen unterscheidet. Kreuze deren Kennziffer auf diesem Prüfbogen an. Falls notwendig, probiere die Getränke erneut.

O **(Kennziffer)** O **(Kennziffer)**

O **(Kennziffer)**

Die getroffene Entscheidung ist O **richtig**

O **falsch**

Abb. 3: Arbeitsbogen Rangordnungsprüfung

Arbeitsbogen Rangordnungsprüfung

Bogen Nr.

Wir wollen 5 Cola-Sorten miteinander vergleichen. Wir müssen zunächst die Anordung auswürfeln. Dazu dienen die Ziffern hinter den Produkten. Anschließend müssen wir für jede Probe eine zweistellige Zufallszahl auswürfeln.

Classic Cola (1)
Coca Cola (2)
Hartinger Cola (3)
Pepsi Cola (4)
River Cola (5)

Produkt					
Kennziffer					

Zur späteren Auswertung werden nach dem Versuch die Ergebnisse festgehalten, indem den einzelnen Produkten ein Rang zugeordnet wird.

Rang					

(Bestes Produkt erhält Rang 5, schlechtestes Produkt erhält Rang 1)

Abb. 4: Prüfbogen Rangordnungsprüfung

Prüfbogen Rangordnungsprüfung

Bogen Nr.

Name der Prüfperson:

Schule:

Probiere zuerst die 5 Getränke in der vorgegebenen Reihenfolge von links nach rechts. Ordne die Getränke so um, daß das am meisten geschmacklich favorisierte Getränk ganz links steht und die anderen Getränke in absteigender Reihenfolge sortiert sind. Falls notwendig, probiere die Getränke erneut und ändere gegebenenfalls die Anordnung. Schreibe die endgültige Reihenfolge in die unten angegebene Tabelle.

am meisten favorisiert				am wenigsten favorisiert