Ich erkenne das Geschlecht an der Handschrift, denn Mädchen schreiben schöner als Jungen. Blankes Vorurteil oder bittere Realität?

WOLFGANG RIEMER, KÖLN

Zusammenfassung: *Allzu oft beschränkt sich Stochastik-Unterricht auf das Lösen eingekleideter Aufgaben. Der Beitrag stellt ein kurzweiliges Experiment vor, das im Gegensatz dazu in einer Doppelstunde spannende Fragen aufwirft, die sich mit gängigen statistischen Methoden beantworten lassen. Dabei ist es Geschmackssache, ob man die Thematik an den Anfang stellt und daran Theorie entwickelt oder anders herum als Anwendung zuvor erarbeiteter Theorie - etwa zur Prüfungsvorbereitung - nutzt.*

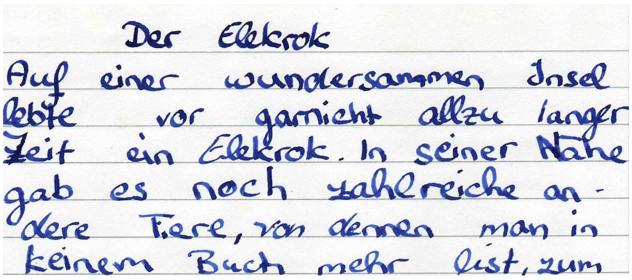
1 Didaktischer Hintergrund

Beurteilende Statistik mit dem Testen von Hypothesen und dem Schätzen von Parametern über Konfidenzintervalle gilt als krönender Abschluss des Stochastikunterrichts. Weil diese Themen zum Bestandteil zentraler Prüfungen gehören, werden sie bis zur Routine eingeschliffen. Dabei kommen oft eingekleidete (mitunter recht fragwürdige) Kontexte zum Einsatz. Authentische Fragestellungen mit selbst formulierten Hypothesen, die mit Hilfe realer Daten geprüft/getestet werden, sind - allen lernpsychologischen Paradigmen zum Trotz - im schulischen Alltag die Ausnahme. Das hier diskutierte Beispiel gestattet, in motivierender Weise viele Aspekte beurteilender Statistik bis hin zu Grundgedanken Bayesscher Statistik so lebendig werden zu lassen, dass man sich auch lange nach der Schulzeit gerne daran erinnert.

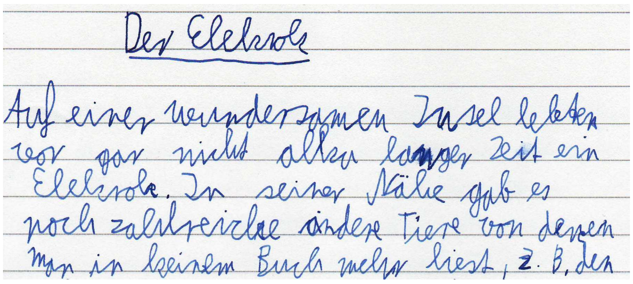
2 Fragestellung

Man schaut auf eine liegen gebliebene Heftseite und ist sich sicher: Das kann nur ein Mädchen geschrieben haben. Aber wie sicher kann man sich da sein? Und schreiben Mädchen wirklich so viel schöner als Jungen? Spannende Fragen, bei deren Beantwortung statistische Werkzeuge helfen können.

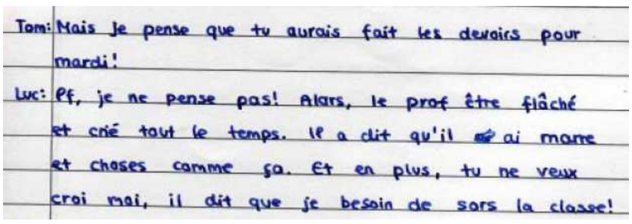
Wie in Abschnitt 5 gezeigt wird, kann man mit *Beschreibender Statistik* die Vorurteile, die wir über die Geschlechterabhängigkeit von Schriftbildern in uns tragen, sichtbar machen und auch, ob die Vorurteile zu Recht bestehen. Mit *Beurteilender Statistik* lässt sich anschließend prüfen, ob man bei der Geschlechtserkennung aus dem Schriftbild signifikant besser ist als jemand, der „blind rät“ (einseitiger Test) oder ob man die eigene Treffsicherheit akzeptabel einschätzt (zweiseitiger Test). Im Prinzip sind das die Standard-Testprobleme, die man üblicherweise im Zusammenhang mit „gezinkten Münzen“ thematisiert. Aber hier ist der Kontext ungleich interessanter. Das lohnt den bescheidenen Mehraufwand.


Der Elekrok
Auf einer wundersammen Insel
lebte vor garnicht allzu langer
Zeit ein Elekrok. In seiner Nähe
gab es noch zahlreiche an-
dere Tiere, von denen man in
keinem Buch mehr list, zum

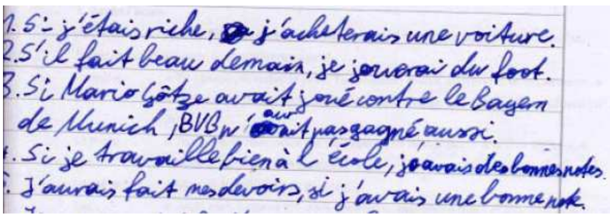
70% vermuten: Mädchen - mittlere Sicherheit 74%


Der Elekrok
Auf einer wundersamen Insel lebten
vor gar nicht allzu langer Zeit ein
Elekrok. In seiner Nähe gab es
noch zahlreiche andere Tiere von denen
man in keinem Buch mehr liest, z. B. den

65% vermuten: Junge - mittlere Sicherheit 69%


Tom: Mais je pense que tu aurais fait les devoirs pour
mardi!
Luc: Pf, je ne pense pas! Alors, le prof être fâché
et crié tout le temps. Il a dit qu'il ai marre
et choses comme ça. Et en plus, tu ne veux
croi moi, il dit que je besoin de sors la classe!

75% vermuten: Mädchen - mittlere Sicherheit 80%


1. S: j'étais riche, j'achèterais une voiture.
2. S'il fait beau demain, je jouerai du foot.
3. Si Mario Götze avait joué contre le Bayern
de Munich, BVB n'aurait pas gagné aussi.
4. Si je travaille bien à l'école, j'aurais des bonnes notes.
5. J'aurais fait mes devoirs, si j'avais une bonne note.

100% vermuten: Junge - mittlere Sicherheit 88%

Abb. 1a: typisch Mädchen?

Abb. 1b: typisch Junge?

3 Vorbereitung

Der Versuchsleiter fotografiert oder scannt einen Ausschnitt aus einer Klassenarbeit wie in Abb. 1 und erstellt daraus eine Schriftproben-Präsentation und einen Fragebogen mit abgeknickter Lösungsspalte, auf der das tatsächliche Geschlecht eines jeden Autors markiert ist (vgl. Abb. 2, hier sind die drei ersten der insgesamt 30 Schriftproben zu bewerten). Bevor es losgeht, schätzt man im Plenum, mit welcher Wahrscheinlichkeit p* man glaubt, das Geschlecht nur am Schriftbild erkennen zu können. Damit hat man eine Hypothese aufgestellt. Die Einschätzung p* hängt von der Klassenstufe ab, weil Handschriften in Klasse 5 noch nicht so ausgeschrieben sind wie in höheren Jahrgangsstufen. Eine Gruppe aus 36 Bewertern (Schülerinnen und Schüler einer Klasse 11) schätzte für ein Diktat in Klasse 5 p* = 0,75.

Lösung			Bewertungsbogen										
Geschlecht (tatsächlich)		Treffer	Probe	Geschlecht (vermutet)		Sicherheit				Note			
Mädchen	Junge			Mädchen	Junge	50%	65%	80%	95%	1	2	3	4
x			1										
x			2										
	x		3										

Abb. 2: Die ersten drei Zeilen eines Bewertungsbogens mit so vielen Zeilen, wie es Textproben gibt.

4 Versuchsdurchführung

Der Versuchsleiter zeigt die Schriftproben in einem Zeittakt von ca. 10 Sekunden.
Derweil notieren die Versuchsteilnehmer gemäß Abb. 2

a) das vermutete Geschlecht
b) die subjektive Sicherheit bei ihrer Entscheidung. (Hier werden als Entscheidungshilfe Vorgaben zum Ankreuzen gemacht, sonst quantifizieren Schüler große Unsicherheit durch Trefferquoten in der Nähe von 0% , nicht 50%)
c) eine Note (zwischen 1 und 4) für das *Schriftbild*, nicht den Inhalt.

Bei 30 Klassenarbeiten dauert die Datenerhebung ca. 6-8 Minuten.

5 Beschreibende Statistik

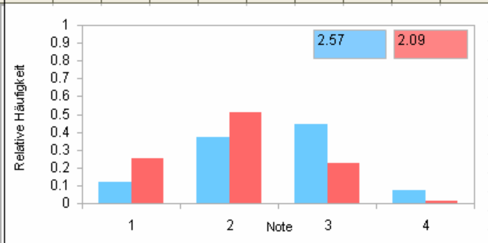
Die Auswertung organisiert man wie folgt:
a) *Vorurteile*
Noch *bevor* das tatsächliche Geschlecht bekannt gegeben wird, bestimmt jeder Bewerter seinen Notenspiegel samt Mittelwert für die Schriftbilder, hinter denen er
- Jungen vermutet (vgl. Abb. 3 Spalten B bis F)
- Mädchen vermutet (Abb. 3, Spalten G bis K).

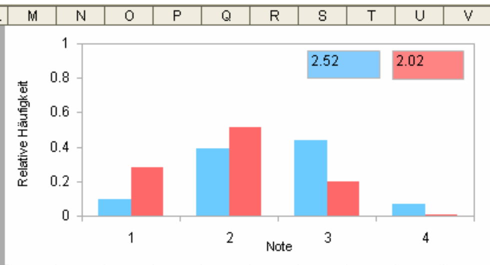
Falls die vermuteten Mädchen deutlich besser bewertet werden (Abb. 3, Spalte K und F) *glaubt* der betreffende Bewerter fest dran, dass Mädchen schöner schreiben. Damit wird ein Vorurteil quantitativ sichtbar, denn ein Urteil würde ja Kenntnis des Geschlechts voraussetzen.
b) *Urteile*
Das Geschlecht eines jeden Autors wird (z. B. durch Aufdecken der Lösungsspalte in Abb. 2) bekannt gegeben. Analog zu a) bestimmt man nun die Notenspiegel für die
- *tatsächlichen* Jungen (Abb. 3 Spalten M bis Q)
- *tatsächlichen M*ädchen (Abb. 3 Spalten R bis V).

c) *Trefferquoten*
anschließend ermittelt jeder *für sich selber*
- die mittlere Sicherheit p**, mit der er während der Versuchsdurchführung glaubte, das Geschlecht zu erkennen (Abb. 3 Spalte X)
- die Trefferquote h, mit der er das Geschlecht richtig erkannt hat (Abb. 3 Spalte Y).

6 Ergebnisse

Wie man aus den kumulierten Daten in Abb. 3 Zeile 16 entnimmt, schreiben die vermuteten Jungen eine halbe Notenstufe schlechter als die vermeintlichen Mädchen (2,57 gegen 2,09). In dieser Größenordnung liegen auch die Unterschiede zwischen den tatsächlichen Jungen und den tatsächlichen Mädchen (2,52 gegen 2,02). Kurz:
a) Das Vorurteil: Jungen schreiben schlechter als Mädchen ist deutlich ausgeprägt
b) Das Vorurteil entspricht den Tatsachen
c) Die Testteilnehmer glaubten, das Geschlecht mit durchschnittlich 74%iger Sicherheit erkennen zu können (Abb. 3 Zelle X16). Tatsächlich lag die Trefferquote bei „nur“ 66% (Abb. 3 Zelle Y16).
Die Angaben beziehen sich auf 20 Bewertungen von 30 Schriftproben einer Klasse 5. Untersuchungen der Schriftbilder in Matheklausuren (Kl. 11) und von Französisch-Arbeiten in Klasse 9 lieferten ähnliche Ergebnisse, wobei deutlich höhere Trefferquoten bei der Geschlechtserkennung nicht ungewöhnlich sind.





	Vermutete Jungen					Vermutete Mädchen						Tatsächliche Jungen					Tatsächliche Mädchen						Vermutete Sicherheit	Tatsächliche Trefferquote
	1	2	3	4	μ	1	2	3	4	μ		1	2	3	4	μ	1	2	3	4	μ		74%	66%
	46	200	278	48	2.57	99	271	123	11	2.09		45	187	245	31	2.52	113	276	104	9	2.02		74%	66%
1	0	17	2	0	2.11	1	9	1	0	2.00	1	0	13	2	0	2.13	1	13	1	0	2.00	1	67%	73%
2	1	7	7	0	2.40	3	11	1	0	1.87	2	1	10	4	0	2.20	3	9	3	0	2.00	2	77%	60%
3	4	5	6	0	2.13	6	5	2	0	1.69	3	3	7	5	0	2.13	8	4	3	0	1.67	3	72%	53%
4	0	6	9	3	2.83	0	4	8	0	2.67	4	0	3	9	3	3.00	0	7	8	0	2.53	4	71%	43%
5	6	7	4	0	1.88	7	3	3	0	1.69	13	3	6	6	0	2.20	10	4	1	0	1.40	13	74%	60%
6	1	6	5	1	2.46	9	5	3	0	1.65	14	2	7	5	1	2.33	8	4	3	0	1.67	14	77%	80%

Abb. 3: Auswertung von 30 Diktaten einer Klasse 5. Die Notenspiegel von 6 Bewertern stehen in den Zeilen 17-22. Die kumulierten Verteilungen sind Zeile 16 zu entnehmen. Aus dieser Zeile schätzt man auch die bedingten Wahrscheinlichkeiten, mit denen Jungen/Mädchen die Noten 1 bis 4 erhalten (vgl. Abschnitt 8).

7 Beurteilende Statistik

Nach der Überzeugung des Autors ist die Arbeit mit Konfidenzintervallen für Lernende erhellender als das Testen von Hypothesen, weil sich der Einfluss des Stichprobenumfangs an der Länge der Konfidenzintervalle unmittelbar ablesen lässt. Im Gegensatz dazu fordern viele zentrale Prüfungen (aus zufälligen historischen Gründen) Vertrautheit mit rechtsseitigen, linksseitigen und zweiseitigen Hypothesentests. Nur deswegen (nicht aus innerer Überzeugung) soll an drei Aufgabenstellungen demonstriert werden, wie man die Daten des Schriftbild-Experiments nutzen kann, um „alle Spielarten“ des Hypothesentestens bei Bernoulliketten abzudecken.

Aufgabe:
p bezeichne die unbekannte Wahrscheinlichkeit, mit der Sie das Geschlecht am Schriftbild erkennen. Testen Sie für sich selber:

a) Ich kann H_o: p = 0,5 gegen H_1 p>0,5 verwerfen (rechtsseitiger Test der Hypothese „Ich rate nur“).
b) Ich kann H_o: p=p* (0,75) gegen die Alternative H_1 p<p* verwerfen (linksseitiger Test der Hypothese „Wir haben uns vorab im Plenum nicht überschätzt“).
c) Ich kann H_o: p = p** gegen H_1 p ≠ p** verwerfen (zweiseitiger Test der Hypothese: „Ich habe meine subjektive Sicherheit während der Testdurchführung richtig eingeschätzt“).
d) Bestimmen Sie das 95%-Konfidenzintervall für Ihre Trefferwahrscheinlichkeit.

Abb. 4 vermittelt einen Eindruck von den zu erwartenden Ergebnissen (0 bedeutet Hypothese kann nicht verworfen werden). So schätzte der sechste Bewerter (Zelle B21) seine Treffsicherheit auf p**=74% ein. Tatsächlich erreichte er die Trefferquote 80%. Damit kann er H_o: p = 0,5 gegen H_1 p>0,5 verwerfen (1 in Zelle D21) und braucht weder H_o: p=p* gegen die Alternative H_1 p<p* noch H_o: p = p** gegen H_1 p ≠ p** zu verwerfen.

	Vermutete Sicherheit p**	Tatsächliche Trefferquote	p=0,5 gegen p>0,5 verwerfen?	p=0,75 gegen p<p* (=0,75) verwerfen?	p=p** gegen p<>p** verwerfen?
μ->	72%	64%	54%	29%	26%
1	67%	73%	1	0	0
2	77%	60%	0	1	1
3	72%	53%	0	1	1
4	71%	43%	0	1	1
5	74%	60%	0	1	0
6	77%	80%	1	0	0
7	76%	57%	0	1	1
8	74%	80%	1	0	0
9	74%	53%	0	1	1
10	70%	73%	1	0	0
11	64%	65%	1	0	0
12	77%	53%	0	1	1
13	82%	80%	1	0	0
14	73%	67%	1	0	0
15	71%	67%	1	0	0
16	71%	63%	0	0	0
17	77%	67%	1	0	0
18	72%	70%	1	0	0
19	67%	53%	0	1	0
20	72%	67%	1	0	0
21	71%	50%	0	1	1
22	79%	67%	1	0	0
23	79%	63%	0	0	1
24	79%	77%	1	0	0
25	81%	72%	1	0	0
26	77%	70%	1	0	0
27	80%	63%	0	0	1
28	77%	83%	1	0	0
29	75%	70%	1	0	0
30	78%				
31	69%	63%	0	0	0
32	75%	63%	0	0	0
33	71%	80%	1	0	0
34	75%	60%	0	1	0
35	65%	67%	1	0	0
36	72%	70%	1	0	0

Abb. 4: Testergebnisse. 30 Diktate, Klasse 5 wurden von 36 Personen bewertet.

8 Bayessche Statistik

Obwohl es momentan nicht zum Standardrepertoire gehört, kann man das Handschriftexperiment hervorragend nutzen, um einen Blick in die Grundgedanken Bayesscher Statistik zu werfen. Dabei geht es nicht mehr um das Testen von Hypothesen, sondern um Wahrscheinlichkeiten, die man Hypothesen in Abhängigkeit von hinzukommenden Informationen beimisst.
Objektivisten lehnen eine solche Sichtweise ab, weil in ihrem Weltbild Hypothesen entweder wahr oder falsch sind und nicht mit subjektiven Wahrscheinlichkeiten bewertet werden. Hypothesen sind keine Zufallsgrößen. Das entspricht aber nicht wirklich der Alltagssicht, denn täglich erleben wir, dass Menschen in Situationen subjektiver Unsicherheit „Experten" befragen und sich nach deren Urteil ausrichten. Hinzukommende Informationen/Expertisen ändern unsere Bewertungen, sie machen für uns manche Alternativen wahrscheinlicher, andere unwahrscheinlicher.
Diesen Prozess des „Erkenntnisgewinns" kann man gemäß Abb. 5 im Handschriften-Experiment mit Hilfe der Regel von Bayes qualitativ verstehen: A priori, d. h. ohne Vorinformation glauben wir bei einem Schriftstück je mit 50% an einen Jungen bzw. ein Mädchen. Wir erfahren, dass der erste „Experte" das Schriftbild mit 1 bewertet hat. Das spricht sehr für ein Mädchen, denn bei Mädchen tritt die 1 mit 23%, bei Jungen nur mit 9% auf. Die Wahrscheinlichkeit für Mädchen steigt nach der Bayesschen Regel auf 71,8%. Zur Sicherheit fragen wir weitere Bewertungen ab. Abb. 5 zeigt, wie sich unsere Einschätzung verändert, wenn der zweite und der dritte „Experte" je eine 2 vergeben, der dritte eine 4 und der vierte wieder eine 1. Nach den ersten drei Urteilen glauben wir mit 85% Sicherheit an ein Mädchen. Die Note 4 verunsichert uns (62,5%), aber die folgende 1 macht uns wieder sicherer (80,9%). Würden wir uns nach 5 Schritten auf „Mädchen" festlegen, würden wir die subjektive Irrtumswahrscheinlichkeit 19,1% in Kauf nehmen, denn wir glauben ja nur mit 80,9% Sicherheit an ein Mädchen. In der Regel führt man diesen Prozess solange fort, bis eine der Hypothesen eine sehr hohe Wahrscheinlichkeit bekommt, man sich also sehr, sehr sicher ist.

	A	B	C	D
1			Junge	Mädchen
2	0		50.0%	50.0%
3	1	1	28.2%	71.8%
4	2	2	20.9%	79.1%
5	3	2	15.0%	85.0%
6	4	4	37.5%	62.5%
7	5	1	19.1%	80.9%

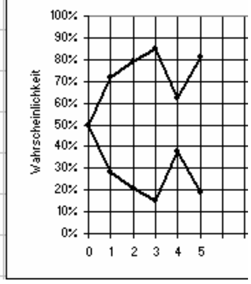


Abb. 5: Durch „Expertenurteile" (Spalte B) ändern sich die Wahrscheinlichkeiten, die man den beiden konkurrierenden Hypothesen (Junge/Mädchen) beimisst. Man erhält eine neue Tabellenzeile aus der darüber stehenden durch wiederholte Anwendung der Regel von Bayes. Z. B. ergibt sich die Wahrscheinlichkeit der Hypothese „Mädchen" in der Zeile 2 (Zelle D4) aus der darüber stehenden Zeile nach der Bayesschen Regel so:

$$p_2(M) = \frac{0{,}718 \cdot 0{,}55}{0{,}282 \cdot 0{,}37 + 0{,}718 \cdot 0{,}55} \approx 0{,}791$$

Der Nenner gibt an, mit welcher Wahrscheinlichkeit man die Note 2 nach der Pfadregel erwartet, der Bruch ist der Anteil, den der Pfad über „Mädchen" an dieser Wahrscheinlichkeit besitzt.

Resümee

Eine verantwortungsvolle Abiturvorbereitung erfordert das Einschleifen von Algorithmen. Darüber sollte man aber nicht vergessen, dass Stochastikunterricht von authentischen Fragestellungen lebt, die beim Sammeln von realen Daten einen erheblichen Einfluss auf Motivation und Nachhaltigkeit besitzen. Die hier entwickelte Fragestellung ist so nahe liegend und besitzt so viele Facetten, dass sie eine Vielzahl statistischer Aspekte lebendig werden lässt. Die investierte Doppelstunde zahlt sich in jedem Fall aus.

Anschrift des Verfassers

Dr. Wolfgang Riemer
August-Bebel-Str. 80
50259 Pulheim
w.riemer@arcor.de
www.riemer-koeln.de