

Statistik mit Red Bull

Pharmaforschung im Klassenraum

LERNGRUPPE: 7.–8. Schuljahr (beschreibende) oder 12. Schuljahr (beurteilende) Statistik

IDEE: Wir prüfen: Verbessern Energy-Drinks das Reaktionsvermögen?

ONLINE-MATERIAL: Excel-Programme zur Messung und Auswertung von Reaktionszeiten

ZEITBEDARF: ca. 4 Stunden, davon 2 Stunden Experiment, 2 Stunden Theorie

In Statistikbüchern „wimmelt“ es von Aufgaben, in denen mithilfe fingierter Daten Medikamente oder medizinische Behandlungsmethoden auf ihre Wirksamkeit hin „untersucht“ werden sollen.

Ich möchte Sie ermutigen, so eine Untersuchung tatsächlich einmal mit Ihrer Klasse durchzuführen. Sie brauchen eine Doppelstunde, eine Kiste Energy-Drink, vielleicht auch eine Packung Traubenzucker ... und Statistik bekommt – auch Jahre nach dem Ende der Schulzeit – einen ungewohnt süßen, sinnvollen Nachgeschmack.

Welche statistischen Methoden man bei der Auswertung einsetzt (oder noch

besser: an diesem Beispiel in sinnstiftendem Kontext erarbeitet), hängt von der Jahrgangsstufe ab, in der man experimentiert. Lassen Sie die Dosen „knacken“ und stürzen Sie sich in die Statistik der Pharmaforschung!

Von der Idee zum Experiment

Die Initiative „Hart am Limit“ (HALT, <http://www.halt-projekt.de>) gegen das Alkohol-Kampfrinken hat sich bundesweit einen Namen gemacht. Warum sollte man nicht auch im Mathematikunterricht den negativen Einfluss von Alkohol auf die Reaktionsfähigkeit untersuchen? Schließlich sind Bremsweg und Reaktionszeiten (in Abhängigkeit von der Geschwindigkeit) ein beliebtes Thema in den Klassen 8 und 9. Doch stößt der Alkoholkonsum in der Schulzeit auf nahe liegende Schwierigkeiten, auch wenn er im Dienste der Wissenschaft erfolgen würde. So wurde die Idee geboren, das genaue Gegenteil zu untersuchen: Doping mit Energy-Drinks. Denn offensichtlich glau-

ben nicht nur Formell-Piloten sondern auch Sportler und Klausuren schreibende Schüler an die beflügelnde Wirkung von Koffeinbräusen.

Die Versuchsdurchführung ist einfach. Eine Doppelstunde (**Abb. 1**) reicht. Wenn man auf die Untersuchung der Langzeitwirkung verzichtet, kommt man mit einer Einzelstunde aus.

Bitten Sie Ihre Schülerinnen und Schüler, morgens keine Muntermacher zu konsumieren und in „nüchternem“ Zustand ihre Reaktionszeiten zu messen, indem sie 200 Mal mit der Maus auf ein zufällig am Bildschirm auftauchenden Button klicken. Das dauert 10 Minuten. Jeder speichert seine 200 Daten ab. Um anonym zu bleiben, verwendet man am besten einen Codenamen wie den Vornamen der Oma. Die Datei heißt dann z. B. Christa-1.xls.

Nun bekommt jeder Versuchsteilnehmer zwei Esslöffel Traubenzucker und spült das Pulver mit einer Dose Energiebräuse hinunter. Nach 15 Minuten wird die zweite Messreihe aufgenommen und abgespeichert (Christa-2.xls). Nach weiteren 20 Minuten, in denen man die zuvor aufgenommenen Daten inspizieren kann, folgt die dritte Messreihe (Christa-3.xls) zur Untersuchung der „Langzeit-Wirkung“.

Beschreibende Statistik

Wenn man das Doping-Experiment im Rahmen eines Kurses zur beschreibenden Statistik einsetzen und das Auswerten/Visualisieren/Vergleichen der Reaktionszeiten zum zentralen Lerngegenstand machen möchte, bekommen die Lernenden ihre Reaktionszeiten nur als Zahlenlisten. (abgespeckte Variante: Datei reaktion-o.xls).

Die Berechnung statistischer Kennwerte, das Visualisieren und das Be-

08.10 Uhr
die Computer sind hochgefahren und startklar für den ersten Reaktionstest ...

200 Daten speichern (christa-1.xls)

08.25 Uhr
zwei Esslöffel Traubenzucker mit Energiebräuse herunterspülen ...
Wirkung abwarten, dabei Daten inspizieren

08.40 Uhr
zweiter Reaktionstest, 200 Daten speichern (christa-2.xls)

09.10 Uhr
dritter Reaktionstest („Langzeitwirkung“) 200 Daten speichern (christa-3.xls)

09.30 Uhr
Computer sind heruntergefahren



Foto: Matthias Schiller

Abb. 1: Ablauf der Doppelstunde zur Datenerhebung

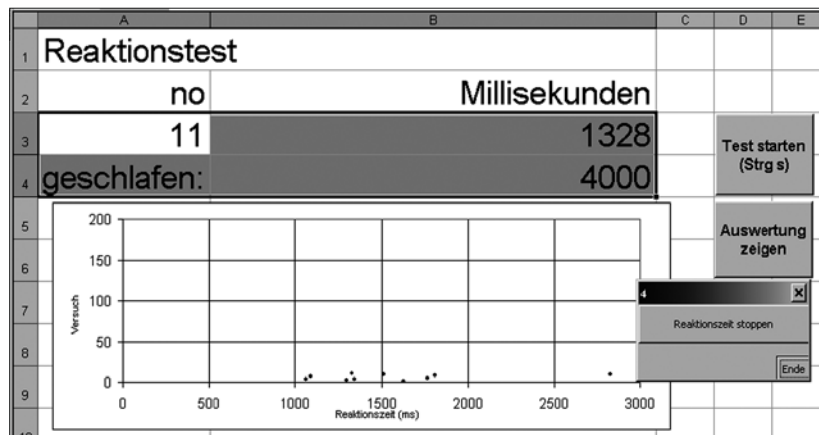


Abb. 2: Beim Reaktionszeiten-Messprogramm (Excel) muss man mit der Maus möglichst schnell auf den an einer Zufallsposition erscheinenden Knopf „Reaktionszeit stoppen“ klicken. Nach 200 Versuchen beendet man das Experiment durch Klick auf „Ende“ und erhält eine erste statistische Auswertung mit Kenngrößen, Boxplot, Histogramm und Trendanalyse

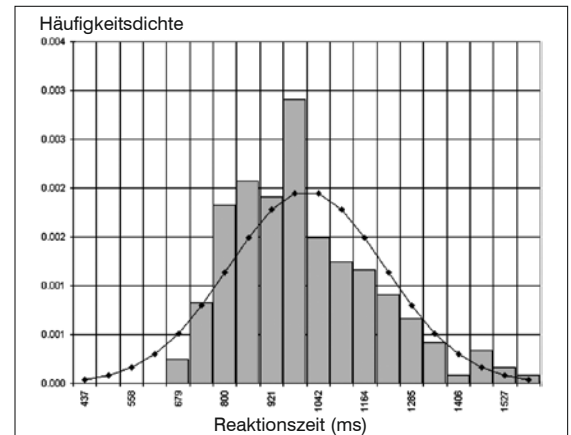
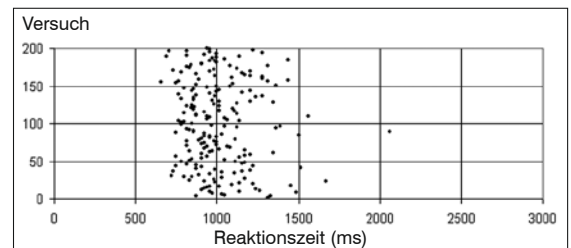


Abb. 3: Während des Tests werden die Reaktionszeiten als Punkte einer Wolke, die nach oben hin wächst, sichtbar und am Ende in der Auswertung als Histogramm (und auch als Boxplot) präsentiert

werten/Präsentieren der Ergebnisse ist für Mittelstufenschüler eine Herausforderung, in der alle Werkzeuge der beschreibenden Statistik zum Einsatz kommen. Selbstverständlich kann man kleine Reaktionsdatensätze vorab auch nutzen, um Schülerinnen und Schüler an Kennwerte und graphische Aufbereitungen (händisch rechnend) heranzuführen.

Wenn man die Wirksamkeit des Doping mit beurteilender Statistik untersuchen möchte, wird man die Programmversion reaktion.xls einsetzen, welche die Aufbereitung der Daten abnimmt.

Auswertung der Daten

Mit welchen Ergebnissen man rechnen kann, zeigt **Abb. 3**. Das arithmetische Mittel der Reaktionszeit sinkt bei „Christa“ nach dem Doping von 1012 ms auf 863 ms (Verbesserung um 15%) und nach der 20-minütigen Wartepause weiter auf 820 ms (Verbesserung um weitere 5%). Die Verkürzung der Reaktionszeit ist auch am Absinken des Medians abzulesen, wobei der Median stets unter dem arithmetischen Mittel liegt. Das liegt daran, dass die glockenförmige Verteilung der Reaktionszeiten Ausreißer nach oben („Schlafmütze“), aber nicht nach unten haben kann.

Der Median ist im Gegensatz zum arithmetischen Mittel Ausreißern ge-

genüber unempfindlich. Diese leichte Unsymmetrie spiegelt sich auch in den Boxplots wider.

Nur ein Übeeffect?

Pfiffige Schüler kommen beim Vergleich der Ergebnisse auf die Idee, dass die Verbesserung der Reaktion nur ein Übungseffekt sein könnte, also nicht notwendig auf der „medizinischen“ Wirkung des Doping beruht. Wenn man sie auffordert, diese Hypothese an den vorliegenden Daten zu prüfen, machen sie in der Regel den Vorschlag, die Serien aus 200 Reaktionszeiten jeweils in zwei 100er Serien zu zerlegen. (Christa-1.xls wird in Christa-1a.xls und Christa-1b.xls zerlegt). Man erhält dann drei Zwillingssparchen, deren Mittelwerte **Abb. 4a** (s. S. 56) veranschaulicht.

Bei Vorliegen eines Übungseffektes müssten die Mittelwerte der jeweils ersten 100 Reaktionszeiten deutlich unter denjenigen der ersten 100 Reaktionszeiten liegen. Das ist aber nicht der Fall, wie **Abb. 4b** und noch deutlicher **Abb. 5** (S. 56, 15 Schüler zusammen) zeigen. Die Auswertungsvorlage.xls nimmt die Zerlegung in 100er-Serien und weitere Berechnungsschritte ab.

Das Ergebnis

Durch Doping mit zwei Esslöffeln DextroPur und einer Dose RedBull

verbessert sich die Reaktionszeit eines Jugendlichen um durchschnittlich ca. 10%. Diese Verbesserung beruht nicht auf einem Übungseffekt.

Beurteilende Statistik

In der Qualifikationsphase wird man bei der „Doping-Studie“ die Standard-Werkzeuge beurteilender Statistik erproben, vielleicht auch „neue“ kennen lernen – wie

- das Konfidenzintervall für den Median (wodurch endlich einmal eine Brücke geschlagen wird zwischen „beschreibender Boxplot-Statistik“ und beurteilender Statistik),
- den Mood'schen Median-Test oder
- den t-Test.

Die Tatsache, dass der t-Test in Tabellenkalkulationsprogrammen und graphischen Taschenrechnern (meist unentdeckt) schlummert, macht ihn für Referate oder Facharbeiten höchst attraktiv (vgl. Sachs 1999, S. 351 Abschnitt 36).

Vorzeichentest: einfacher geht's nicht

Mit dem Vorzeichentest prüft man, ob zwei Datensätzen die gleiche Wahrscheinlichkeitsverteilung zugrunde liegt wie folgt: Man subtrahiert den Messwert des zweiten (Christa-3) von dem entsprechenden Messwert des ers-

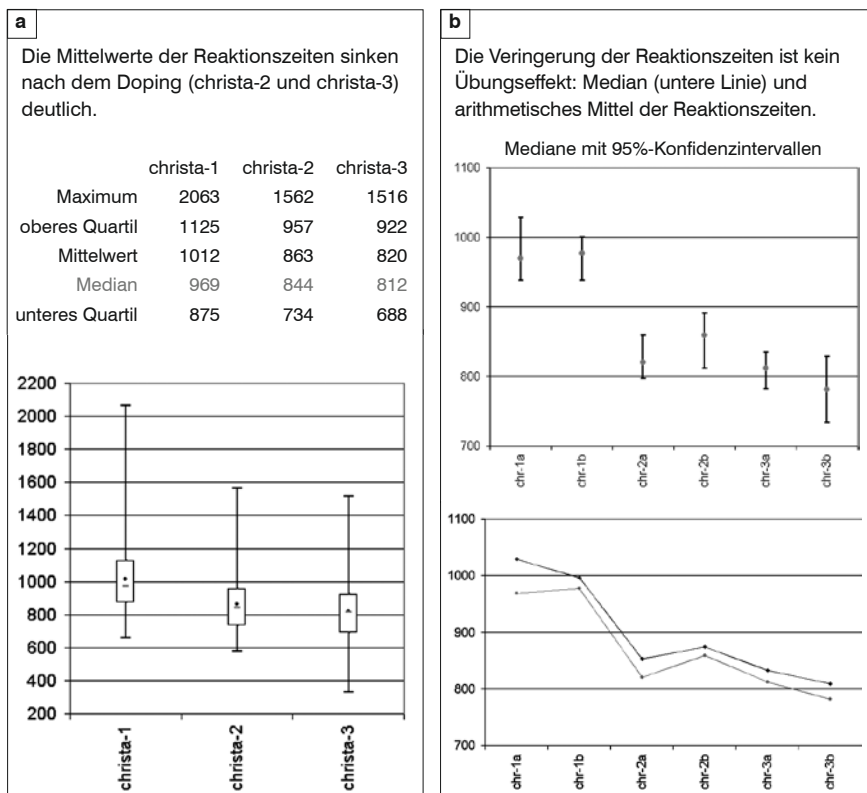


Abb. 4: Ergebnisse des Probanden mit Codename „Christa“ aus drei Serien mit je 200 Reaktionszeiten werden durch Zerlegung sechs 100er-Serien

ten Datensatzes (Christa-1) und notiert das Vorzeichen (+/–) der Differenz. Gleiche Reaktionszeiten werden einfach überschlagen. Hätte das Doping keinen Einfluss, wären die Vorzeichen $B(n; 0,5)$ -verteilt wie die Seiten einer fairen Münze. Bei dem in **Abb. 6** abgebildeten Datensatz mit $n = 200$ Paaren hatte man 162 mal „+“ (und 38 mal „–“). Mit $\mu = 0,5 \cdot 200 = 100$ und $\sigma = \sqrt{200 \cdot 0,5 \cdot 0,5} \approx 7,07$ ergibt die Sigmaregel, dass die Anzahl der „+“ mit 95 % Wahrscheinlichkeit im Intervall

$[\mu - 1,9\sigma; \mu + 1,9\sigma] = [87; 113]$ liegen müsste.

Da die beobachtete Anzahl 162 weit außerhalb liegt, kann man die Hypothese, das Doping habe keinen Einfluss auf die Reaktionszeit für „Christa“ hoch signifikant zurückweisen. Der „P-Wert“, also die Wahrscheinlichkeit, dass man bei Gültigkeit der Hypothese einen noch höheren Testwert erhielte, ist mit $1 - P(38 < X < 162) = 1 \cdot 10^{-15}$ verschwindend klein. Wenn man die beiden Datensätze Christa-1a

und Christa-1b (nüchtern) mit je 100 Reaktionszeiten vergleicht, ergibt sich mit 53 mal „+“ kein signifikanter Unterschied, ein Übungseffekt ist damit nicht nachweisbar.

Da man beim Vorzeichentest auf viele Informationen, die im Datensatz stecken, verzichtet, ist er nicht sehr trennscharf: Die Hypothese der Gleichheit beider Verteilungen wird tendenziell zu häufig akzeptiert.

Test und Vertrauensintervall für den Median

Mit der Idee des Vorzeichentests testet man auch Hypothesen über den Median v einer Wahrscheinlichkeitsverteilung, bestimmt Konfidenzintervalle für v und schlägt so eine Brücke zwischen der beschreibenden und der beurteilenden Statistik.

Der Gedankengang wird am Datensatz Christa-1a vom Umfang $n = 100$ erläutert. Der aus diesem Datensatz geschätzte Median \hat{v} hat hier den Wert 969 ms. Natürlich wird dieser Schätzwert von Versuch zu Versuch schwanken und zwar um den theoretischen (aber unbekannten) Median v der Wahrscheinlichkeitsverteilung. Wir fragen: „Ist die Hypothese $v = 930$ ms mit dem beobachteten Datensatz auf dem 5 % Signifikanzniveau vereinbar?“

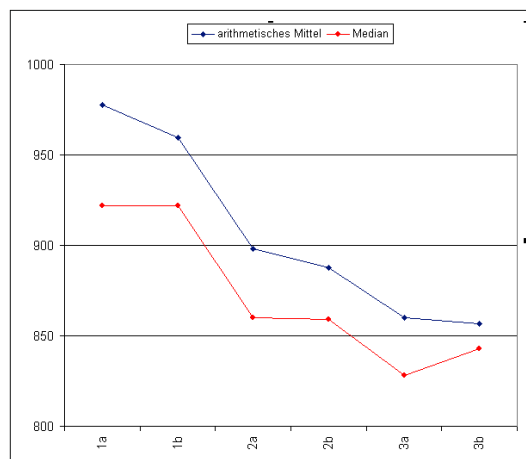
Antwort: Wenn der Median der theoretischen Reaktionszeitenverteilung den Wert $v = 930$ ms hätte, wäre bei vielen Wiederholungen des Reaktionstests die Anzahl k der Reaktionszeiten unter 930 ms $B(100; \frac{1}{2})$ -binomialverteilt, denn jede einzelne Reaktionszeit liegt nach Definition mit Wahrscheinlichkeit 0,5 unter dem angenommenen Median. Dann müsste k nach der Sigmaregel mit 95 %-iger Wahrscheinlichkeit im $1,96\sigma$ -Intervall um den Erwartungswert 50 liegen:

$$k \in [50 - 1,96\sigma; 50 + 1,96\sigma] \approx [40; 60].$$

Tatsächlich liegen aber nur 38 Reaktionszeiten unter 930 ms und die Hypothese muss (knapp) verworfen werden (vgl. **Abb. 6, 7**).

Alle Hypothesen, die man auf diese Weise nicht verwerfen kann, bilden das 95 % Konfidenzintervall für v . Wenn man die 100 Reaktionszeiten aufsteigend sortiert, dann liegt das 95 % Konfidenzintervall zwischen der

Abb. 5: Gruppenergebnisse – in Zwillingspaare aufgeteilte Reaktionszeiten von 15 Schülern. Jeder der sechs Mediane (unten)/arithmetischen Mittelwerte (oben) basiert auf 1500 Reaktionszeiten.



40ten und der 60ten Reaktionszeit. Es gilt: $I_{0,95} = [t_{40}; t_{60}] = [938; 1016]$. Für $n = 200$, wie z. B. bei dem kompletten Datensatz Chista-1 (mit $\sigma = 7,07$), müsste für k analog mit ca. 95 %-iger Wahrscheinlichkeit gelten $k \in [100 - 1,96\sigma; 100 + 1,96\sigma] \approx [86; 114]$. Das 95 %-Konfidenzintervall für den Median ist also $I_{0,95} = [t_{86}; t_{114}] = [953; 1000]$.

Das Ablesen der Grenzen des Konfidenzintervalls aus den sortierten Datenlisten ist mit dem Computer kein Problem, man kann sich die Grenzen aber auch ohne vorher zu sortieren als die 0,4 / 0,6-Quantile ($n = 100$) bzw. 0,43 / 0,57-Quantile ($n = 200$) ausgeben lassen.

Mitunter wird empfohlen, das 95%-Konfidenzintervall des Medians aus den Quartilabstand ($Q_3 - Q_1$), also der Länge der Box im Boxplot zu schätzen nach der Faustregel:

$$I_{0,95} \approx [\hat{v} - 1,58(Q_3 - Q_1)/\sqrt{n}; \hat{v} + 1,58(Q_3 - Q_1)/\sqrt{n}].$$

Diese Faustregel (vgl. Sachs S. 228, Abschnitt 314) liefert für die 100 Datensätze von Chista-1a $I_{0,95} = [930; 1009]$ und für die 200 Datensätze von Chista-1 $I_{0,95} = [941; 997]$ im Vergleich mit den exakten Konfidenzintervallen recht brauchbare Näherungswerte.

Vierfeldertafeln: Mood's Median-Test

Da Vierfeldertafeln derzeit im Stochastikunterricht populär sind, sei am Rande auch Mood's Test auf die Gleichheit zweier Mediane erwähnt: Man fasst die zweimal 200 Reaktionszeiten ohne Doping und mit Doping (Chista-1 und Chista-3) zu einem Satz mit 400 Reaktionszeiten zusammen und bestimmt den gemeinsamen Median (hier 906) vgl. Abb. 8. Dann zählt man, wie viele Reaktionszeiten von Chista-1 unter und wie viele über dem gemeinsamen Median liegen, ebenso bei Chista-3. Die Ergebnisse werden in den beiden Spalten der Vierfeldertafel notiert. Die Tatsache, dass bei Chista-1 „fast alle“ Reaktionszeiten über und bei Chista-3 „fast alle“ Reaktionszeiten unter dem gemeinsamen Median liegen spricht für sich. Wer den Unterschied statistisch sauber belegen möchte, verwendet den Chi-Quadrat- Unabhängigkeitstest (Riemer 1990, Lamba-

	chr-1	chr-3	>		
median	969	812		Anzahl +	162
1	1313	1063	+	Anzahl -	38
2	969	718	+	Summe	200
3	1328	1047	+	Anteil +	0,81
4	875	641	+	Anteil -	0,19
5	1046	750	+		
195	953	781	+		
196	813	1344	-		
197	703	1297	-		
198	1219	843	+		
199	953	579	+		
200	937	766	+		

Abb. 6: Wirkung der Brause: Vergleich von Chista-1 und Chista 3 mit Vorzeichentest (200 Datenpaare)

	chr-1a	chr-1b	>		
	969	976,5		Anzahl +	53
1	1313	781	+	Anzahl -	47
2	969	797	+	Summe	0,53
3	1328	1140	+	Anteil +	0,81
4	875	766	+	Anteil -	0,47
5	1046	1062	-		
95	1125	953	+		
96	1047	813	+		
97	1391	703	+		
98	968	1219	-		
99	782	953	-		
100	844	937	-		

Abb. 7: Übungseffekt? Vergleich von Chista-1a und Chista-1b mit Vorzeichentest (100 Datenpaare)

906	chr-1	chr-3	61,45
<med	60	138	198
<med	135	58	193
	195	196	391

Abb. 8: Vierfeldertafel mit Chi-Quadrat-Testwert

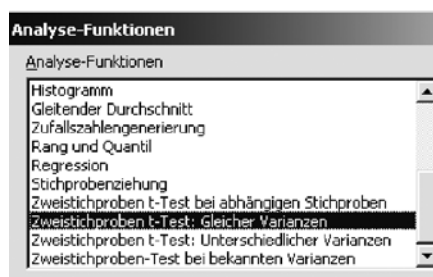
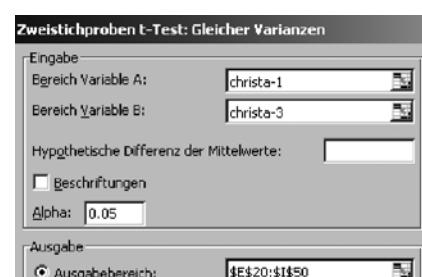


Abb. 9: t-Test als Excel-Analysefunktion



Zweistichproben t-Test unter der Annahme gleicher Varianzen

	Chista-1	Chista-3
Mittelwert	1012	820
Varianz	40694	33383
Beobachtungen	200	200
Gepoolte Varianz	37038	
Hypothetische Differenz der Mittelwerte	0	
Freiheitsgrade (df)	398	
t-Statistik	9.96	
P(T<=t) einseitig	2.57E-21	
Kritischer t-Wert bei einseitigem t-Test	1.65	
P(T<=t) zweiseitig	5.13E-21	
Kritischer t-Wert bei einseitigem t-Test	1.97	

Abb. 10: t-Test: Ergebnisse

Zweistichproben t-Test unter der Annahme gleicher Varianzen

	Chista-1a	Chista-1b
Mittelwert	1028	996
Varianz	47591	33711
Beobachtungen	100	100
Gepoolte Varianz	40651	
Hypothetische Differenz der Mittelwerte	0	
Freiheitsgrade (df)	198	
t-Statistik	1.10	
P(T<=t) einseitig	1.36E-01	
Kritischer t-Wert bei einseitigem t-Test	1.65	
P(T<=t) zweiseitig	2.73E-01	
Kritischer t-Wert bei einseitigem t-Test	1.97	

Hintergrund des t-Tests: Arbeiten mit der Normalverteilung

Man geht von der Modellannahme aus, dass die Reaktionszeiten eines Probanden normal verteilt sind: Ohne Doping nach $N(\mu_1; \sigma)$ und mit Doping nach $N(\mu_2; \sigma)$ mit jeweils der gleichen (aber unbekannten) Standardabweichung σ . Dann ist der Mittelwert aus 200 Reaktionszeiten ohne Doping $N(\mu_1; \frac{\sigma}{\sqrt{200}})$ - normalverteilt und derjenige aus 200 Reaktionszeiten mit Doping $N(\mu_2; \frac{\sigma}{\sqrt{200}})$ - normalverteilt. Die Differenz der Mittelwerte $\bar{x}_2 - \bar{x}_1$ ist dann nach

$N(\mu_2 - \mu_1; \sqrt{\frac{\sigma^2}{200} + \frac{\sigma^2}{200}}) = N(\mu_2 - \mu_1; \frac{\sigma}{\sqrt{100}})$ normalverteilt.

Wenn beide Erwartungswerte gleich wären – und davon geht man bei der zu prüfenden Hypothese aus – dann wäre die Differenz $\bar{x}_2 - \bar{x}_1$ also $N(0; \frac{\sigma}{\sqrt{100}})$ - normalverteilt.

Die normierte Testgröße $t = \frac{(\bar{x}_2 - \bar{x}_1) \sqrt{100}}{\sigma}$ wäre $N(0; 1)$ standard-normalverteilt und müsste nach der Sigmaregel mit 95%iger Wahrscheinlichkeit zwischen $-1,96$ und $+1,96$ (bzw. bei einseitiger Fragestellung mit 95% Wahrscheinlichkeit unter $1,65$) liegen.

Da die Standardabweichung σ nicht bekannt ist, schätzt man sie durch die aus beiden Teil-Stichproben gemittelte Stichprobenstandardabweichung und erhält den Testwert $t = \frac{(1012 - 820) \cdot 10}{192,45} = 9,69$, der im Ausgabefenster als

„Test-Statistik“ auftaucht.

Dadurch, dass man im Testwert t die theoretische Standardabweichung durch die Stichprobenstandardabweichung s ersetzt, ändert sich die Verteilung von t „ein klitzekleines bisschen“. Aus der Standard-Normalverteilung mit Varianz 1 wird die Student'sche t-Verteilung (hier mit 398 Freiheitsgraden), die auch glockenförmig und symmetrisch zu 0 ist, aber mit $398/396 = 1,005$ eine unmerklich höhere Varianz besitzt.

Auch die t-Verteilung mit 198 Freiheitsgraden und der Varianz $198/196 = 1,007$, die beim Vergleich der 100

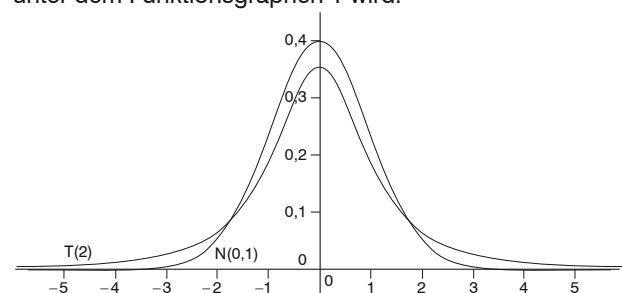
Reaktionszeiten Christa-1a und Christa-1b zum Einsatz kommt, unterscheidet sich praktisch nicht von der Standard-Normalverteilung. Für praktische Zwecke kann man bei den hier verwendeten Stichprobenumfängen die t-Verteilung bedenkenlos durch die Normalverteilung ersetzen. Schließlich kommt es beim „Verstehen stochastischer Aussagen“ sehr viel mehr auf die Interpretation von Testgrößen an, als auf technische Details.

Die t-Verteilung

Die t-Verteilung wurde von William Sealy Gosset, einem Angestellten der irischen Guinness-Brauerei entdeckt. Da sein Arbeitgeber die Veröffentlichung nicht gestattete, veröffentlichte Gosset sie 1908 unter dem Pseudonym Student. Seitdem ist sie in der Statistik als die Student'sche t-Verteilung bekannt. Die t-Verteilung mit n Freiheitsgraden hat die Dichte $t_n(x) = \frac{c}{(1 + \frac{x^2}{n})^{\frac{n+1}{2}}}$

mit dem Erwartungswert $\mu=0$ und der Standardabweichung $\sigma = \sqrt{\frac{n}{n-2}}$ ($n > 2$). Je größer n desto ähnlicher wird sie der Standard-Normalverteilung $N(0, 1)$.

c ist eine Konstante, die dafür sorgt, dass die Fläche unter dem Funktionsgraphen 1 wird.



Dichte der t-Verteilung mit 2 Freiheitsgraden im Vergleich mit der Standard-Normalverteilung

cher-Schweizer 2012, S. 111). Hier ergibt sich der Testwert 61,45, der weit über der kritischen Grenze 3,8 liegt.

Wer – statt nur die Anzahl der Daten über oder unter dem Median zu zählen – zusätzlich auch deren Rangzahlen berücksichtigt, also mehr von der Information nutzen möchte, die in den Messwerten steckt, dem sei der *Mann-Whitney Test* empfohlen. Um ihn zu verstehen, muss man allerdings die Normalverteilung kennen (Sachs, S. 380 f.).

Der noch effektivere t-Test nutzt die volle Information der Daten aus,

wobei man allerdings von der Hypothese ausgeht, dass die Reaktionszeiten normal verteilt sind. Wegen der leichten Unsymmetrie ist diese Hypothese aber mit einer gewissen Vorsicht zu genießen.

t-Test: Gleichheit zweier Mittelwerte

Der t-Test ist „DAS Referenzwerkzeug“, um zu prüfen, ob die Erwartungswerte μ_1 , μ_2 zweier normalverteilter Zufallsgrößen (mit gleicher Standardabweichung σ) übereinstimmen. Da man „die volle Information“, also die Messwerte selber und nicht nur

Vorzeichen von Differenzen oder Rangzahlen nutzt (und auch von der Annahme einer Normalverteilung Gebrauch macht), ist dieser Test trennschärfer als der Vorzeichentest. Zunächst wird am Beispiel von „Christas“ Reaktionszeiten erläutert, wie man z. B. in Excel mit diesem Test arbeiten kann.

Zunächst wählt man unter Extras – Analysefunktionen die Option Zweistichproben t-Test: Gleiche Varianzen (Abb. 9, S. 57) und gibt im sich öffnenden Fenster die zu vergleichenden Datenbereiche christa-1 (200 Reaktionszeiten nüchtern) und christa-3 (200 Re-



meinUnterricht.de

Auf den Lehrer kommt es an.

Doppelt-Blindtest – denn „Der Glaube versetzt Berge“

Professionelle Untersuchungen sind als „Doppelt-Blind-Tests“ angelegt: Dabei wissen weder Versuchsleiter noch Probanden („beide sind blind“), ob sie den Wirkstoff (hier Koffein/Taurin) oder ein wirkungsloses Placebo erhalten.

Damit soll die psychologische Wirkung ausgeschlossen werden, die allein aus „dem Gefühl“ resultiert, ein wirksames „Medikament“ bekommen zu haben. Dass auch Placebos wirken können kommt im Sprichwort zum Ausdruck: „Der Glaube versetzt Berge“.

Da koffeinfreies RedBull nicht angeboten wird und koffeinfreie Cola sich von normaler Cola geschmacklich unterscheiden lässt, ist ein „doppelt-blind“ Design hier nicht möglich. Der reale Versuch schärft aber durch Nachdenken über Fehlereinflüsse ein kritisches Bewusstsein für sachgerechte Interpretation von Versuchsergebnissen. Und auch das gehört – neben einer Kenntnis von Verfahren – zur statistischen Allgemeinbildung.



Mit meinUnterricht.de bereiten Sie Ihren Unterricht schnell, fundiert und digital vor!



Virtueller Schreibtisch:
Planen Sie Ihren Unterricht auf einem digitalen, mobilen Schreibtisch



Intelligente Suche:
Finden Sie einfach und treffsicher tausende hochwertige Materialien renommierter Verlage



Ihr Fach ist dabei:
Auch mit einer großen Auswahl an Mathematik- und Physikmaterialien

Bislang 25.000 Seiten
Wir erweitern unsere Bibliothek laufend!



Verlosung

Gewinnen Sie 1 von 3 Tablets!

Um am Gewinnspiel teilzunehmen, geben Sie unter www.meinUnterricht.de/gewinn den Code **gwn201231** ein.

Spezial-Angebot

10 für 12!
Sichern Sie sich 2 kostenlose Monate und sparen Sie zusätzlich Monat für Monat in der Einführungsphase.

Unter www.meinUnterricht.de/spezial erhalten Sie nur mit dem Code **Friedrich201216** Ihre exklusiven Vorteile!

aktionszeiten – Langzeitdoping) sowie das gewünschte Signifikanzniveau ein. Im Ausgabebereich (**Abb. 10**, S. 57) werden die mittleren Reaktionszeiten, deren Varianzen und auch der (bei unterschiedlichen Stichprobenumfängen gewichtete) Mittelwert der Varianzen als „gepoolte Varianz“ ausgegeben. Wenn der Testwert („t-Statistik“, hier 9,96) über dem kritischen t-Wert (hier 1,97) liegt, muss man die Hypothese auf dem angegebenen Signifikanzniveau (hier 5%) verwerfen.

Die Unterschiede zwischen den ungedopten und den gedopten Reaktionszeiten sind beim zweiseitigen Test wegen $9,69 > 1,97$ signifikant. Die Wahrscheinlichkeit, dass man bei Gültigkeit der Hypothese ($\mu_1 = \mu_2$) einen noch größeren Testwert erhielte – Statistiker bezeichnen diese Wahrscheinlichkeit als P-Wert – ist mit $P = 2,57 \cdot 10^{-21}$ praktisch 0. Er ist deutlich kleiner als der entsprechende P-Wert beim Vorzeichentest. Wenn man zum Vergleich die Mittelwerte der jeweils 100 Reaktionszeiten am Anfang und am Ende der Messung „nüchtern“ miteinander vergleicht, ergeben sich dagegen keine signifikanten Unterschiede – genau wie beim Vorzeichentest.

Resümee

Die Verbesserung der Reaktionszeiten durch Doping mit Energiebrausen erweist sich (entgegen anfänglicher Skepsis) als statistisch höchst signifikant. Sie ist kein Produkt reiner Zufallsschwankungen, was sich auch in vielen anderen Versuchsgruppen bestätigt hat. Dass sie von manch einem Vereinssportler auch für relevant gehalten wird, belegen Interviews und leere Aludosen in manch einer Umkleidekabine.

Was die „Pharmaforschung im Klassenraum“ für den Mathematikunterricht so interessant macht, ist der absolut authentische Kontext und sein statistisches Spektrum, das von der beschreibenden Statistik der Mittelstufe über die abiturrelevante „Binomialstatistik“ bis zur Normalverteilung und dem t-Test reicht, dessen Durchdringung/Anwendung auch im Rahmen von Facharbeiten höchst motivierend sein kann.

Literatur

Sachs, L. (1999): Angewandte Statistik. – 9. Aufl. Springer, Berlin, Heidelberg.
Riemer, W. (1990): Der Chi-Quadrat-Unabhängigkeitstest. – In: MU 5-1990, S. 50 – 66.
Lambacher-Schweizer Stochastik (2012) 735710, Klett Stuttgart.
www.riemer-koeln.de