

Beurteilende Statistik, die Spaß macht: Was Aufbewahrungsboxen mit 100 Büroklammern erzählen können¹

SUDIPTA ROY, JOILET, ILLINOIS, USA

¹ aus: *Teaching Statistics* 41 (2019) 1, S. 25–29.
Übersetzung und Bearbeitung: ANDREAS PRÖMMEL UND
MARKUS VOGEL

Zusammenfassung: In diesem Artikel wird ein Klassenraum-Experiment zu Aufbewahrungsboxen mit 100 Büroklammern vorgestellt. Die Schüler müssen dabei selbst aktiv werden und verschiedene thematische Aspekte der Beurteilenden Statistik anwenden: Hypothesentesten, Vertrauensintervalle und Toleranzintervalle.

1 Einleitung

Experimente sind fester Bestandteil der Bildung und Erziehung in den Naturwissenschaften. Im Gegensatz dazu sind Experimente im Statistikkunterricht nicht standardmäßig vorgeschrieben und liegen somit im Ermessen der Lehrperson. Die Argumente für die Einbindung von Zufallsexperimenten in den Statistikkunterricht sind naheliegend: Eigene Erfahrungen der Schüler mit realen Daten und in realen Kontexten ergänzen sinnvoll Lehrbuchaufgaben und können die manchmal abstrakt wirkenden Konzepte verständlicher und fesselnder machen. Zu diesem Zweck beschreiben Pfaff und Weinberg (2009) Zufallsexperimente mit Gegenständen wie Karten und Würfeln, um das Gesetz der großen Zahlen bzw. das Konzept von Stichprobenverteilungen zu illustrieren, und Beutel mit Bingo-Chips in verschiedenen Farben, um Hypothesentests und Vertrauensintervalle für Anteile zu veranschaulichen. M&M-Schokolinsentüten sind auch ein beliebtes Objekt für das Testen von Anteilen (Johnson 1993; Fricker 1996; Retseck 2013; Wicklin 2017) sowie für weitere statistische Untersuchungen (Froelich und Stephenson 2012). Zufallsexperimente mit Schokoladen-Chips wurden entworfen, um das Konzept der Qualitätskontrolle (Baker 2014) und das Konzept von Mittelwertsunterschieden (Magel 1998) anschaulich zu vermitteln.

Wir stellen in diesem Artikel ein Zufallsexperiment vor, das sich um das Etikett auf einem verpackten Produkt dreht, d. h. den Anspruch eines Herstellers an sein Produkt widerspiegelt. Als Verbraucher kaufen wir ständig abgepackte Produkte, wie z. B. 250 Zahnstocher, 100 Büroklammern, 80 Trocknertü-

cher, 16 Unzen Fleischbrät, 3 Pfund Kartoffeln. Der Anspruch des Herstellers ist durch das Zähl- (oder Gewichts-) Etikett auf der Verpackung deutlich ausgewiesen und ist ein wesentlicher Faktor bei unserer Kaufentscheidung (wenn wir uns entscheiden, die 80 Packungen Trocknertücher statt der 105 Packungen zu kaufen). Dennoch wiegen wir selten die Tüten mit Kartoffeln oder Äpfeln oder zählen die tatsächliche Anzahl der Büroklammern (oder Zahnstocher oder Trocknertücher). Was wäre, wenn man einmal die tatsächliche Anzahl der Büroklammern in einer Stichprobe von Aufbewahrungsboxen zählen würde? Enthalten die Boxen die Anzahl der auf dem Etikett angegebenen Büroklammern? Wären wir über die Ergebnisse erstaunt? Was könnte man anhand der Stichprobe über die durchschnittliche Anzahl von Büroklammern in einer Box ableiten? Kurz: Wie viele Büroklammern kann ein Verbraucher in einer Box tatsächlich erwarten? Das vorgeschlagene Zufallsexperiment enthält zwei Aspekte nach Pfaff und Weinberg (2009): (1) Es ist von einer „natürlichen“ Frage motiviert und die Schüler sammeln und interpretieren die Daten, und (2) es bezieht sich auf einen Kontext, den die Studierenden als Verbraucher kennen und sie können dazu konkrete Fragen formulieren, wie z. B. die folgenden:

- Wie genau ist das Zählkennzeichen? Wird der Anspruch des Herstellers nach der statistischen Fachterminologie durch die Daten gestützt?
- Was kann man laut Stichprobenergebnis über die durchschnittliche Anzahl der Büroklammern pro Aufbewahrungsbox schließen?
- Wie viele Artikel kann ein Verbraucher in seiner Box erwarten?

Die erste Frage ist mit einem Hypothesentest zu beantworten, und die zweite Frage wird durch die Konstruktion eines Vertrauensintervalls beantwortet. Diese beiden Themen werden in der Regel in Einführungskursen zur Statistik diskutiert. Die dritte Frage erfordert den Aufbau eines Toleranzintervalls (Gitlow und Awad 2013), ein Thema, das dazu beitragen kann, besser zu verstehen, was Vertrauensintervalle sind und was nicht. Denn dieses Experiment bietet einen natürlichen Kontext, um Toleranzin-

tervalle einzuführen und mit Vertrauensintervallen zu vergleichen. Im weiteren werden nun das Experiment mit zwei verschiedenen Marken von Büroklammern beschrieben und entsprechende Analysen präsentiert.

2 Sammeln von Daten

Für das Experiment erhält jeder Schüler eine Box mit 100 Büroklammern. Laut Beschriftung des Etiketts zählt jeder Schüler die Anzahl der Gegenstände in seiner Box. Empfehlenswert ist die Sortierung nach 5er oder 10er Gruppen, um das Zählen zu erleichtern. Jeder Schüler gibt anschließend das Ergebnis seiner Zählung in eine Datei ein, die idealerweise (in Echtzeit) auf eine Projektionsfläche projiziert wird.

Die Reaktionen der Schüler, die den Datensatz mit dem Ergebnis ihrer Zählung füllen, geben einen interessanten Einblick in ihre Erwartungen. „Ich fühle mich betrogen!“, ist eine gängige Bemerkung von denen, die weniger als die 100, wie auf dem Etikett angegeben, finden. Diejenigen, die mehr als 100 haben, sagen, sie haben „Glück gehabt“. Offenbar interessieren sich die Schüler für die durch das Zufallsexperiment wahrgenommene Diskrepanz zwischen dem Etikettaufdruck und der tatsächlichen Anzahl der Büroklammern in ihrer Box.

Die Daten werden in einer geeigneten Form gesammelt, welche die Schüler anschließend analysieren können, z. B. in einer Excel-Datei. Die Datensätze, die in zwei Durchgängen des Zufallsexperiments gesammelt wurden, finden sich in Abbildung 1.

Die durchschnittliche Anzahl der Büroklammern in den 20 Staples®-Boxen beträgt 102,05 mit einer Standardabweichung von 2,48. Der Durchschnitt für die Stichprobe von 22 Office Depot®-Boxen liegt bei 100,5 bei einer Standardabweichung von 2,32.

3 Analyse der Daten

Obwohl die Anzahl der Büroklammern, die pro Schachtel verpackt werden, diskret ist, ist es vielleicht nicht unvernünftig, näherungsweise eine Normalverteilung zu verwenden. Eine explorative Analyse legt nahe, dass die Probanden zwar nicht perfekt symmetrisch in Glockenform sind, aber auch nicht signifikant verzerrt (Abb. 2). In den folgenden Schritten suchen die Schüler Antworten auf die drei gestellten Fragen.

Frage 1: Wie genau ist der Etikettaufdruck „100 Büroklammern“?

Die Schüler werden erkennen, dass die Beantwortung dieser Frage einen Hypothesentest erfordert. Für einen Test ist zunächst zu überlegen, was für den vorliegenden Sachverhalt als Nullhypothese und was als Alternativhypothese festgelegt werden soll. Es ist durchaus denkbar, dass ein Verbraucher erwartet, dass jede Box genau 100 Büroklammern enthält. Obwohl Büroklammern kein teures Produkt sind, würde ein reines Profitmotiv des Herstellers darauf hindeuten, dass dieser aus betriebswirtschaftlicher Sicht die Boxen eher nicht gezielt überfüllt. Die Wirtschaftsethik hingegen fordert, dass Boxen auch nicht gezielt unterfüllt werden. Aus Sicht der Qualitätskontrolle ist daher zu erwarten, dass die Füllmaschine so ka-

Data Set 1: Staples®					Data Set 2: Office Depot®				
100	105	100	101	101	96	101	100	101	
102	98	100	101	97	103	100	106	101	101
101	102	107	100	101	102	97	101		
105	102	101	105	100	98	104	102		
106	102	99	104	99	102	101	99		

Abb. 1: Anzahl von Büroklammern pro Box

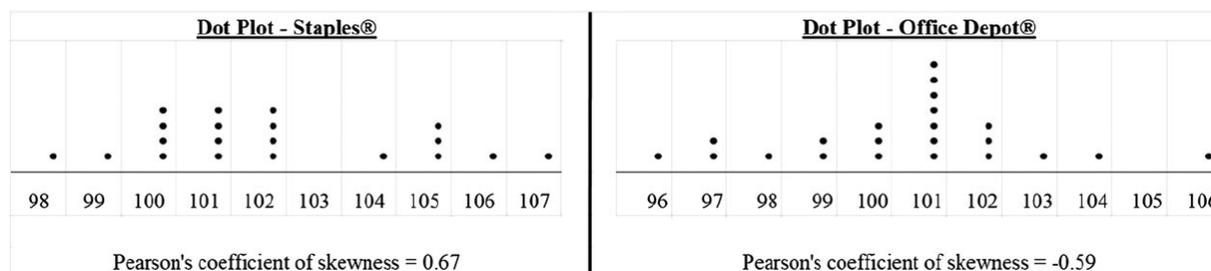


Abb. 2: Explorative Analyse

libriert ist, dass sie eine Box mit (genau) 100 Büroklammern nach der angegebenen Zählung füllt.

Angesichts der oben genannten Argumente erscheint ein zweiseitiger Test am besten geeignet. Die Nullhypothese und die Alternativhypothese(n) sind dementsprechend wie folgt angegeben:

- $H_0: \mu = 100$
- $H_1: \mu \neq 100$

Ein Hypothesentest für einen Mittelwert basiert auf der Auswertung der Stichprobenverteilung der Stichprobenmittelwerte einer Stichprobe, die unter Bezug auf den Zentralen Grenzwertsatz etwa glockenförmig ist. Unter Berücksichtigung der unbekanntes Standardabweichung und des geringen Stichprobenumfangs wird ein t -Test durchgeführt. Abbildung 3 stellt die Ergebnisse vor.

Staples®	Office Depot®
Test statistic = 3.70	Test statistic = 1.10
p -value = 0.0015	p -value = 0.284
Decision: Reject H_0	Decision: Do not reject H_0

Abb. 3: Ergebnisse des Hypothesentests

Die p -Werte deuten darauf hin, dass es keine Hinweise darauf gibt, dass sich das Office Depot®-Mittel von 100 unterscheidet. Andererseits liefern die Daten für Staples® eine starke Evidenz dafür, dass der Mittelwert nicht 100 ist.

Story 1: Im Ergebnis des Hypothesentests lässt sich ableiten, dass die Staples®-Boxen im Durchschnitt keine 100 Clips enthalten. Dass die Office Depot®-Boxen durchschnittlich 100 Clips enthalten, kann nicht ausgeschlossen werden.

An dieser Stelle des Experiments verstehen die Schüler, dass sich das Etikett aus der Perspektive der Qualitätskontrolle auf eine durchschnittliche und nicht auf eine individuelle Zählung bezieht. Als Verbraucher dürfen sie mit den Testergebnissen für Office Depot® nicht unzufrieden sein. Mit dem Ergebnis von Staples® werden sie sicher sehr zufrieden sein.

Darüber hinaus werden sie den Grund dafür erkennen, dass der Hersteller es nicht so sehr ernst mit der Genauigkeit nimmt, denn Büroklammern sind ein relativ preiswertes Produkt. Dennoch wäre ein Hinweis darauf, wie viele Büroklammern im Durchschnitt in einer Box sind, hilfreich. Das wirft die nächste Frage auf.

Frage 2: Was lässt sich anhand der Stichprobe über die durchschnittliche Anzahl von Büroklammern, die pro Box verpackt sind, ableiten?

Die Schüler werden erkennen, dass ein Vertrauensintervall konstruiert werden muss, um eine Schlussfolgerung über den wahren Mittelwert zu ziehen, d. h. um abzuschätzen, wie viele Büroklammern im Durchschnitt pro Box abgefüllt werden. Mit Hilfe einer t -Verteilung werden die Schüler das Vertrauensintervall mittels $\bar{X} = t_{\alpha/2} \cdot s / \sqrt{n}$ konstruieren.

Das 95 %-Vertrauensintervall für den Staples®-Datensatz liegt im Durchschnitt bei 100,89 bis 103,21 oder etwa 101 bis 103 Büroklammern. Das 95 %-Vertrauensintervall für Office Depot® beträgt 99,5 bis 101,6 oder etwa 100 bis 102.

Story 2: Die Staples®-Probe dürfte nicht aus einer Charge mit echtem Mittelwert von 100 stammen. Aufgrund der Stichprobe lässt sich ableiten, dass die durchschnittliche Anzahl der pro Box gefüllten Büroklammern zwischen 101 und 103 liegt. Der Etiketaufdruck von 100 liegt außerhalb des Vertrauensintervalls, was durch den vorher durchgeführten Hypothesentest bestätigt wird. Die durchschnittliche Anzahl der Büroklammern in den Office Depot®-Boxen liegt zwischen 100 und 102. In Übereinstimmung mit den Ergebnissen des Hypothesentests enthält dieses Vertrauensintervall die etikettierte Zahl von 100.

Obwohl die Staples®-Boxen im Schnitt überfüllt sind, variiert die tatsächliche Anzahl der Büroklammern zwischen 98 und 107. Auch wenn die Office Depot®-Boxen im Durchschnitt 100 Büroklammern enthielten, enthielten die einzelnen Boxen zwischen 96 und 106. Das wirft die letzte Frage auf:

Frage 3: Wie viele Büroklammern kann ein Verbraucher in seiner Box erwarten?

Die Anzahl der Büroklammern die ein einzelner Verbraucher in seiner Box erwarten kann, kann vom Durchschnitt abweichen. Wie Vardeman (1992) anmerkt, „... viele praktische Probleme werden eher in Bezug auf einzelne Messungen anstatt als Parameter von Verteilungen formuliert.“ Um die dritte Frage zu beantworten, müssen die Schüler Toleranzintervalle konstruieren.

Ein zweiseitiges Toleranzintervall $[L,U]$ hat eine Wahrscheinlichkeit β , dass mindestens der Anteil γ der Population innerhalb $[L,U]$ liegt, wobei L und U die Unter- und Obergrenzen des Toleranzintervalls sind (Wald und Wolfowitz 1946; Gitlow und Awad

2013). Vereinfacht gesagt definiert ein Toleranzintervall Grenzen, innerhalb derer ein bestimmter Anteil der Population erwartet wird. Im Gegensatz dazu definiert ein Vertrauensintervall Grenzen, innerhalb derer der Populationsmittelwert erwartet wird. Der Wert β ist das Vertrauensniveau, das die Schüler kennen. Sie müssen sich nun für den Anteil γ der Population entscheiden, den sie durch ihr Toleranzintervall abdecken möchten.

Falls Toleranzintervalle nicht Gegenstand des Unterrichts sind, benötigen die Schüler ein Grundverständnis für den Aufbau von Toleranzintervallen. Unter der Annahme, dass die zugrundeliegende Variable einer Normalverteilung folgt, ist das Toleranzintervall ein Schätzintervall des wahren Intervalls $\mu \pm z_{\alpha/2} \cdot \sigma$, das den $(1-\alpha)$ -Anteil der tatsächlichen Werte enthält. Bei der Konstruktion des Intervalls werden μ und σ durch den Stichprobenmittelwert \bar{X} und die Stichprobenstandardabweichung s ersetzt, ohne die Variabilität von \bar{X} und s zu berücksichtigen. Stattdessen wird ein Toleranzfaktor k definiert und das Toleranzintervall als $\bar{X} \pm k \cdot s$ konstruiert.

Tabellen von Toleranzfaktoren k können online gefunden werden, aber Schüler benutzen lieber ein Applet (<http://statpages.info/tolintvl.html>). Die erforderlichen Eingaben sind der Stichprobenumfang, der Stichprobenmittelwert und die Standardabweichung sowie die gewählten Werte von β und γ . Das gleiche Vertrauensniveau von 95 %, das für das Vertrauensintervall verwendet wurde, kann auch hier für β verwendet werden. Nehmen wir an, die Schüler wählen einen Wert von 90 % für den Anteil γ . Bei den Staples®-Daten beträgt damit das Toleranzintervall 96,32 bis 107,78. Die Schüler werden daraus schließen, dass ein 95 %-Toleranzintervall für 90 % der Boxen 96 bis 108 Büroklammern beträgt. Auf Basis der Office Depot®-Daten werden sie ebenfalls daraus schließen, dass ein 95 %-Toleranzintervall für 90 % der Boxen 95 bis 106 Büroklammern beträgt (Toleranzintervall 95,25 bis 105,75).

Story 3: Wir hatten zuvor festgestellt, dass die Office Depot®-Boxen im Durchschnitt 100 Büroklammern enthalten. Auf der Grundlage zusätzlicher Informationen aus dem Toleranzintervall können wir nun sagen, dass wir zu 95 % davon überzeugt sind, dass 90 % der Boxen zwischen 95 und 106 Büroklammern enthalten. Unsere Daten hatten zuvor darauf hingedeutet, dass die Staples®-Boxen überfüllt sind und im Schnitt zwischen 101 und 103 Büroklammern pro Box enthalten. Außerdem können wir jetzt auch sa-

gen, dass ein Verbraucher, der nur 96 Büroklammern in seiner Box findet, sich „betrogen“ fühlen könnte. Auf der anderen Seite könnte ein Verbraucher auch „Glück“ haben und 108 Büroklammern statt der 100 finden, die auf der Grundlage des Etiketts zu erwarten wären. Die Geschichten 2 und 3 unterstreichen, warum „Schüler in Einführungskursen zur Statistik sowohl (Intervalle zu) Vertrauen als auch Toleranz brauchen“ (Gitlow und Awad 2013).

Ein Vertrauensintervall schätzt, wie sehr der Mittelwert streut, während ein Toleranzintervall eine Schätzung dafür ist, wie sehr die tatsächliche Anzahl variiert. Toleranzintervalle sind daher in der Regel breiter als Vertrauensintervalle, wie Abbildung 4 zeigt.

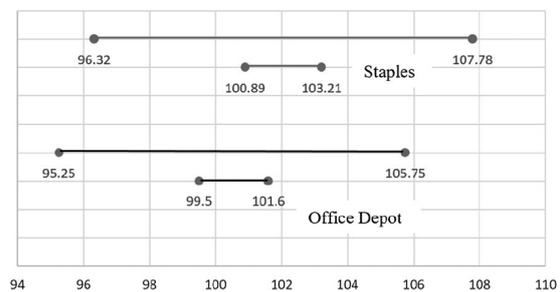


Abb. 4: 95 % Vertrauensintervall und 95 % Toleranzintervall für 90 % der Boxen für beide Hersteller

Nach diesem Experiment wollen Schüler in der Regel wissen, wie die Boxen mit Büroklammern gefüllt werden. Für die statistische Prozesssteuerung in Fertigungsprozessen werden Kontrolldiagramme verwendet, die auf kontinuierlichen Variablen basieren (Heizer und Render 2012). Das bedeutet, dass die Boxen nach Gewicht gefüllt werden. Der aktuelle Stand der Fertigungstechnik legt hingegen nahe, dass eine Füllmaschine Büroklammern auch mit hoher Genauigkeit zählen kann. Unsere Anfrage bei den Herstellern dazu deutet darauf hin, dass einige nach Gewicht füllen, während andere dies nach Zählen tun.

4 Schlussfolgerung

Statistik ist ein leistungsfähiges Werkzeug, um die Daten um uns herum sinnvoll zu verstehen. Experimente im Unterricht durchzuführen, ist eine gute Möglichkeit für Schüler zu sehen, wie Statistik in der realen Welt angewendet wird.

In diesem Artikel habe ich ein einfaches, aber „authentisches“ Experiment vorgeschlagen, das die Anwendung mehrerer Konzepte der beurteilenden Statistik veranschaulicht. Ausgehend von einer „natürlichen“ und „offensichtlichen“ Frage nimmt das

Experiment die Schüler mit auf eine Reise, auf der sie verschiedene Geschichten von einem Herstelleretikett aus aufzeichnen, die vom Standpunkt eines Verbrauchers her bedeutsam sind.

Ich habe sehr positive Erfahrungen damit gemacht, dieses Experiment in meiner Klasse durchzuführen. Auf die Bitte, über ihre Erfahrungen nachzudenken, äußerte ein Schüler: „Durch die Behandlung dieses Hypothesentests, ... bin ich, glaube ich, ein bewusster Konsument geworden.“ Ein anderer sagte: „Aus diesem Experiment habe ich gelernt, dass Statistik nützlich ist und auf viele Dinge im Alltag angewendet werden kann, die mir nicht bewusst waren. Vor diesem Experiment glaubte ich einfach an die Anzahl, die auf den Boxen angegeben wurde.“

Ein weiteres Produkt, das nach Anzahl verpackt ist und sich gut für diese Aktivität eignet, sind Boxen mit Zahnstochern. Produkte, die nach Gewicht verpackt sind, die alternativ verwendet werden könnten, sind Snack-Packs mit Chips, Popcorn oder Nüssen und individuell verpackte Riegel mit Schokolade, Trail-Mix oder Granola.

Als Verbraucher werden die Schüler im Laufe ihres Lebens viele verschieden verpackte Produkte kaufen. Dieses Experiment wird nicht dazu führen, dass sie abgepackte Kartoffeln, Zwiebeln oder Äpfel wiegen, nur um die günstigste Packung für ihr Geld aufzuspüren. Sie werden sicherlich keine Boxen mit Büroklammern, Trocknertüchern oder Zahnstochern öffnen, um zu überprüfen, ob diese wirklich die angegebene Anzahl von Gegenständen enthält, bevor sie dafür bezahlen. Vielleicht werden sie sich aber fragen, ob ihre Box wirklich die Anzahl der auf dem Etikett angegebenen Artikel enthält und sie sollten wissen, dass dies durchaus möglich ist. Dieses Experiment sollte sie gelehrt haben, dass es vom Zufall abhängt, ob man „Glück“ hat oder sich „betrogen“ fühlt.

Literatur

- Baker, A. (2014). Teaching quality control with chocolate chip cookies. *Teaching Statistics*, 36(1), 2–6.
- Fricker, R. D. Jr. (1996). Yikes! Where did all the blue M&M's go? The mysterious case of the blue M&M's®. *Chance*, 9(4), 19–22.
- Froelich, A. G. and Stephenson, W. R. (2012). How much do M&M's weigh? *Teaching Statistics*, 35, 14–20. <https://doi.org/10.1111/j.1467-9639.2012.00515.x>.
- Gitlow, H. and Awad, H. (2013). Intro stats students need both confidence and tolerance (intervals). *The American Statistician*, 67(4), 229–234.
- Heizer, J. and Render, B. (2012). *Operations Management*, Tenth edn. Prentice Hall Pages 179–180.
- Johnson, R. W. (1993). Testing colour proportions of M&Ms. *Teaching Statistics*, 15, 2–4. <https://doi.org/10.1111/j.1467-9639.1993.tb00243.x>.
- Magel, R. C. (1998). Testing for differences between two brands of cookies. *Teaching Statistics*, 20(3), 81–83.
- Pfaff, T. P. and Weinberg, A. (2009). Do hands-on activities increase student understanding?: a case study. *Journal of Statistics Education*, 17(3).
- Retseck, G. (2013). Statistical science: melt-in-your-mouth math. *Scientific American*.
- Vardeman, S. B. (1992). What about the other intervals? *The American Statistician*, 46(3), 193–197.
- Wald, A. and Wolfowitz, J. (1946). Tolerance limits for a Normal distribution. *The Annals of Mathematical Statistics*, 17, 208–215.
- Wicklin, R. (2017). The distribution of colors for plain M&M candies. SAS Blogs. Available at <https://blogs.sas.com/content/iml/2017/02/20/proportion-of-colors-mandms.html>

Anschrift des Verfassers

Sudipta Roy
University of St. Francis
Joliet, IL, USA
sroy@stfrancis.edu