Entdecken von Experimentellem Design: Eine interaktive Unterrichtsübung zum Tee-Verkostungsexperiment von R. A. Fisher[1]

Thomas R. Fanshawe, Oxford

[1] Original: *Teaching Statistics* 43(3), 2021, S. 140–145.
doi.org/10.1111/test.12287
Übersetzung: Manfred Borovcnik

Zusammenfassung: *Wertschätzung des Designs von Versuchen ist ein wichtiger Aspekt der Einführung in die Statistik in einem breiten Spektrum angewandter Disziplinen einschließlich der medizinischen Statistik. Verständnis der Auswirkungen von Designentscheidungen auf die Auswahl der Methode zur Analyse und der anschließenden Interpretation der Ergebnisse können helfen, statistisches Denken in den experimentellen Prozess einzubetten. Ich bespreche eine interaktive Übung, basierend auf R. A. Fishers berühmten „Lady Tasting Tea"-Experiment, die zur Sensibilisierung für Designfragen im Rahmen eines Statistik-Moduls im Bachelorstudium dienen soll.*

Die Übung verfolgt den Ansatz zum entdeckenden Lernen, wobei die Schüler ermutigt werden, mögliche Designs zu identifizieren und Lösungen in Kleingruppen zu erörtern. Dabei sollte die Lehrperson weitgehend im Hintergrund bleiben und nur kleine Denkanstöße geben. Der Wert dieses Lehrstils und möglichen Erweiterungen des Tee-Verkostungsexperiments auf verwandte Themen, die breiter nutzbar sind, werden auch diskutiert.

Schlagwörter: Unterricht, Experimentelles Design, Gruppenunterricht, Studiendesign, Statistikunterricht.

1 Einleitung

Es wurde behauptet, dass „die Wahl eines geeigneten Studiendesigns einer der schwierigsten Aspekte von Problemen in der angewandten Statistik ist" (Gore 1984). Infolgedessen wird eine Würdigung der Versuchsplanung als wertvoller Bestandteil in der einführenden Statistikausbildung angesehen (Easterling 2004), insbesondere in angewandten Disziplinen wie der Medizin, wo mangelhaftes Design zu Studien führen kann, die wissenschaftlich ineffizient oder gar unethisch sind (Appleton 1990). An anderer Stelle wurde argumentiert, dass Lernen, wie man Experimente entwirft, ein nützlicheres Ziel ist als Lernen, Daten aus Experimenten zu analysieren, die von anderen durchgeführt wurden (Cobb 2007).

Dennoch ist es schwierig, Beispiele zu finden, die erfolgreich Themen des experimentellen Designs auf Anfängerniveau einbauen. Es bleibt die Versuchung, eine detaillierte Analyse Fragen zum Design vorzuziehen (MacDougall, Cameron und Maxwell 2020), auch wenn das eigentliche Ziel begriffliches Verständnis ist und eben nicht analytische Kompetenz (Bradstreet 1996). Wie Bland, Altman, und Royston (1990) darauf hinweisen, ist „die Funktion der medizinischen Statistik eben nicht bloß technischer Natur". An sich ist es vorteilhafter, die Rolle der Statistik als integraler Bestandteil des Forschungsprozesses aufzuzeigen, statt eine Reihe von numerischen Berechnungen oder mathematischen Formeln zu forcieren (Wild 1994).

Frühere Arbeiten heben zugängliche Beispiele hervor (Hiebert 2007), wie etwa jene, die sich auf wissenschaftliche Kontroversen beziehen (Bennett 2015). Die Auswahl muss sorgfältig auf die Anforderungen des Kurses und die Fähigkeiten der Studierenden zugeschnitten werden. Andere Strategien für den Unterricht von experimentellem Design konzentrieren sich auf virtuelle Experimente oder Simulation, was für Anwendungen wie Maschinenbau oder Naturwissenschaften besser geeignet sein mag (Anderson-Cook und Dorai-Raj 2001; Darius, Portier und Schrevens 2007).

Eine andere Alternative zur Förderung aktiven Lernens lässt Studierende zunächst ein Experiment entwerfen und dann durchführen, bevor sie die erhaltenen Daten in der Klasse analysieren. Dieser Ansatz, der einfache oder humorvolle Szenarien, welche die Mitarbeit erhöhen, verwenden kann, zielt darauf ab, begriffliches Verständnis für Versuchsmethoden zu verbessern, obwohl dies mehr Zeit benötigt (Montangero et al. 2018; Pyott 2021).

Dieser Beitrag beschreibt eine kurze Übung, welche experimentelles Design besser verstehen lassen soll. Sie wurde im Rahmen eines Medizinstatistik-Moduls für Medizinstudierende durchgeführt, sie erfordert aber wenig Vorkenntnisse und wäre auch für andere Zielgruppen auf einem vergleichbaren oder niedrigeren akademischen Niveau geeignet.

Die Übung baut auf dem berühmten Tee-Verkostungsexperiment von R. A. Fisher auf. Nach einer kurzen Erläuterung des historischen Ursprungs dieses Experiments erkläre ich, wie dies zu einem Unterrichtsexperiment weiterentwickelt wurde, und präsentiere Ergebnisse aus dem Unterricht. Die Diskussion beschreibt mögliche Modifikationen und skizziert die Vorteile dieser Art von Experimenten im einführenden Unterricht.

2 Historischer Hintergrund: Das Tee-Verkostungsexperiment von Fisher

Seit der ursprünglichen Veröffentlichung (Fisher 1935) wurde Fishers Experiment schon viele Male überarbeitet (Basu 1980; Bi und Kuesten 2015; Box 1990; Chadwick und Dudley 1983; Gridgeman 1959; Morton 1975; Neyman 1950; Wrightson 1953); es erscheint sogar im Titel eines Buches zur Geschichte der statistischen Wissenschaft im 20. Jahrhundert (Salsburg 2001). Eine einfache Übersicht wird von Senn (2012) bereitgestellt. Es folgt eine Zusammenfassung der wichtigsten Punkte.

Fishers Experiment zielte darauf ab, zu testen, ob „eine Lady" (später als Muriel Bristol benannt) beim Trinken einer Tasse Tee, „unterscheiden kann, ob die Milch oder der Tee zuerst in die Tasse gefüllt wurde" (Fisher 1935, S. 13). Fishers Ansatz ist der, dass von acht Tassen vier mit „Milch zuerst" (MZ) und vier mit „Tee zuerst" (TZ) zubereitet werden, wobei die Reihenfolge bei der Zubereitung zufällig sein sollte.

Wichtig ist, dass auch festlegt wird, dass man der Lady mitteilt, wie viele Tassen vom jeweiligen Typ sind. Fisher geht davon aus, dass die abgegebenen Einschätzungen damit konform gehen, was bedeutet, dass sie vier Tassen als MZ und vier als TZ klassifiziert. Dieses Design wurde als „doppelte Tetrade [Vierheit]" (Gridgeman 1959) oder „Oktade [Achtergruppe]" (Bi und Kuesten 2015) bezeichnet.

Fisher spezifiziert eine Nullhypothese, nach der das „Subjekt keinerlei „Fähigkeit der sensorischen Erkennung des Unterschieds" besitzt (Fisher 1935, S. 19) und berechnet, basierend auf dieser Annahme, dass die korrekte Klassifizierung aller acht Tassen mit folgender Wahrscheinlichkeit auftreten würde (Erläuterung im zusätzlichen Lehrmaterial):

$$1/\binom{8}{4} = 1/70.$$

Eine ordnungsgemäße Klassifizierung von drei der vier MF und daher auch drei der vier TF, würde danach mit folgender Wahrscheinlichkeit auftreten:

$$\binom{4}{3} \cdot \binom{4}{1} / \binom{8}{4} = 16/70.$$

Da diese Wahrscheinlichkeit nicht klein genug ist, stellt Fisher fest, dass dies „nicht als statistisch signifikanter Nachweis für eine reale sensorische Unterscheidungsfähigkeit eingestuft werden könnte" (Fisher 1935, S. 17).

Fisher hat eine starke Abneigung gegen die Zuweisung der Reihenfolge des Eingießens mit Tee und Milch für alle acht Tassen nach dem Zufallsprinzip mit gleicher Wahrscheinlichkeit, da „es gelegentlich auftreten würde, dass alle Tassen in derselben Reihenfolge befüllt werden, weil dies, abgesehen davon, dass es für die Testperson durch ein unerwartetes Ereignis verwirrend wäre, der Testperson den Vorteil der Beurteilung durch Vergleich nehmen würde" (Fisher 1935, S. 27). Nach Fishers Ansicht ist die Leistung der „Lady" in diesem Experiment unklar, obwohl andere gemeint haben, dass sie zwischen den beiden Zubereitungsmethoden unterscheiden könnte (Box 1990; Senn 2012).

Die Diskussion von Neyman (1950) ist ausführlicher als Fishers Original, obwohl er in erster Linie ein Design betrachtet, bei dem die Tassen als vier Paare präsentiert werden, wobei jedes Paar eine MZ- und eine TZ-Tasse umfasst und dies der Testperson bekannt ist. Neymans Modellierung konzentriert sich auf die Angabe zulässiger Hypothesen und die Bestimmung der Wahrscheinlichkeit, dass die einzelnen Klassifizierungen richtig sind, mit anderen Worten, der Kapazität der Testperson, die beiden Zubereitungen zu unterscheiden. Dies wurde auch auf mehrere Prüfer erweitert (eine „hypothetisch unendliche Population", Wrightson 1953).

3 Das Unterrichtsexperiment

Kontext

Das Experiment wurde im Rahmen einer Einführung in Medizinstatistik durchgeführt (im präklinischen Bachelorstudium der Medizin an der Universität Oxford). Das gesamte Modul besteht aus 16 Stunden Lehrveranstaltungen und umfasst Themen wie Konfidenzintervalle, Schätzung, Hypothesenprüfung, Korrelation, die Normal- und die Binomialverteilung (erstes Jahr), t-Test, Poisson-Verteilung, lineare Regression, Überlebensdaten und Studiendesign zur kritischen Bewertung von medizinischen Studien (zweites Jahr). Die Studierenden wurden in der Präsenzlehrveranstaltung in Gruppen von 25 bis 30 (etwa ein Sechstel der Jahreskohorte) unterrichtet.

Ziel des Moduls ist es, dass die Studierenden Verständnis für die Statistik angewandt auf biomedizinische Wissenschaften erwerben. Der Schwerpunkt liegt auf der Interpretation und der Anwendung von Methoden statt auf mathematischen Details. Die Übung, die hier beschrieben wird, wurde am Beginn des zweiten Studienjahres durchgeführt und war nicht Bestandteil ihrer summativen Beurteilung. Die meisten anderen Unterrichtseinheiten konzentrieren sich auf die statistische Analyse von Daten. Die Ergebnisse beziehen sich auf zwei Kohorten von Studierenden (2018 und 2019).

Beschreibung des Unterrichtsexperiments

Die Studierenden wurden gebeten, Gruppen von 4 bis 6 Personen zu bilden und einen Sprecher zu benennen, der die Rolle des „Hauptuntersuchungsleiters“ übernimmt. Die verbale Beschreibung des Ziels lautete:

> Muriel Bristol behauptet, einen Unterschied erkennen zu können zwischen Tassen von Tee, bei denen zuerst Milch eingegossen wurde und jenen, bei denen zuerst Tee in die Tasse gegossen wurde. Ronald Fisher ist skeptisch. Die Herausforderung besteht darin, ein Experiment zur Prüfung der Behauptung zu entwerfen.

Weitere Einzelheiten wurden auf ein Minimum beschränkt, obwohl die Stichprobengröße (Tassenzahl) festgelegt wurde:

> Sie können das Experiment beliebig gestalten, wie Sie dies für angemessen halten. Die einzige Einschränkung ist, dass aus praktischen Gründen ein Maximum von acht Verkostungen zugelassen ist.

Den Studierenden wurde gesagt, sie könnten jegliche Ausrüstung benutzen wie Standard-Tee-Zubereitungsgeräte, Tassen und Becher, Milch und Zucker (obwohl sie darauf aufmerksam gemacht wurden, dass sie das Experiment nicht wirklich durchführen würden). Die Gruppen hatten 20 Minuten, um die Aufgabe zu erörtern und einen Konsens zu erzielen über das zu verwendende Design und wie die Ergebnisse interpretiert werden könnten. Der Sprecher schlichtete Meinungsverschiedenheiten und hatte die Verantwortung für die endgültigen Entscheidungen. Die Gruppen wurden gebeten, zu diskutieren, sich zu einigen und ihre Antworten auf diese Fragen niederzuschreiben (siehe auch siehe das zusätzliche Lehrmaterial):

1) Wie würden Sie das Experiment gestalten? (Antworten Sie so detailliert wie möglich)
2) Welche Informationen würden Sie der Testperson (dem Verkoster) geben?
3) Welche Ergebnisse würden Sie brauchen, um die Behauptung der Testperson akzeptieren zu können?

Während dieser Diskussionsphase gingen zwei Lehrkräfte im Raum herum, um Fragen zu beantworten und anzuregen, dass die Gruppen Begründungen für ihre Entscheidungen angeben. Da diese Sitzung mehrmals stattfand (für die verschiedenen Gruppen), waren mehrere Lehrkräfte involviert. Zuvor hatte das Lehr-Team an einem Probelauf teilgenommen (anfänglich noch im Unklaren über die möglichen Lösungen), was half, die Fragen an die Studierenden zu klären und Übereinkunft zu erhalten, wie viel Hilfestellung man den Studierenden bieten sollte. Infolgedessen versuchten die anwesenden Lehrkräfte zu vermeiden, die Studierenden während der Sitzung in Richtung bestimmter Antworten zu beeinflussen. In der Diskussionsphase gab es keinen Internetzugang.

Am Ende dieser Phase wurden die Gruppen gebeten, ihre Entscheidungen der ganzen Klasse zu erklären, bevor man ihnen die Wahl, die Fisher getroffen hatte, und seine Begründung dazu skizzierte. Die Übung dauerte 30 bis 35 Minuten. Da die Zeit für zusätzliche Diskussion während der Sitzung begrenzt war, wurde eine schriftliche Beschreibung der verschiedenen von den Studierenden gewählten Optionen für das Experiment nach der Sitzung allen zur Verfügung gestellt. Feedback mittels eines anonymen Fragebogens wurde am Ende des Moduls eingeholt.

4 Ergebnisse

Studentische Antwort auf die Aufgabe

Die meisten Gruppen machten sich rasch mit der Aufgabenstellung vertraut und benötigten wenig Hilfe, um sie zu bearbeiten, obwohl einige eine Erinnerung brauchten, um einen Konsens zu erreichen und ihre Entscheidungen innerhalb der vorgeschriebenen Zeit zu protokollieren. Viele Diskussionen waren lebhaft, besonders wenn es innerhalb der Gruppe Uneinigkeit gab. Insgesamt gaben 61 Gruppen ihre Zustimmung, dass ihre Ergebnisse in die Studie aufgenommen werden (30 in 2018 und 31 in 2019).

Antworten auf Frage 1

Viele Antworten zeigten, dass sie sich der Notwendigkeit bewusst waren, die experimentellen Faktoren zwischen verschiedenen Verkostungen konstant zu halten. Dazu gehören:

- Brühzeit (oft numerisch angegeben);
- Menge und Art des Tees und der Milch (häufig numerisch angegeben);
- Zuckermenge (in der Regel kein Zucker);
- Temperatur (z. B. kontrolliert durch ein Wasserbad);
- Die Becher oder Tassen, in denen der Tee serviert wurde;
- Konsistenz des Rührens (z. B. „10 Umrührungen im Uhrzeigersinn“);
- Anzahl, Umfang und Standardisierung der zulässigen Schlucke (z. B. „sie muss an jeder Tasse einmal nippen und ausspucken“);
- Zeit zwischen den Verkostungen.

Einige Gruppen bemühten sich, die Auswirkungen äußerer Faktoren zu verringern, um die Glaubwürdigkeit zu verbessern. Vorschläge umfassten die Bereitstellung von Wasser zum „Auswaschen“ oder zur „Gaumenreinigung“ zwischen den Verkostungen; Darreichen des Tees an verschiedenen Tagen oder „jeden Morgen nach einem standardisierten Frühstück“ (vgl. Neyman 1950, S. 272); oder die Einschränkung sensorischer Eindrücke mittels einer Augenbinde, Nasenclip oder Geräuschunterdrückung durch Kopfhörer. Gruppen, die mehr als eine Testperson (einen Verkoster) einsetzen wollten (vgl. Wrightson 1953), wurden daran erinnert, dass sich die Behauptung einer besonderen Unterscheidungsfähigkeit nur auf eine Person bezieht.

Die andere Überlegung bezog sich darauf, wie die Verkostungen dem Typ MZ oder TZ zuzuordnen sind. Die meisten Gruppen waren sich einig, dass dies nach dem Zufallsprinzip erfolgen sollte, und viele Gruppen empfahlen eine doppelte Verblindung (des Versuchsleiters und des Verkosters), aber die Auswahl variierte. Die häufigste Wahl waren acht 50:50 zufällige Zuweisungen von MZ und TZ, aber etwa ein Drittel der Gruppen wählte das Fishersche Design. Fast alle Gruppen zogen mehrere Optionen in Betracht, bevor sie ihre endgültige Wahl trafen. Mehrere unterschiedliche Designs wurden weniger häufig vorgeschlagen (Tabelle 1; ich habe jeweils Namen beigefügt).

Das Vier-Paar-Design ist dasselbe wie das von Neyman (1950) vorgeschlagene. Das Subsampling-Design ist fast gleichwertig mit dem vollständig zufälligen Design. Wenn der Testperson das gewählte Design bekannt ist, sind Doppel-Tetrade und Pentade-Triade Sonderfälle innerhalb der „M + N“-Familie (Bi und Kuesten 2015).

Antworten auf Frage 2

Einige Gruppen sahen anfangs keinen Grund, warum irgendwelche Informationen an die Testperson (Verkoster) weitergegeben werden müssen. Als sie darauf hingewiesen wurden, dass zumindest genügend Informationen benötigt werden, um sicherzustellen, dass sie ihre Zustimmung zur Teilnahme geben kann, erkannten die meisten, wie dies mit ihrer Designentscheidung zusammenhängt.

Nahezu alle Gruppen entschieden, dass der Testperson nicht gesagt werden sollte, wie viele Tassen vom Typ MZ und wie viele vom Typ TZ sind, und zogen es vor zu sagen, dass sie nach dem Zufallsprinzip ausgewählt wurden (manchmal sogar, wenn dieses Design nicht gewählt worden war) oder dass jede Tasse entweder vom Typ MZ oder TZ sein könnte, ohne die Auswahlmethode zu nennen. Eine Ausnahme war die eine Gruppe, die das gepaarte Design wählte mit dem Hinweis, dass dieser Versuchsplan genau erklärt werden sollte, bevor das Experiment beginnt.

Antworten auf Frage 3

Viele Gruppen stellten die Verbindung zwischen dem gewählten Design und einer einfachen Methode zur Analyse her, die entweder die Binomialverteilung oder kombinatorische Überlegungen verwendet. Weitere Einzelheiten zu den Berechnungen für die jeweiligen Designs sind im zusätzlichen Lehrmaterial enthalten.

Design-Name	*Beschreibung*	*Häufigkeit*
Vollständig zufällig	Acht unabhängige 50:50 zufällige Zuordnungen von MZ oder TZ	28
Doppelte Tetrade	Vier MZ, vier TZ, in zufälliger Reihenfolge	20
Pentade-Triade	Fünf MZ, drei TZ (oder umgekehrt), in zufälliger Reihenfolge	3
Teilstichproben	Zufällige Auswahl von acht Tassen aus einer vorbereiteten größeren Gruppe (20 Tassen mit je 10 MZ und TZ, oder 50 bzw. 100 jeweils mit gleicher Anzahl der beiden Zubereitungen MZ und TZ.	3
Vier Paare (Neyman)	4 Paare, absichtlich zusammengestellt als MZ/MZ, TZ/TZ, MZ/TZ und TZ/MZ Oder „Wir würden ein paarweises Experiment machen … die Tassen in jedem Paar auf beide Arten zubereitet.“	3
Täuschung	„Sagen Sie der Testperson, dass sie acht Tassen Tee verkosten wird, wobei jede Tasse auf eine der beiden Arten zubereitet wird. In alle Tassen wird dann zuerst Tee eingefüllt.“ Oder „Geben Sie dem Verkoster acht gleiche Tassen Tee: alle entweder zuerst mit Milch eingefüllt oder zuerst mit Tee.“	2
Eingeschränkt zufällig	Nach dem Zufallsprinzip zugewiesen mit einer Einschränkung von mindestens zwei MZ und zwei TZ	1
Zufall + Kontrolle	Eine Kontrollprobe mit Tee ohne Milch, sonst zufällig MZ oder TZ	1

Tab. 1: Zusammenfassung vorgeschlagener Designs

Einige Gruppen argumentierten zum Beispiel, dass bei einem völlig zufälligen Design die Anzahl X korrekt identifizierter Tassen, wenn man rein zufällig „rät“, einer Binomialverteilung folgen würde ($n = 8$, $p = 0{,}5$). Da $P(X \geq 7) = 0{,}035$, könnten mindestens sieben richtige Vermutungen einen ausreichenden Nachweis liefern. Einige meinten, dass sie schon bei sechs richtigen Klassifikationen überzeugt wären, egal ob mit oder ohne Berechnung der zugehörigen Wahrscheinlichkeit. Gruppen, die sich für das doppelte-Viereinheit-Design entschieden, führten die Berechnung nach Fisher i. A. nicht durch, vielleicht, weil einige Gruppen, die dieses Design wählten, es nicht für nötig hielten, diese Entscheidung offenzulegen.

Feedback und Evaluation

In den Rückmeldungen aus dem ersten Jahr, in dem die Übung eingesetzt wurde, bewerteten die Studierenden die Übung als mäßig unterhaltsam, aber schätzten ihren Nutzen nicht (Durchschnittswerte von 3,3 bzw. 2,3 auf einer Skala von 1–5). In Freitextkommentaren merkten einige Studierende an, dass die Übung zu lang war und keinen klaren Bezug zum Rest des Kurses hatte.

Bevor die Übung im folgenden Jahr eingesetzt wurde, überprüften die Lehrkräfte diese Rückmeldungen und nahmen einige Änderungen an der Art und Weise vor, wie die Übung durchzuführen ist. Die Zeit, die den Gruppen für die Entscheidungsfindung eingeräumt wurde, wurde auf 20 Minuten begrenzt und die Gesamtzeit für die Übung einschließlich der Diskussion auf 35 Minuten.

Die Lehrkräfte wurden ermutigt, sich proaktiver einzubringen, um sicherzustellen, dass die Gruppen in der vorgegebenen Zeit zu klaren Entscheidungen kommen, und das Zeitlimit klarer zu kommunizieren, denn beim ersten Durchgang benötigten einige Gruppen für diesen Schritt 30 bis 40 Minuten, was bedeutete, dass die gesamte Übung fast eine ganze Unterrichtsstunde ausfüllte. In der abschließenden Diskussion stellten die Lehrkräfte eine klarere Verbindung zu anderen Teilen des Kurses her und eine Zusammenfassung der Schlussfolgerungen wurde nach der Sitzung ausgeteilt (im zusätzlichen Lehrmaterial enthalten). Bei zweiten Durchgang stiegen die Durchschnittswerte für Spaß und Nützlichkeit auf 4,3 bzw. 3,5, und die Freitextkommentare waren fast durchweg positiv, insbesondere in Bezug auf die Steigerung des Interesses am Thema.

5 Diskussion

Ziel dieser Übung war es, die Beschäftigung mit dem Thema Studiendesign und das Verständnis dafür in einem Modul der medizinischen Statistik zu verbessern. Die Übung war so konzipiert, dass die Bedeutung von Designentscheidungen und deren Auswirkungen auf die Analysemethoden aufgezeigt wird. Die Ergebnisse deuten auf eine Verbesserung des Engagements und des Verständnisses nach der Übung hin. Dies wurde durch die Wahrnehmungen der Lehrkräfte bestätigt, die die Übung durchführten.

Das Unterrichtsexperiment folgte dem Ansatz entdeckenden Lernens: Den Studierenden wurden nur wenige Informationen zur Verfügung gestellt, so dass sie von Anfang an ihre eigenen Entscheidungen treffen konnten. In diesem Sinne ähnelt sie einer problemorientierten Lernaktivität in kleinem Maßstab (Schmidt, Rotgans und Yew 2011). Aus der Lernperspektive ist der Prozess der Diskussion und Entscheidungsfindung mindestens so wichtig wie die Entscheidungen selbst.

Dies deckt sich mit Empfehlungen über die Bedeutung von interaktiven Übungen zur Entwicklung begrifflichen und statistischen Denkens für den Statistikunterricht für Studierende in angewandten Disziplinen (Bradstreet 1996; Garfield und Ben-Zvi 2009). Da nur wenig Zeit zur Verfügung steht, ist die Verwendung eines Beispiels wichtig, das ohne Vorwissen leicht verständlich ist, auch wenn das Teeverkostungsexperiment in vielerlei Hinsicht als untypisch angesehen werden kann. Dass der Kontext dieses speziellen Beispiels nicht direkt mit der medizinischen Wissenschaft zu tun hatte, schien kein Hindernis für die Teilnahme zu sein.

Die abweichenden Meinungen, die über die Wahl des Designs aufkamen, entsprechen einem Mikrokosmos von Diskussionen, der bei der Planung realer experimenteller Studien üblicherweise entsteht. Viele Gruppen von Studierenden waren in der Lage, unaufgefordert die Bedeutung mehrerer Punkte zu erkennen, die seit Fishers ursprünglicher Darstellung Debatten provoziert haben. Zum Beispiel interpretierten einige die ursprüngliche Formulierung der Behauptung der Testperson (der Lady) dahingehend, dass sie immer zwischen MZ und TZ unterscheiden könnte und nicht nur, dass ihre Wahrscheinlichkeit einer korrekten Klassifizierung größer als 0,5 wäre. Folglich sahen zwei Gruppen einen Vorteil darin, alle acht Tassen auf die gleiche Weise zuzubereiten (das „Täuschungs“-Design) und argumentierten, dass eine solche unerwartete Konstellation nicht stören sollte, wenn die Testperson perfekt unterscheiden kann.

Zwei praktische Punkte sollten beachtet werden, bevor man diese Übung durchführt. Erstens, wenn die Übung von verschiedenen Lehrkräften mit unter-

schiedlichem Erfahrungsstand durchgeführt wird, ist ein Probelauf sinnvoll, da dieses informelle Training dazu beitragen kann, Probleme aufzuzeigen, die bei der Durchführung der Übung auftauchen könnten. Zweitens, ist es hilfreich, im Voraus zu vereinbaren, wie stark die Lehrkräfte eingreifen sollten, anstatt die Studierenden alle Entscheidungen ohne Eingriffe alleine treffen zu lassen. Eine nützliche Strategie für die Diskussionen ist es, wenn die Lehrkraft den designierten Sprecher fragt, welche Entscheidungen getroffen wurden und dann den Rest der Gruppe fragt, ob alle mit diesen Entscheidungen einverstanden sind. Wenn es, wie es oft vorkommt, keinen vollständigen Konsens gibt, bietet dies der Gruppe eine Möglichkeit, die Gründe für ihre Entscheidungen zu klären.

Einige Einschränkungen bei der Evaluation dieser Übung sollten beachtet werden. Da es schwierig war, eine direkte Beurteilung des Verständnisses der Studierenden in das bestehende Beurteilungsschema einzufügen, entsprechen viele der vorgestellten Messgrößen eher einer Selbsteinschätzung der Studierenden und spiegeln keine objektive Messung statistischer Kompetenz. Es war nicht möglich, einen randomisierten Vergleich zwischen verschiedenen studentischen Gruppen durchzuführen, um die Wirksamkeit zu bewerten. Die Übung wurde nur im Präsenzunterricht eingesetzt. Obwohl es möglich ist, sie für den Fernunterricht anzupassen, zum Beispiel mit Hilfe von Online-Breakout-Diskussionsräumen, wurde dies nicht durchgeführt.

Diese Übung kann nur eine begrenzte Anzahl von Designfragen ansprechen, die Lehrkräfte üblicherweise behandeln möchten. Sie wäre wohl nicht geeignet für Themen wie optimale und faktorielle Versuchspläne. Sie war auch nicht dazu gedacht, Hypothesentests, *p*-Werte oder Missverständnisse in Bezug auf diese Themen zu unterrichten (Vallecillos 2000). Dennoch bot die Übung eine nützliche Möglichkeit für die Studierenden, sich mit Randomisierung und Verblindung vertraut zu machen für einen späteren Kurs über medizinische Studiendesigns (randomisierte kontrollierte Studien, Kohorten- und Fall-Kontroll-Studien). Sie bot auch eine Gelegenheit, einen wichtigen allgemeinen Punkt bezüglich Design und Analyse kennenzulernen – nämlich, dass die Analyse vom Design abhängig ist und daher beide von Anfang an mitbedacht werden sollten. Dies stärkt die Einsicht, dass statistisches Denken im gesamten Forschungsprozess und nicht nur in der Analysephase wichtig ist.

Erweiterungen sind möglich, um verschiedenen Zielgruppen gerecht zu werden. Die Übung könnte zum Beispiel verwendet werden, um die Konzepte der statistischen Macht (Power, Neyman 1950) und des Stichprobenumfangs sowie die hypergeometrische Verteilung zu demonstrieren. Unsere studentischen Kohorten hatten bereits Kenntnisse über die Binomialverteilung, sodass die Übung eine Gelegenheit bot, dieses Wissen zu vertiefen. Für Studierende, die mit Wahrscheinlichkeitsrechnung oder analytischen Fragen weniger vertraut sind, könnte die dritte Frage gestrichen werden, die Diskussion von Designfragen bliebe immer noch relevant.

Ein Artikel beschreibt die Nutzung dieses Experiments zur Motivation Bayesianischen Denkens anzuwenden, ohne zu erklären, wie dies am besten umgesetzt werden sollte (Lindley 1993). Für fortgeschrittenere Studierende könnte dies bedeuten, dass sie eine a priori-Verteilung auf dem Intervall [0; 1] oder [0,5; 1] spezifizieren für die Wahrscheinlichkeit, dass die Testperson eine korrekte Klassifizierung vornimmt, und diese Wahrscheinlichkeitsverteilung zusammen mit der Likelihood (bedingt auf das Ergebnis des Verkostungsexperiments) nutzen, um eine a posteriori-Verteilung zu berechnen. Dies könnte dazu verwendet werden, um die Auswirkungen der Wahl der a priori-Verteilung zu veranschaulichen, wenn die Studierenden eine unterschiedliche Einschätzung der Gültigkeit der ursprünglichen Behauptung (einer Fähigkeit, die Reihenfolge der Zubereitung der Teetassen zu erkennen) haben und in der Lage sind, dies durch geeignete a priori-Verteilungen auszudrücken.

Generell wird ein anwendungsorientierter Ansatz von den Studierenden eher geschätzt und unterstützt das Engagement für das Fach (Nolan und Speed 1999). Für diese studentische Gruppe sind reale Anwendungen in der Regel effektiver als Beispiele, die abstrakt sind oder Computersimulationen verwenden.

Meine Erfahrung mit Gruppenübungen zeigt, dass Beispiele, die für Statistiker intellektuell interessant sein könnten, als Lehrbeispiele bei Studierenden im Bachelorstudium nicht immer gut ankommen. Ein gewisses Maß an Versuch und Irrtum ist daher unvermeidlich, und kleine Anpassungen waren sowohl nach einem Probelauf als auch nach dem ersten Durchlauf dieser Lehrveranstaltung erforderlich.

Danksagung

Ich bedanke mich bei Lehrteam und anderen Forschern des Nuffield Dep. of Primary Care Health Science, Universität Oxford, die halfen, die Lehrübung zu testen und durchzuführen, sowie den Schülern, die teilnahmen und Feedback gaben. Die ethische Zulassung (CUREC Ref. R58986/RE001) wurde eingeholt, um anonyme Antwor-

ten aufzuzeichnen und zur Veröffentlichung zu nutzen. Nur die Ergebnisse derjenigen, die sich dafür entschieden haben, werden hier einbezogen.

Literatur

Anderson-Cook, C. M.; Dorai-Raj, S. (2001): An active learning in-class demonstration of good experimental design. In: *Journal of Statistics Education* 9, S. 1.

Appleton, D. R. (1990): What statistics should we teach medical undergraduates and graduates? In: *Statistics in Medicine* 9, S. 1013–1021.

Basu, D. (1980): Randomization analysis of experimental data: the Fisher randomization test. In: *Journal of the American Statistical Association* 75, S. 575–582.

Bennett, K. A. (2015): Using a discussion about scientific controversy to teach central concepts in experimental design. In: *Teaching Statistics* 37, S. 71–77.

Bi J.; Kuesten, C. (2015): Revisiting Fisher's ‚Lady Tasting Tea' from a perspective of sensory discrimination testing. In: *Food Quality and Preference* 43, S. 47–52.

Bland, J. M.; Altman, D. G.; Royston, J. P. (1990): Statisticians in medical schools. In: *Journal of the Royal College of Physicians of London* 24, S. 85–86.

Bradstreet, T. E. (1996): Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. In: *American Statistician* 50, S. 69–78.

Chadwick, S. J. D.; Dudley, H. A. F. (1983): Can malt whisky be discriminated from blended whisky? The proof. A modification of Sir Ronald Fisher's hypothetical tea tasting experiment. In: *British Medical Journal* 287, S. 1912–1913.

Cobb, G. W. (2007): One possible frame for thinking about experiential learning, In: *International Statistical Review* 75, S. 336–347.

Darius, P. L.; Portier, K. M.; Schrevens, E. (2007): Virtual experiments and their use in teaching experimental design. In: *International Statistical Review* 75, S. 281–294.

Easterling, R. G. (2004): Teaching experimental design. In: *American Statistician* 58, S. 244–252.

Fisher, R. A. (1935): The design of experiments. Edinburgh: Oliver and Boyd.

Fisher Box, J. (1990): R. A. Fisher and the design of experiments, 1922–1926. In: *American Statistician* 34, S. 1–7.

Garfield, J; Ben-Zvi, D. (2009): Helping students develop statistical reasoning: implementing a statistical reasoning learning environment. In: *Teaching Statistics* 33, S. 72–77.

Gore, S. M. (1984): Teaching experimental design: Prescribed by a medical statistician. In: *Journal of the Royal Statistical Society D* 33, S. 243–247.

Gridgeman, N. T. (1959): The lady tasting tea, and allied topics. In: *Journal of the American Statistical Association* 54, S. 776–783.

Hiebert, S. M. (2007): Teaching simple experimental design to undergraduates: Do your students understand the basics? In: *Advances in Physiological Education* 31, S. 82–92.

Lindley, D. V. (1993): The analysis of experimental data: The appreciation of tea and wine. In: *Teaching Statistics* 15, S. 22–25.

MacDougall, M; Cameron, H. S.; Maxwell, S. R. J. (2020): Medical graduate views on statistical learning needs for clinical practice: A comprehensive survey. In: *BMC Medical Education* 20, S. 1.

Montangero, S.; Vittone, F.; Olderbak, S.; Wilhelm, O. (2018): Exploration of experimental design and statistical methods using the stick-on-the wall spaghetti rule. In: *Teaching Statistics* 40, S. 40–45.

Morton, R. (1975): On the efficiency of Fisher's tea-tasting designs. In: *Journal of the Royal Statistical Society B* 37, S. 49–53.

Neyman, J. (950): First course in probability and statistics. New York, NY: Henry Holt and Company.

Nolan, D; Speed, T. P. (1999): Teaching statistics theory through applications. In: *American Statistician* 53, S. 370–375.

Pyott, L. (2021): Tennis anyone? Teaching experimental design by designing and executing a tennis ball experiment. In: *Journal of Statistical Data Science Education* 29, S. 22–26.

Salsburg, D. (2001): The Lady tasting tea: How statistics revolutionized science in the twentieth century. New York, NY: Henry Holt and Company.

Schmidt, H. G.; Rotgans, J. I.; Yew, E. H. J. (2011): The process of problem-based learning: what works and why. In: *Medical Education* 45, S. 792–806.

Senn, S. (2012): Tea for three: Of infusions and inferences and milk in first. In: *Significance* 9, S. 30–33.

Vallecillos, A. (2000): Understanding of the logic of hypothesis testing amongst university students. In: *Journal für Mathematikdidaktik* 21, S. 101–123.

Wild, C. J. (1994): Embracing the ‚wider view' of statistics. In: *American Statistician* 48, S. 163–171.

Wrightson, R. F. (1953): The theoretical basis of the therapeutic trial. In: *Acta Genetica et Statistica Medica.* 4, S. 312–343.

Zusätzliches Lehrmaterial

onlinelibrary.wiley.com/doi/10.1111/test.12287.

Wiley *(Teaching Statistics)* hat dem Autor die Publikation einer deutschen Fassung in Stochastik in der Schule genehmigt.

Anschrift des Verfassers

Thomas R. Fanshawe
Nuffield Dep. Primary Care Health Sciences
University of Oxford, Oxford OX2 6GG, UK
thomas.fanshawe@phc.ox.ac.uk