Rezension: Bernhard Hauser und Winfried Humpert:

Signifikant? Einführung in statistische Methoden für Lehrkräfte.

Zug: Klett und Balmer Verlag 2009

MANFRED BOROVCNIK, KLAGENFURT

Zusammenfassung: *Die Autoren wenden sich an Personen, die an der Grundschule oder in der Sekundarstufe I unterrichten oder in entsprechender Ausbildung stehen. Grundidee ist, mit wenig formalem Aufwand Prinzipien empirischer Forschung und Schlüsselkonzepte wie das der statistischen Signifikanz zu erklären. Dem Werk ist eine CD mit Übungen und Demonstrationen angeschlossen.*

Zum Inhalt des Buches

1 Wissenschaft und ihre Arbeitsweisen (8)

Ein wesentliches Ziel der Autoren ist es zu vermitteln, wie man die Qualität von Untersuchungen (wie PISA) beurteilt, wie man Leistungstests (Klassencockpit, Intelligenztests) versteht, oder wie man selbst kleine Untersuchungen zur Beurteilung des Unterrichts durchführt. Dazu führen sie in Grundbegriffe empirischer Forschung ein: Wie präzisiert man Hypothesen, wie beachtet man Maßeinheiten und Bezugssysteme und wie erkennt man den Unterschied zwischen der Prüfung von Unterschieden bzw. von Zusammenhängen?

2 Wissenschaftliche Verfahren der Datenerhebung (13 Seiten)

Zur Überprüfung von Hypothesen sind Daten zu erheben; an Methoden werden Beobachtung, Gestaltung von Fragebögen und Experiment vorgestellt. Bei der Elimination von Zufallseinflüssen fokussieren die Autoren auf die Ausschaltung von Störfaktoren durch die Versuchsbedingungen.

3 Grundlagen des Messens (18)

Die Autoren behandeln den Ablauf von Forschung und gehen auf die Formulierung von Hypothesen, auf Untersuchungseinheiten sowie Typen von Merkmalen ein. Bei der Messung sprechen sie die üblichen Gütekriterien (Validität, Reliabilität, Objektivität) an und stellen den Zusammenhang von Skalenniveaus (nominal, ordinal, Intervall, Verhältnis) mit erlaubten Rechenoperationen dar.

4 Beschreibende Statistik (46)

Breiten Raum nimmt die Darstellung von Daten in Tabellen und Graphen ein. Auf die Realisierung in EXCEL wird eingegangen. Typen von Verteilungen werden schematisch illustriert. In einem Schülerexperiment wird der Mittelwert als Ausgleichswert erklärt; daneben wird der Median als „mittlerer Rangplatz“ eingeführt. Als Maße der Streuung werden Varianz, Standardabweichung und Spannweite besprochen. Dabei wird die Formel der Varianz „vereinfacht“ – im Nenner steht die Anzahl der Fälle.

Ausführungen zur Normalverteilung ergänzen das Kapitel zur Beschreibenden(!) Statistik; inferentielle Statistik wird durch drei Fragen charakterisiert:

- „Gilt ein Ergebnis einer Stichprobe für die gesamte Population?“
- „Sind die Unterschiede […] zufällig oder sind es echte Unterschiede?“
- „Sind Zusammenhänge […]?“

Die Standardisierung von Daten einer Normalverteilung wird als Prozentrang interpretiert.

5 Schließende Statistik (30)

Die Generalisierbarkeit von Stichprobenergebnissen hängt – nach den Autoren – wesentlich von drei Faktoren ab: Konkurrierende Hypothesen, Anlagefehler, Zufalls- oder Stichprobenfehler.“ (S. 101). Als konkurrierende Hypothesen werden Kovariable (können die Zielvariable beeinflussen) und ein allfälliger Placebo-Effekt (Wirkung auch bei Scheinbehandlung) herangezogen. Anlagefehler sind Verzerrungen der Messung, welche durch den Erhebungszeitpunkt und das Verfahren der Messung (z. B. Sensibilisierung) entstehen.

Jede Testgröße hat als Zufallsvariable eine Standardabweichung, die in der Literatur auch als Standardfehler bekannt ist. Diesen Standardfehler ziehen die Autoren als zentralen Bezug durch. Der Zentrale Grenzwertsatz (ZGS) wird zitiert; damit wird die näherungsweise Normalverteilung der Mittelwerte von Stichproben begründet. Für den Standardfehler des Mittelwerts wird eine Formel angegeben. Damit wird der Test für den Mittelwert (ein- und zweiseitig) erläutert. Abschließend werden Konfidenzintervalle besprochen.

Schlüssel für die Interpretation statistischer Tests im Buch ist die Signifikanz: „Signifikant ist ein Ergebnis, wenn die Irrtumswahrscheinlichkeit klein genug ist.“ (S. 108).

6 Ausgewählte Verfahren (50)

Auf 50 Seiten werden die üblichen Methoden der Angewandten Statistik präsentiert. Genauer besprochen werden der t-Test, auch für zwei Stichproben (abhängig und unabhängig), der F-Test für Varianzen, die Prüfung von Korrelationskoeffizienten auf Signifikanz, sowie einige nicht-parametrische Verfahren. Im Zusammenhang mit dem Mittelwertvergleich wird auch auf die Effektstärke eingegangen; das ist eine Schätzung der Stärke der Unterschiede, die berechnet wird, wenn das Ergebnis des Tests signifikant ist.

7 Testen und Leistungen messen in der Schule (20)

Ein fiktives Verfahren zur Entwicklung eines Erhebungsinstruments wird beschrieben; einige größere Projekte zur Leistungsfeststellung wie PISA werden kursorisch vorgestellt.

8 Interpretieren (5)

Die enorme Bedeutung der Interpretation von empirischen Studien wird hervorgehoben.

Einige große Linien

Wissenschaftliches Arbeiten – Anwendungen

Die Tücke von Anwendungen zeigt sich von Anbeginn, etwa in der Präzisierung der Aussage „Meine Klasse war heute morgen wieder so unruhig“ durch „Ein [Kind] hat mehrere Male (z. B. zweimal pro Unterrichtsstunde) mit dem Banknachbarn geredet und dabei unterrichtsfremde Inhalte getauscht.“ (S. 16). Dem Rezensenten fallen ad hoc Lärmpegel, emotionale Beteiligung, Dauer der Störung, Auswirkung auf andere Kinder ein. Diese Merkmale sind schwieriger zu messen, ermöglichen aber eine bessere Präzisierung und Überprüfbarkeit der vorgegebenen Aussage.

Bei der Illustration von Zusammenhangshypothesen: „Je länger die Schulzeit, desto höher die Bildung“ (S. 19/20) werden die Merkmale als *Anzahl der Schuljahre* sowie als *Art des Abschlusses* operationalisiert. Damit wird Bildung abhängig von der Schuldauer definiert. Will man eine Hypothese über den Zusammenhang von Bildung und Schulzeit überprüfen, muss man aber Bildung unabhängig von der Schuldauer definieren und messen – etwa durch einen genormten Test.

Dass beim Experiment die randomisierte Aufteilung von Versuchs- und Kontrollgruppe gar nicht angesprochen wird, ist als grobes Versäumnis zu werten. Gerade daran mangeln viele empirische Untersuchungen und das verursacht, dass die Ergebnisse wertlos werden.

Angesichts des Ziels, große empirische Studien lesen zu lernen, ist es bemerkenswert, dass die Autoren auf die Besprechung von Konstruktvariablen verzichten. Indirekt greifen sie diesen Komplex in Kapitel 7 auf, ohne jedoch auf dessen überragende Bedeutung für die empirische Forschung einzugehen.

Bei *Stichprobe* nehmen die Autoren wiederkehrend auf eine einzelne Schulklasse Bezug: Stichprobe ist ein Terminus technicus und sollte auf *zufällig* ermittelte Teilmengen beschränkt bleiben; die Verwirrung um diesen Begriff stellt eine wesentliche Ursache für die Fehleinschätzung von empirischen Studien sowie der Relevanz der daraus resultierenden Ergebnisse dar.

Beschreibende Statistik

Boxplots fehlen unverständlicherweise; das muss als grober Mangel eines praktisch orientierten Lehrbuches empfunden werden. Die Beschreibung von Verteilungstypen ist verbesserungsbedürftig, so sind die Flächen der Verteilungen verschieden.

Es wird lamentiert, dass *Laien* dazu neigen, „dem Durchschnitt eine zu hohe Wichtigkeit zuzuschreiben.“ (S. 77). Experten machen doch denselben Fehler; etwa sollte man in der PISA-Studie auch den oberen 10 %-Punkt der Leistungsmessung in verschiedenen Ländern vergleichen. Der Vergleich von *Verteilungen* aufgrund *einer* Kennziffer wird immer zu kurz greifen.

Die Notwendigkeit eines Maßes der Streuung wird durch ein Beispiel (S. 81) begründet, in welchem der Mittelwert eine schlechte Beschreibung der Daten liefert. Zielführender wäre es, ein Beispiel zu nehmen, in dem der Mittelwert versagt, zwei Datensätze miteinander zu vergleichen. Damit die Standardabweichung deskriptiv sinnvoll interpretiert werden kann, muss der Mittelwert eine gute Beschreibung der Daten liefern.

Normalverteilung

Es wird ein naives Verständnis für die Normalverteilung aufgebaut: „In der Natur kommt [sie] recht häufig vor. [… Sie] entsteht, wenn ein Merkmal bei sehr vielen Merkmalsträgern gemessen wird. Für die Entstehung müssen viele Faktoren, die unabhängig voneinander sind, einwirken. Der Einfluss dieser Faktoren muss dabei unsystematisch und zufällig sein. Ein Beispiel hierfür ist die Intelligenz.“ (S. 88)

In der Natur kommt keine Normalverteilung vor; diese entsteht – nach dem zentralen Grenzverteilungssatz

(ZGS) *ideell* durch additive Überlagerung vieler Einflussgrößen. Allein durch Messung „bei sehr vielen Merkmalsträgern“ entsteht keine Normalverteilung. Das Konstrukt Intelligenz ist normalverteilt, weil der Test so kalibriert ist, dass bei der Normalbevölkerung eine Normalverteilung herauskommt. Es handelt sich somit um ein Artefakt der Testung.

Der ZGS rechtfertigt Normalverteilung; allerdings gilt dies nicht für die Verteilung der Merkmale, wie fälschlicherweise oft unterstellt wird, sondern erst für die Verteilung von (aus einer *zufälligen* Stichprobe) abgeleiteten Kennziffer wie dem Mittelwert. Die meisten Kennziffern sind durch eine Summe (über alle Objekte der Stichprobe) definiert und daher annähernd normalverteilt – vorausgesetzt, die Stichprobe ist zufällig.

Die Beurteilung der Differenz von *Mittelwerten* kann daher immer auf die Normalverteilung rekurrieren; dazu bedarf es keiner Normalverteilung des Merkmals (und die fehlt i. A. auch). Allerdings ergeben sich bei der *Interpretation* der Unterschiede – sofern sie signifikant sind – erhebliche Vorteile, wenn das Merkmal selbst normalverteilt ist. Dann stellt ein Unterschied im Erwartungswert (bei gleicher Varianz) einfach eine Verschiebung der ganzen Verteilung nach oben oder unten dar, sodass man den Unterschied zweier Verteilungen durch *eine* Kennziffer erfassen kann. In vielen Studien (auch PISA) ist die Verteilung des Merkmals schief und in den zu vergleichenden Gruppen unterschiedlich, sodass eine Reduktion der Unterschiede auf Unterschiede im Mittelwert am Problem vorbei geht.

Zufällige Stichproben – Vergleich mit Zufall

Wie man die Grundfragen der statistischen Inferenz ohne den Begriff der zufälligen Stichprobe erarbeiten kann, ist dem Rezensenten ein Rätsel. Entsprechend werden gegebene Schulklassen mit dem nationalen Schnitt verglichen und auf signifikante Abweichungen davon geprüft. Bereits ein oder zwei Schüler, die Besonderheiten nach oben oder nach unten aufweisen, machen einen solchen Vergleich obsolet und führen das „Niveau“ vielleicht auf bestimmte pädagogische Maßnahmen zurück, die durchgeführt worden sind.

Stichproben werden mit einer Analogie zur Käseproduktion motiviert (S. 99); da wird die Masse vorher gut durchmischt und ist daher homogen – es ist dann (fast) egal, wo man die Teilmenge entnimmt. Das ist in der Statistik anders; damit die untersuchte Stichprobe als repräsentativ für die Population gelten kann, ist es notwendig, besondere Vorkehrungen zu treffen, wie man die Stichprobe zieht. Zufälliges Ziehen einer Stichprobe ist das Mittel der Wahl; in der Meinungsforschung weicht man auf so genannte Quotenstichproben aus. Jedenfalls ist die im Buch fortdauernd anzutreffende Gleichsetzung einer Schulklasse, die untersucht werden soll, mit einer Stichprobe aller vergleichbaren Schüler keineswegs zulässig.

Zum Zufallsfehler äußern sich die Autoren so (S. 105): „Selbst wenn wir eine Studie ganz seriös geplant haben, ist es möglich, dass ein Zusammenhang […] nichts anderes als ein Glückstreffer ist. […] Dieser Zusammenhang besteht […] ganz einfach deshalb, weil die Stichprobe (zufällig) sehr atypisch und wenig repräsentativ ist.“

Eine solche Feststellung kann man aber nur treffen, wenn man den Auswahlvorgang der Stichprobe über den Zufall kontrolliert, also zufällig auswählt. (Oder die Zuordnung von Personen zu Behandlungs- und Kontrollgruppe durch Zufall steuert.) Aber davon ist im ganzen Buch nirgends die Rede. Es werden durchgehend Gelegenheitsteilmengen untersucht und als Zufallsstichproben angesehen. Ein Vergleich mit dem Zufall kann nur erfolgen, wenn die Stichprobe selbst zufällig gezogen wurde.

Die übliche Fehlinterpretation von Signifikanz

Bei der „Irrtumswahrscheinlichkeit“ wird außer Acht gelassen, dass es sich um eine *bedingte* Wahrscheinlichkeit – bedingt auf das gewählte Modell (inklusive der Nullhypothese) – handelt.

„Signifikant heißt bedeutsam und meint, dass ein Ergebnis mit einer bestimmten (hohen) Wahrscheinlichkeit nicht zufällig zustande gekommen ist.“ (S. 116). Hier tritt eine „Umkehrung“ der bedingten Wahrscheinlichkeiten zutage: Wird E (oder „extremeres“) beobachtet, so hat man bei Signifikanz zum Niveau α: $P(E|H_0) < \alpha$. Die Gleichsetzung $P(H_0|E) = P(E|H_0)$ und der daraus folgende Schluss $P(H_0|E) < \alpha$ ist aber *falsch*. Die Wahrscheinlichkeit, dass die „Nullhypothese“ dennoch zutrifft, wenn ein beobachtetes Ergebnis signifikant ist, kann man erst mit weiteren Annahmen berechnen.

„Signifikant“ und „bedeutsam“ muss man sorgsam trennen, will man die Begriffe wirklich aufbauen. Im Fall von signifikanten Ergebnissen gilt es, den Effekt – also die Größe tatsächlicher Unterschiede – zu schätzen. Das tun die Autoren später, umso mehr überraschen solche naiven Gleichsetzungen.

Die übliche Fehlinterpretation des Konfidenzniveaus schließt lückenlos an die falsche Interpretation des Signifikanzniveaus an: „Daraus können wir folgende Schlussfolgerung ziehen: […] können wir mit einer

95%igen Sicherheit davon ausgehen, dass der Populationsmittelwert […] innerhalb des Konfidenzintervalls von 3,00 bis 3,33 […] liegt." (S. 124). Es kann nicht oft genug wiederholt werden, aber die Übertragung der Überdeckungswahrscheinlichkeit aus dem Verfahren zur Gewinnung von Konfidenzintervallen auf die Überdeckungswahrscheinlichkeit einzelner Intervalle – so wünschenswert das wäre – entbehrt jeder methodologischen Begründung.

Effektstärke – gut, aber kein Allheilmittel

Stellt man signifikante Unterschiede (Zusammenhänge) fest, so gilt es in einer zweiten Runde, deren Größe zu beurteilen. Die direkte Differenz der Mittelwerte kann im Kontext beurteilt werden. Die Effektstärke beurteilt die „Verschiebung", indem sie den Unterschied mit der Standardabweichung der einzelnen Verteilungen normiert.

Allerdings setzt die Interpretation der Effektstärke (*ES*) normal verteilte Merkmale (mit gleichen Varianzen) voraus. Ansonsten macht ein „normierter" Mittelwertunterschied keinen Sinn. Wenn Normalverteilungen im Spiel sind, so gibt die Effektstärke eine geschätzte Verschiebung der Verteilung von der einen zur anderen Versuchsbedingung an. In vielen Zusammenhängen kann man die Normalverteilung wohl für die Testgrößen (die Mittelwerte) in Anspruch nehmen, die Merkmale folgen jedoch anderen Verteilungen. Das erschwert die Interpretation der Effektstärke.

Da hilft dann nur die Macht eines Tests, das ist die Wahrscheinlichkeit des Tests, die Nullhypothese abzulehnen in Abhängigkeit von Werten für die Parameter aus der Alternativhypothese. Implizit haben Überlegungen zur Macht zu den vagen Interpretationsregeln geführt, wonach etwa $ES = 0{,}8 \cdot \sigma$ als starker Effekt aufgefasst wird.

Effektstärke ist ein Schritt in die richtige Richtung, aber kein Ersatz für die Macht eines Tests, welche ihre Interpretierbarkeit behält, wenn man von der Normalverteilung oder den gleichen Varianzen als Voraussetzungen abweicht, was häufig der Fall ist.

Für den Korrelationskoeffizienten versagt das Konzept der Effektstärke entsprechend, weil zwar die Verteilung unter $\varrho = 0$ (keine Korrelation) symmetrisch aber für $\varrho \neq 0$ schief ist. Der Feststellung der Autoren (S. 159): „Für die Korrelation muss die Effektstärke nicht zusätzlich noch errechnet werden, weil der Korrelationskoeffizient per Definition ein Effektstärkemaß darstellt […]" muss man entgegen halten: Natürlich ist es wichtig zu wissen, welche Chance der Test bietet, die Nullhypothese abzulehnen, wenn man als tatsächliche Korrelation etwa $\varrho = 0{,}3$ unterstellt.

Bei den Beispielen ist bedauerlich, dass die Autoren ihr eigenes Konzept vernachlässigen; sie berechnen in keinem der Fälle die Effektstärke.

Formeln vermeiden oder vereinfachen

Als Folge der anfänglichen Vereinfachung der Festlegung der Standardabweichung müssen alle Formeln neu geschrieben werden und sind dann eigentlich komplizierter. Als Leser gerät man in Schwierigkeiten, wenn man Standardtexte liest. Die Vermeidung von griechischen Buchstaben macht sich schmerzhaft bemerkbar. Damit wird die Unterscheidung zwischen Parametern und Schätzgrößen bzw. Schätzwerten erschwert.

EXCEL und die CD

Bei der Behandlung der Datenanalyse mit EXCEL werden wichtige Dinge genannt, aber nicht immer erklärt. Schlimmer jedoch wiegt, dass die Anleitungen Lücken haben; Nebensächliches wird ausgeweitet, was den Überblick stört. Es mag verwundern, dass die Autoren die Möglichkeiten von EXCEL ignorieren, Daten zu simulieren und damit Konzepte verständlich zu machen.

Die beiliegende CD enthält neben Übungsbeispielen auch sechs Animationen. Hervorzuheben sind die Animationen zur Verteilung der Augensumme mehrerer Würfel sowie zur Stichprobenverteilung von Mittelwerten (in Javascript, nicht in EXCEL), die den ZGS illustrieren sollen. Der Begleittext dazu ist wenig hilfreich, die Graphiken sind verbesserungsbedürftig. Eine Einbindung in den Buchtext zur *Entwicklung* der Begriffe fehlt.

Resümee

Der Rezensent betrachtet wesentliche Ziele der Autoren wie das der verständlichen Erklärung wissenschaftlicher Ergebnisse aus Studien als gescheitert. Möglichkeiten einer Tabellenkalkulation wurden kaum genutzt. Vage Formulierungen und Ungenauigkeiten stören. Die durchgehende Gleichsetzung von Gelegenheitsdaten mit zufälligen Stichproben versetzt der Absicht, empirische Forschung verstehen und bewerten zu lernen, den k.o.-Schlag. Selbst wenn man das Zielpublikum vor Augen hat, ist vom Buch abzuraten.

Anschrift des Verfassers

Prof. Dr. Manfred Borovcnik
Alpen-Adria-Universität Klagenfurt
Institut für Statistik
manfred.borovcnik@uni-klu.ac.at