Vom Sinn und Unsinn der Lehrevaluationen an deutschen Hochschulen – Über den Missbrauch von Statistik


Frank Brand

Prof. Dr.-Ing. Dipl.-Phys.
Frank Brand
Professor für Mathematik und Statistik
Fachbereich 1 Wirtschaftswissenschaften
Hochschule für Wirtschaft und Recht Berlin
frank.brand@hwr-berlin.de
www.hwr-berlin.de/service/kontakt/personen/detail/frank-brand/

Seit einigen Jahren werden viele Details des Hochschullebens einer Bewertung unterzogen – so vor allen Dingen auch die Lehrveranstaltungen. Dieses Verfahren trägt den Namen Lehrevaluation. Zur Auswertung der Daten werden statistische Fachkenntnisse benötigt. Bei heutzutage durchgeführten Evaluationen werden leider fast immer Mittelwerte von den erhobenen Daten berechnet, obwohl das für diesen Datentyp nicht erlaubt ist.

Wir müssen uns fragen: Auf welche Bereiche können sich die derart gewonnenen Evaluationsergebnisse auswirken? Da ist zum einen der in vielen Hochschulen vorhandene Leistungsrat, der in Ermangelung anderer Daten auch Evaluationsergebnisse im Rahmen der Festsetzung von Leistungszulagen für W-Besoldete verwendet. Darüber hinaus kann die Weiterbeschäftigung von Lehrbeauftragen abhängig von ihren persönlichen Evaluationsergebnissen sein. Nicht zuletzt können Evaluationen Auswirkungen auf Rankings der Hochschule haben.

Nicht nur diese wenigen angeführten Punkte, sondern auch die Verfahrensweisen ausgerechnet an wissenschaftlich orientierten Institutionen rechtfertigen die Forderung nach wissenschaftlicher Korrektheit!

Statistische Ausgangssituation

Wenn es um Veröffentlichungen in der Ökonomie (und nicht nur da) geht, wird in unglaublicher Weise auf statistische Auswertung fokussiert – man hört auch schon mal den Hinweis eines Herausgebers: „Wir können den Artikel nur dann publizieren, wenn noch etwas mehr Statistik hineingesteckt wird.“ Dort wie auch bei Evaluationen in Hochschulen wird hingenommen, dass mit unangemessenen oder teilweise sogar falschen Methoden Daten ausgewertet werden und dann inhaltlich natürlich nur unrichtige Schlüsse gezogen werden können.

Besonders peinlich ist, dass Studierende in einer Statistikvorlesung ja bereits in der zweiten Semesterwoche damit vertraut gemacht werden, welche Bedeutung sogenannte Skalierungen in der Statistik für die Erstellung und Auswertung von Fragebögen haben. Sie sind dann in der Lage, die Unstimmigkeiten praktizierter Evaluationsauswertungen zu verstehen.

Worum geht es? Besinnen wir uns: Zahl ist nicht gleich Zahl – jedenfalls nicht in der Statistik. Dazu einige Beispiele.

Mit einer Telefonnummer kann man nichts weiter tun, als diese einer Person zuzuordnen oder eben nicht. Sachlogisch nicht sinnvoll ist es, verschiedene Telefonnummern verschiedener Nutzer mathematisch vergleichen zu wollen. Die Summe zweier Telefonnummern zu bilden, ist sachlogisch jedenfalls nicht zulässig. Betrachten wir nun Schulnoten, dann kommt dabei der Verwendung von Zahlen keine besondere Bedeutung zu – ebenso gut kann man nämlich auch die Buchstaben A bis E (oder, wie in vielen Teilen der USA

Die Mittelwertbildung in der Auswertung von Fragenbögen liefert keine interpretierbaren Informationen. Neben der Kritik an der aktuellen Vorgehensweise wird aufgezeigt, wie eine korrekte Auswertung aussehen sollte.

üblich, die Buchstaben A, B, C, D und F) verwenden. Dieses Vorgehen ist möglich, da sowohl den Zahlen als auch den Buchstaben eine Ordnung innewohnt. Sachlogisch ist es daher möglich (und sinnvoll), Schulnoten in eine Reihenfolge zu bringen. Gibt es nun Operationen, die bei Schulnoten nicht erlaubt sind, aber zum Beispiel bei Temperaturangaben in Grad Celsius durchgeführt werden dürfen?

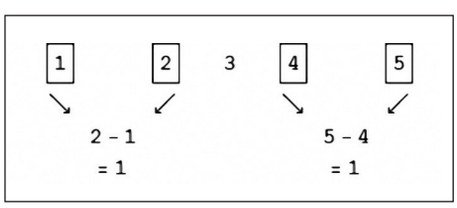

Abbildung 1: Nicht erlaubte Bildung von Differenzen bei Schulnoten

Betrachten wir die beiden Differenzen zwischen den Schulnoten fünf und vier bzw. zwei und eins. In beiden Fällen ist die Differenz rechentechnisch gleich eins (siehe Abbildung 1). Dass die beiden Differenzen sachlogisch die gleiche Bedeutung haben, muss aber verneint werden. Während die Noten eins und zwei die Abkürzungen für sehr gut und gut darstellen, ist man mit einer fünf durch das Examen gefallen, während man es mit einer vier noch besteht. Vergleicht man dagegen unterschiedliche Temperaturwerte, ist die Differenzbildung erlaubt.

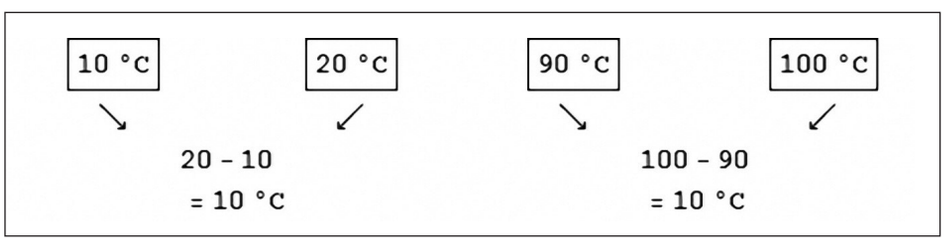

Abbildung 2: Differenzbildung von Temperaturen ist erlaubt

Wir sehen also: Die verwendeten Zahlen sind nichts weiter als Labels, die ebenso gut durch Buchstaben ersetzt werden können. Was bei Telefonnummern nicht erlaubt ist, darf z. B. bei Zahlen, die Temperaturangaben in Grad Celsius darstellen, aber getan werden; denn die beiden Differenzen von 10 Grad Celsius (siehe Abbildung 2) haben sachlogisch die gleiche Bedeutung!

Wozu sind all diese Betrachtungen und Unterscheidungen gut? Die Zuordnung verschiedenen Zahlenmaterials zu verschiedenen erlaubten Auswertungen hat weitreichende Bedeutung. Und eben nicht nur in der Statistik. Für Schulnoten dürfen keine Durchschnitte gebildet werden – für Temperaturen und Preise ist dies hingegen erlaubt. So kann man einer (fälschlicherweise berechneten) Durchschnittsnote von sagen wir 2,45, die durch die Mittelung von ganzzahligen Einzelnoten entstanden ist, überhaupt keine sachlogische Bedeutung beimessen. Denn niemand könnte erklären, welche Bedeutung dem Nachkommateil von 0,45 in der Note 2,45 zuzuordnen bzw. wie diese sprachlich von den ganzzahligen Noten noch zu unterscheiden wäre; noch weniger wären diese Differenzen zwischen Durchschnittsnoten, z. B. 2,45 und 2,72, erklärbar.

Kritische Würdigung der bisher praktizierten Vorgehensweise

Wir stellen fest, dass die Schulnoten von eins bis sechs als Ausprägungen eines ordinal skalierten Merkmals anzusehen sind. Für ordinal skalierte Merkmale ist als Lageparameter ausschließlich die Berechnung des Medians erlaubt. Die Berechnung des arithmetischen Mittelwertes ist nicht erlaubt und führt nicht zu sinnvoll interpretierbaren Ergebnissen. Zusätzlich zur statistischen Unmöglichkeit wird bei der Berechnung einer Note wie 2,45 durch die Angabe von zwei Nachkommastellen eine Genauigkeit von einem Hundertstel impliziert. Wer sich mit Messproblemen auskennt, weiß, dass dies im Kontext der Lehrevaluation grober Unfug ist.

Ein weiteres Resultat der Datenerhebung im Rahmen der Lehrevaluation ist die sogenannte Profillinie für eine spezielle Lehrperson im Vergleich (auch hier wieder) der gemittelten Daten über alle Lehrenden des gleichen Faches. Diese Vorgehensweise ist natürlich mit dem oben Gesagten ebenso unsinnig, denn auch diese Differenz kann mit keiner sinnvollen Bedeutung assoziiert werden. Es bleibt festzuhalten, dass Methoden zur Anwendung auf kardinale Daten (bspw. Durchschnittsbildung) nicht auf ordinale Daten anwendbar sind.

Korrekte Auswertung

Nehmen wir an, dass die den Studierenden vorgelegten Fragen tatsächlich die Qualität der Lehre messen, und wählen

solche Fragen aus, bei denen Schulnoten als Antworten erlaubt sind, so wird der Median bestimmt als das mittlere Element in der zuvor geordneten Datenreihe. Liegt eine Datenreihe mit einer geraden Anzahl von Werten vor, so wird der Median als Mittelwert aus den zwei mittleren Werten berechnet. Aus diesem Grund kann dieser also auch einen nicht ganzzahligen Wert annehmen, z. B. (5 + 4) / 2 = 4,5, wenn die beiden mittleren Werte 4 und 5 sind. Die einzige zulässige Interpretation des Medians liegt darin, dass sowohl links als auch rechts vom Median die gleiche Anzahl von Daten liegen – und zwar unabhängig von ihren reinen Zahlenwerten! Die Häufigkeiten der gegebenen Anworten sollten grafisch in einem Histogramm dargestellt werden, welches neben der Gesamtzahl der Antworten die prozentualen Verteilungen ebenso zeigt wie den Median und die Spannweite, also die Differenz zwischen minimalem und maximalem Wert (siehe Abbildung 3).

Werden von allen Fragen in einem Evaluationsbogen die Mediane bestimmt und grafisch dargestellt, kommt man zu den in Abbildung 4 dargestellten Kurven. Der durchgezogene Linienzug ist dabei der Median der Medianwerte aller Lehrender, der gestrichelte Linienzug stellt jeweils eine individuelle Datenmenge dar. Erst jetzt ist es inhaltlich überhaupt möglich, einen Vergleich anzustellen und kritische Punkte über die gestellten Fragen zu identifizieren und kollegial zu besprechen.

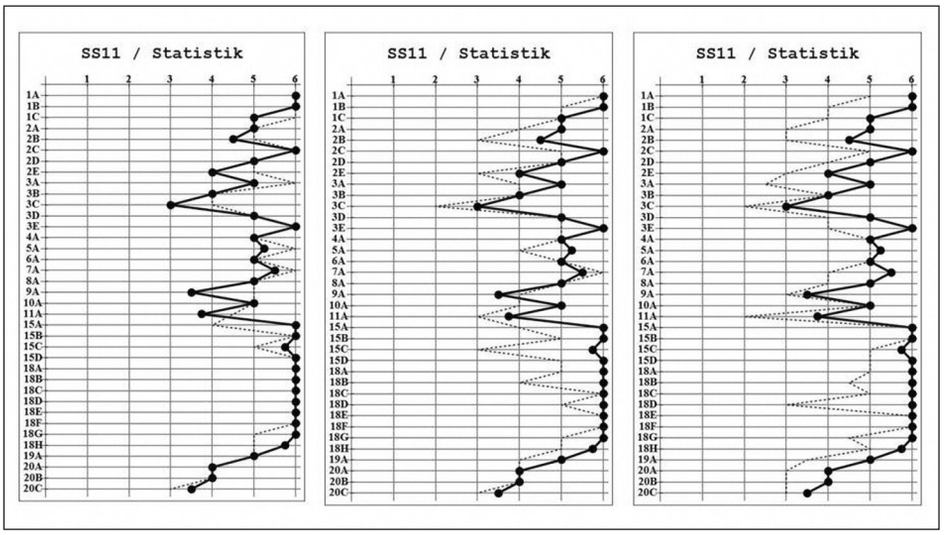

Abbildung 4: Vergleich der Profillinien für mehrere einzelne Lehrende (gestrichelter Linienzug) mit der Medianlinie der Mediane aller Statistik-Lehrenden

Die bisherige Vorgehensweise führt viel zu häufig einerseits zu Diskussionen ohne wirklichen Inhalt oder andererseits zu Analysen über Differenzen, die tatsächlich gar nicht existieren.

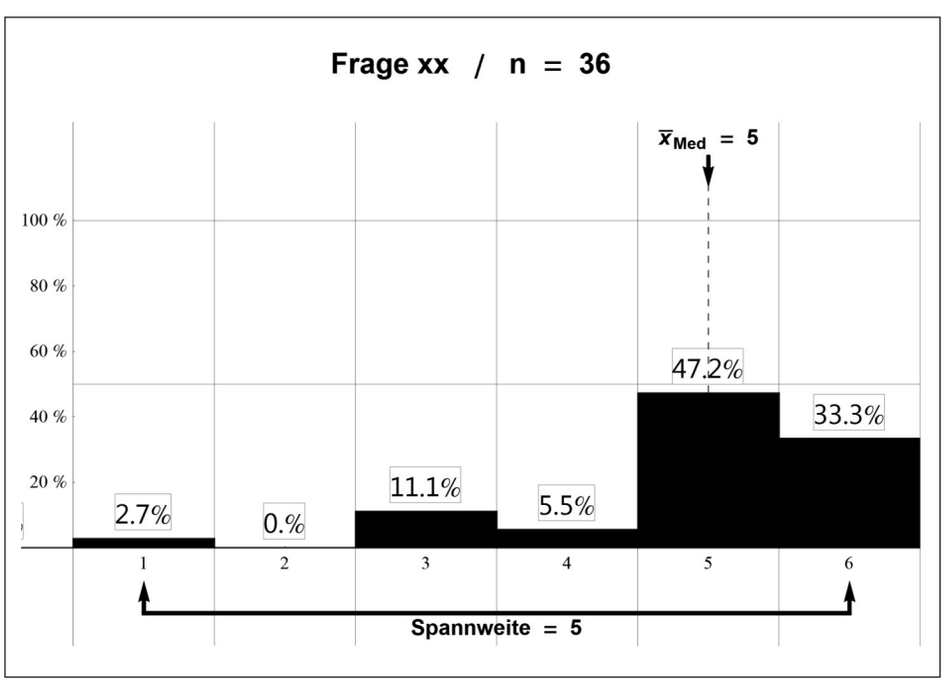

Abbildung 3: Statistisch korrekte Datenauswertung in Form eines Histogramms inklusive der numerischen Werte für Median und Spannweite (1 = schlechtes, 6 = gutes Ergebnis)

Verbesserungsvorschläge

Es ist erforderlich, auf die sinnlose Berechnung und Darstellung der Kenngrößen Mittelwert und Standardabweichung zu verzichten. Stattdessen sollten Median und Spannweite bestimmt und das Histogramm angesehen und interpretiert werden. Ein weiteres allgemeines Problem vieler verwendeter Fragebögen ist darüber hinaus, dass sie eher Themen wie Zufriedenheit oder Dienstleistungsqualität bei den Studierenden abfragen und nicht wirklich die Qualität der Lehre messen. Zum Beispiel fehlen auch Fragen wie „Haben Sie in dieser Veranstaltung etwas gelernt?“. Es findet normalerweise auch keine Verknüpfung der Auswertungen mit den Ergebnissen der Klausuren statt.

Weitere Gebiete mit falscher statistischer Vorgehensweise

Fraglich ist, ob die nicht zulässige Art und Weise der Anwendung statistischer Methoden beschränkt ist auf die Evaluation an Hochschulen. Dies ist mitnichten der Fall!

So ist die Angabe einer Durchschnittsnote neben der Notenübersicht in den Klausuren unserer Schulkinder ebenso unsinnig wie die Angabe einer durch-

schnittlichen Abiturnote. Dabei sei an dieser Stelle noch kein Wort verloren über die Noteninflation an Schulen. Eine besondere Brisanz erfährt diese Problematik bei relativen Noten und allgemein der Umrechnung von Noten zwischen verschiedenen Notensystemen.

Auch viele Marketingbücher enthalten schlicht falsche Informationen zu diesem Thema. Den Autoren ist gemeinsam, dass sie zu begründen versuchen, warum ein ordinalskaliertes Merkmal wie eine Schulnote doch als intervallskaliert aufzufassen sei. Erst mit dieser Eigenschaft lassen sich Methoden wie Faktoranalysen etc. durchführen. Zwei willkürlich gewählte Beispiele belegen dies eindrucksvoll (Berekoven et al. 2006, Schilke 2007). Es scheint ein in sich geschlossenes System von nicht zutreffenden Annahmen zu bestehen, da sich Schilke in einer Fußnote auf Seite 168 wiederum auf Berekoven et al. bezieht mit dem Hinweis: „Ein wesentlicher Vorteil dieser Art der Item- bzw. Antwortformulierung ist, dass die Daten als intervallskaliert angesehen werden können, was die Voraussetzung für die Anwendbarkeit verschiedener Analysemethoden (...) darstellt."

Ein weiteres Beispiel mag das jährlich durchgeführte DHV-Ranking sein. Dabei werden Rektoren und Wissenschaftsminister von den Mitgliedern des Deutschen Hochschulverbandes bewertet. Wieder werden Mittelwerte mit einer Genauigkeit von zwei Stellen nach dem Komma in einer Tabelle abgedruckt. Die beiden ersten Plätze (für die Erhebung im Jahr 2012) weisen aber den gleichen Medianwert auf. Die Plätze 3 bis 14 müssten bei korrekter statistischer Vorgehensweise exakt einem Platz zugeordnet werden!

Hochschulpolitische Bedeutung

Was sind die Konsequenzen dieser beschriebenen Vorgehensweise?

Es ist zu fragen, ob es bei der heutzutage durchgeführten Art der Evaluation noch um Qualität geht oder doch mehr um die Beruhigung der bildungspolitischen und Hochschul-Öffentlichkeit und darum, Akkreditierungskommissionen zufriedenzustellen. Tatsächlich wird einem geistlosen Berichtswesen genügt.

Erinnern wir uns kurz. Mit welchem Ziel sind Evaluationen denn so großflächig über den Hochschulen und Universitäten dieses Landes ausgeworfen worden? Die hehre Idee war, Studierenden die Aussicht auf eine bessere Lehre zu bieten. Das kann diese sinnlose Verwendung und Auswertung von Fragebögen nicht leisten. Werden sechsstellige Beträge pro Jahr und Hochschule in die Hand genommen, darf man wirklich etwas mehr erwarten. Die Beachtung der einfachsten Zusammenhänge in der beschreibenden Statistik könnte helfen und wäre schon mal ein Beginn, um die aktuelle Situation zu verbessern. ■

Danksagung: Ich möchte mich bei meinen Kollegen Karl Moog, Siegbert Preuß, Bruce Spear und Axel Zernack für sehr fruchtbare Diskussionen bedanken.

Literatur

Berekoven, Ludwig; Eckert, Werner; Ellenrieder, Peter: Marktforschung. Methodische Grundlagen und praktische Anwendungen. Betriebswirtschaftlicher Verlag Gabler, 11. Auflage, 2006.

Krüger, Thomas; Rudinger, Georg: Rektor und Wissenschaftsminister des Jahres 2012. Ergebnisse des DHV-Rankings. In: Forschung & Lehre, Nr. 5, Jg. 19 (2012), S. 292–295.

Schilke, Oliver: Allianzfähigkeit – Konzeption, Messung, Determinanten, Auswirkungen. Deutscher Universitäts-Verlag, Dissertation Universität Witten/Herdecke, 2007.

HAW Hamburg/VR China

China baut Fachhochschulsektor auf/Besuch in Hamburg

Die Volksrepublik China will das Hochschulwesen stärker auf die Bedürfnisse des Arbeitsmarkts ausrichten. Dabei sollen Universities of Applied Sciences als ein neuer Hochschultyp eingeführt werden. Etwa 600 Universitäten, die derzeit in Trägerschaft der chinesischen Provinzen geführt werden, sind für eine Umwandlung vorgesehen.*

In diesem Zusammenhang besuchte eine hochrangige Delegation aus der chinesischen Provinz Yunnan im Oktober 2014 die Hochschule für Angewandte Wissenschaften Hamburg. Unter Leitung des Vizegouverneurs Gao Feng und in Begleitung des chinesischen Generalkonsuls in Hamburg Yang Huiqun informierten sich Mitglieder von Hochschulleitungen und Mitarbeiter des Hochschulsekretariats der Provinzregierung über die Arbeitsweise und die charakteristischen Merkmalen einer deutschen Fachhochschule. Empfangen wurden sie von der Präsidentin der HAW Hamburg Prof. Dr. Jaqueline Otten und dem China-Beauftragten der Hochschule Prof. Dr. Reinhard Völler. Die HAW Hamburg bietet bereits seit über 20 Jahren Studiengänge in China an. So dürften die Gespräche nützliche Anregungen für den geplanten Aufbau der „West Yunnan University of Applied Sciences" geliefert haben.

* Zha, Qiang und Wang, Chuanyi: Systematic Changes in China, International Higher Education, 76 (2014), S. 14–16.