Grundlagen der Statistik nicht nur lernen sondern begreifen

Excel-Programme zur verständnisintensiven Nachbearbeitung
des Seminarstoffs zur deskriptiven und operativen Statistik

Autor:

Dr. Helmut Stauche

Institut für Erziehungswissenschaft

Universität Jena

November 2007

Unser Programm des Einführungsseminars

Im modularisierten Magisterstudium der Erziehungswissenschaft stehen am Institut für Erziehungswissenschaft der Friedrich-Schiller-Universität Jena zwei Semesterwochenstunden für ein einführendes Seminar zur allgemeinen Epistemologie, zur Praxis der Datenerhebung sowie zu grundlegenden quantitativen Auswerteverfahren zur Verfügung.

Im Bereich der deskriptiven Statistik wird u. a. auf zentrale Werte und Streuungswerte in monovariaten Verteilungen, auf die Darstellung bivariater Verteilungen im Diagramm und in der Kreuztabelle sowie auf verschiedene Assoziationsmaße zur mathematischen Beschreibung bivariater Verteilungen eingegangen. Für das Verständnis von Kovarianz und Korrelation wird die Punktewolke im Koordinatensystem (Menge der Messwertpaare) betrachtet und interpretiert. Im Anschluss daran wird auch auf die lineare Regression eingegangen und die Entstehung der Regressionsgeraden sowie deren Bedeutung erläutert.

In den Sitzungen zur operativen Statistik werden zunächst die Grundaxiome der Wahrscheinlichkeitstheorie, der Zusammenhang empirischer und theoretischer Verteilungen als Basis der Signifikanzprüfung, die Gauß'sche Normalverteilung $N(\mu,\sigma^2)$, die Interpretation der Flächen unter der Gauß-Funktion, die Standardisierung der Gauß'schen Normalverteilung zu N(0;1) sowie der Zusammenhang von Fehlerwahrscheinlichkeit und Ergebnisgenauigkeit behandelt, bevor die Schrittfolge zur Prüfung statistischer Hypothesen erläutert und auf den z-Test, den unabhängigen t-Test und den χ^2-Test angewendet wird. Anhand des unanhängigen t-Tests wird auch auf einseitiges und zweiseitiges Testen eingegangen.

In beide Seminarteile zur Statistik werden Aufgaben aus dem Einsatzfeld eines Erziehungswissenschaftlers einbezogen und diese sowohl im Seminar als auch zu Hause gelöst.

Das Testat am Ende des Semesters findet computergestützt statt. Zugelassene Hilfsmittel sind ein Taschenrechner, eine Formelsammlung, die Verteilungen von χ^2 und t sowie die Zufallshöchstwerte für die Signifikanz von r.

Eingabeformen sind Single und Multiple Choice, Eingabe von Zahlen, Ergebnissen von Aufgaben (ohne Rechenweg – die gestellten Aufgaben werden zuvor auf bereit gestelltem Papier gelöst), Begriffen und Wortgruppen.

Insgesamt gibt es zwei Parallelversionen des Tests, darüber hinaus werden die Aufgaben an benachbarten Rechnern in Zufallsreihenfolge gezeigt. Nach Ablauf der Zeit wird das Ergebnis als Punktestand und Zensur bekannt gegeben.

Verständnisintensives Erlernen des Stoffes

Angestrebt wird eine Beschäftigung mit dem Stoff, die zu anwendungsbereitem und zugleich problemoffenem Wissen führt. Von verständnisintensivem Lernen kann man dann sprechen, wenn die Erfahrungen und Vorstellungen, das Begreifen und die Metakognition zusammenwirken. Grundlegend für das verständnisintensive Lernen ist die Bedeutung des Verstehens. „Verstehen“ wird nach Fauser[1] als ein Lernen begriffen, bei dem Erfahrung und eigenes Tun, Vorstellungsbildung, Begreifen und kritische Reflexion zusammenwirken. Fauser entwickelte dieses Konzept in Zusammenarbeit mit dem Thüringer Kultusministerium[2].

Neuartig am Konzept ist „die Betonung der Erfahrung eigenen Tätigseins im Verhältnis zu Erfahrungen aus zweiter und dritter Hand, die Bedeutung des Denkens in und mit Vorstellungen im Verhältnis zu anderen Formen des Denkens, die Entwicklung und Begründung von Regeln und Gesetzen gegenüber deren bloßer Anwendung, die Aufmerksamkeit für Lernen als Prozess im Verhältnis zu dessen Zielen sowie die aktiv-konstruktive gegenüber der reproduktiven Qualität des Lernens.“ (Fauser 2003[3], S. 260)

[1] Prof. Peter Fauser ist Inhaber des Lehrstuhls für Schulpädagogik und Schulentwicklung am Institut für Erziehungswissenschaft der FSU Jena und Leiter des Imaginata e.V.

[2] vgl. Robert-Bosch-Stiftung: Verstehen lehren – Unterrichtsentwicklung durch verständnisintensives Lernen: http://www.bosch-stiftung.de/content/language1/html/997.asp (aufgesucht am 28.11. 2007)

[3] Fauser, P.: Lernen als innere Wirklichkeit. Über den Zusammenhang zwischen Imagination, Lernen und Verstehen. In: Rentschler, Madelung, Fauser (Hrsg.): Bilder im Kopf. Texte zum Imaginativen Lernen. Seelze-Velber 2003, S.242-286.

Wahrgenommene Erfahrungen und ein Lösungsansatz

Seit der Verbreitung der computergestützten Statistik ist in der statistischen Grundlagenausbildung im Ergebnis eines von mir angestellten Vergleichs des eigenen mit dem Vorgehen einiger anderer Einrichtungen gleicher oder ähnlicher Zielstellung zu konstatieren, dass das statistische Auswerteverfahren häufig auf

- dessen Problem angemessene Auswahl,
- die formale Durchführung der Prozedur (fast immer mit Hilfe des SPSS-Menüs) und
- die Interpretation des erhaltenen Ergebnisses

reduziert wird.

Ich halte dies für einen Mangel, weil damit das oben geforderte Begreifen der Prozedur nicht hinreichend unterstützt wird. Deshalb wird in meiner Lehre bewusst nicht auf das Berechnen statistischer Maßzahlen mit konventionellen Mitteln (Taschenrechner, Tabellenmaterial etc.) verzichtet, wenngleich die Beispiele zu einem Aufgabentyp bezüglich des Umfangs und der Komplexität der Daten einfach gehalten werden. Zielführend kann ein derartiges Vorgehen jedoch nur dann sein, wenn eine Formel nicht nur algorithmisch abgearbeitet wird sondern wenn während des Lösungsprozesses die Sinnerfülltheit ihres Zustandekommens und ihrer Struktur erfasst bzw. vertieft wird. Günstige Ansatzpunkte hierfür bieten zum Beispiel die Varianz, die Kovarianz, der Pearson'sche Maßkorrelationskoeffizient sowie die χ^2-basierten Assoziationsmaße Φ, C und Cramer's V.

Allerdings muss sich dieses Verständnis unterstützende Arbeiten auf das jeweils ausgewählte Beispiel beschränken, Modifikationen durch den Austausch von Werten wären nur mit verhältnismäßig großem Aufwand möglich und aus Zeitgründen nicht realisierbar.

Einen Ausweg aus dem Dilemma und damit eine Hinwendung zum verständnisintensiveren Erfassen der Auswerteprozedur sehe ich in einem zwischen den beiden erwähnten Ansätzen

- konventionelle Lösung einer Aufgabe laut Formel und
- Input -Output-Betrieb eines PC-Statistikprogrammes

liegenden Vorgehen.

Zur Vertiefung der Seminararbeit und zur individuellen Vorbereitung auf das Testat stelle ich den TeilnehmerInnen am Seminar sechs Excel-Files zur Verfügung, die teils mit Tabellenblattoperationen, teils mit im Hintergrund laufenden VBA-Programmen[4] bedient werden.

Im Gegensatz zur Arbeit mit SPSS wird MS Excel veranlasst, neben dem Endergebnis auch sämtliche Zwischenergebnisse, die durch die relevanten Terme der Formel abgebildet werden, zu berechnen und anzuzeigen.

Eine Variation der eingegebenen Daten führt zur simultanen Veränderung von Zwischen- und Endergebnissen.

Beide Features zusammen bieten die Möglichkeit, Reserven für das Begreifen der Prozeduren aufzuschließen und in fast spielerischem interaktivem Umgang mit den In- und Outputs konstruktiv Wissen aufzubauen.

Die sechs Excel-Dateien und ihre Potenzen

Erste Datei *rechenhilfen1.xls* zur deskriptiven Statistik

Für die bereits erwähnten Beispiele aus der deskriptiven Statistik heißt das:

1. Nach der Eingabe der (bis zu 25) Werte einer monovariaten Verteilung zeigt das Tabellenblatt die Spannweite der Verteilung, den Wert des arithmetischen Mittels, die Abweichungen aller Messwerte vom Mittel, deren Quadrate und die daraus gebildete Summe der Abweichungsquadrate sowie deren Absolutbeträge und die Summer derselben.

[4] Zum Aktivieren der VBA-Programme ist vor dem Öffnen der Excel-Dateien Folgendes zu tun:
– Bei Office 2003: *Extras => Makro => Sicherheit* bedienen und dort Punkt bei Stufe *Mittel* setzen. Nach erneutem Öffnen der Datei *Makros aktivieren* wählen.
– Bei Office 2007: *Entwicklertools => Makrosicherheit* und dort Punkt bei *Alle Makros mit Benachrichtigung deaktivieren* setzen. Nach erneutem Öffnen der Datei auf *Optionen* neben der Sicherheitswarnung klicken und *Diesen Inhalt aktivieren* wählen. Sollte in der Funktionsleiste *Entwicklertools* nicht angezeigt werden, klicken Sie auf das Symbol links oben, danach unten rechts auf *Excel-Optionen* und setzen den Haken bei *Entwicklertools in der Multifunktionsleiste anzeigen*.

Schließlich werden unter Bezugnahme auf den Freiheitsgrad bzw. auf den Stichprobenumfang die Standardabweichung sowie die Mittlere Abweichung ausgegeben.

2. Der Programmteil Kovarianz und Pearson'scher Korrelationskoeffizient r) wurden bewusst in einem Teilprojekt verarbeitet, denn schließlich ist Pearson's r nur eine die Kovarianz auf den Wertebereich $-1 \leq r \leq 1$ normierende Erweiterung. Alle zur weiteren Berechnung nötigen Zwischenergebnisse, d.h. beide arithmetische Mittel und beide Standardabweichungen werden angezeigt, weiterhin die Differenzen der einzelnen Werte zu ihrem Mittel und das Produkt der Differenzen für jedes einzelne Wertepaar. Aus dem Vorzeichen und dem Betrag der Differenzen kann nun augenfällig abgelesen werden, ob der Zähler der Formel positiv oder negativ ausfallen wird und damit ein positiver oder ein negativer Zusammenhang bestimmter Stärke zustande kommen wird.

3. Zur Berechnung der χ^2-basierten Assoziationsmaße Φ , C bzw. Cramer's V einer Kreuztabelle werden – insbesondere zum Verständnis der Berechnung von χ^2 – nach Eingabe der beobachteten Werte f_b die Erwartungswerte jeder Zelle f_e sowie die aus f_b und f_e berechneten Zellensummanden $(f_b - f_e)^2 / f_e$ gezeigt.

4. Ein weiterer Programmteil zeichnet nach Vorgabe der jeweiligen Minima und Maxima zweier Likert-Skalen diese gleich breit untereinander. Nach Angabe eines arithmetischen Mittels auf der Skala A wird das dazu analoge Mittel auf der Skala B berechnet und eine senkrechte rote Linie vom Programm gezeichnet, die die Lage beider Mittelwerte veranschaulicht.
   Des weiteren rechnet das Programm die Skalenwerte von A nach B um und macht auf dabei entstehende Rundungsfehler aufmerksam.

Über das Beschriebene hinaus existieren Programmteile für den Median und den Modalwert, für die Rangkorrelation, die punktbiseriale Korrelation, für den tetrachorischen Koeffizienten sowie für die Flächen unter der standarisierten Normalverteilung und die Umrechung von z-Werten in andere Normwertskalen.

In jedem der zehn Teile des Excelblattes – nicht allein bei den vier Beispielen – ist ein anschauliches Variieren der Rohwerte möglich, alle Zwischenwerte und das Endergebnis reagieren darauf simultan.

Zweite Datei *rechenhilfen2.xls* zur operativen Statistik

Im Bereich der operativen Statistik wird für jede der oben erwähnten Prüfungen gegen die Nullhypothese zunächst die Prüfgröße nachvollziehbar berechnet und der Freiheitsgrad des Systems ausgewiesen. Zeitgleich greift das Programm auf eine in einem ausgeblendeten Tabellenblatt befindliche Datenbank zurück, deren Inhalt aus Clauß/Ebner 1978[5] gescannt wurde. Aus dem Vergleich des berechneten Wertes der Prüfgröße und des von Fehlerwahrscheinlichkeit und Freiheitsgrad abhängigen kritischen Wertes aus der Datenbank wird auf die Verifikation der Nullhypothese geschlossen.

Beim unabhängigen t-Test zum Prüfen von Mittelwertunterschieden werden zwei Berechnungen angeboten. Für den Fall, …

1. dass die gemeinsame Standardabweichung aller zu beiden Gruppen gehörigen Fälle bekannt ist und
2. dass die Standardabweichungen in beiden unabhängigen Gruppen einzeln bekannt sind. Im letzteren Fall wird die gemeinschaftliche Standardabweichung aus der gewogenen Mittelung der beiden Einzelwerte bezogen.

   Auch hier bietet sich augenfällig die Möglichkeit, den Einfluss der Differenz zwischen den beiden Gruppengrößen auf das gewogene Mittel auszuprobieren und damit den Unterschied des gewogenen Mittels zum arithmetischen Mittel zu erfassen.

Schließlich wird in einem weiteren Teilprojekt der Zusammenhang zwischen Fehlerwahrscheinlichkeit und Genauigkeit des Ergebnisses verdeutlicht. Ausgehend von Elementen der Stichprobentheorie wird sowohl für einen Schluss von der Grundgesamtheit auf eine Stichprobe als auch umgekehrt anhand selbst gewählter Daten aufgezeigt, dass

1. das Ergebnis nie ein genauer Wert ist sondern immer ein Ergebnisintervall (Vertrauens- oder Konfidenzintervall genannt) sein muss und dass
2. dieses Intervalls bei einer vorgegebenen Fehlerwahrscheinlichkeit von 1% breiter ausfällt als bei 5%.

---

[5] Clauß, G & Ebner, H.: Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen, Volk und Wissen Volkseigener Verlag Berlin 1978.

Alles in allem sollten diese beiden Excel-Files selbsterklärend sein, so dass an dieser Stelle keine weitere Anleitung gegeben werden muss.

Dritte Datei *korrelationsdiagramm.xls*

Die Excel-Datei zeigt den ersten Quadranten eines Y-X-Koordinatensystems. Das Doppelklicken auf beliebige Zellen bewirkt das Entstehen eines Punktes und damit eines Wertepaares.[6]

Nach dem Verlassen der jeweils letzten doppelgeklickten Zelle wird vom Programm simultan der Pearson'sche Maßkorrelationskoeffizient berechnet und angezeigt.

Wiederholtes Doppelklicken auf die Zelle entfernt den Punkt und verändert damit simultan den Wert von r.

Auch diese Programmierung kann – wenngleich in symbolisierender Weise nur Werte für Y und X aus dem Bereich der natürlichen Zahlen bis 20 möglich sind – zum tieferen Verständnis des Zusammenhangs zwischen dem Aussehen der die Messwerte verkörpernden Punktewolke und dem dazugehörigen Korrelationskoeffizienten beitragen.

Zuzüglich zur r-Berechnung aus Messwertpaaren gibt es zwei Aufgaben, deren Lösungen eingesehen werden können:

1. Zeichnen Sie alle Wertepaare in eine Zeile bzw. in eine Spalte. Warum entsteht die Fehlermeldung?
2. Stellen Sie mehrere Möglichkeiten für Null-Korrelationen dar!

Vierte Datei *regressionsgeraden.xls*

Ähnlich aufgebaut wurde dieses Projekt. Eine X-Y-Wertetabelle[7] wird mit bis zu 21 Wertepaaren gefüllt. Verarbeitet werden der Einfachheit und Übersichtlich-

[6] Um das Einklicken der Punkte zu ermöglichen, erfolgt die Einteilung des Quadranten in der Art der „Rechenkästchen". D.h. der das Messwertpaar abbildende Punkt liegt jeweils in der Mitte eines solchen Kästchens. Von daher gibt es die Werte X=0 und Y=0 nicht, die Werte beginnen bei 1. Für die richtige Berechnung von r ist dies ohne Belang.

[7] In diesem Projekt wird ein korrektes Koordinatensystem verwendet (vgl. Fußnote 6), da andernfalls das Zeichnen der Geraden nicht möglich wäre.

keit zuliebe nur natürliche Zahlen. Mit Mausklicks werden die Wertepaare als übergroße Punkte im ersten Quadranten des X-Y-Koordinatensystems angezeigt, die Parameter für die Geradengleichungen $y=a_1+b_1x$ sowie $x=a_2+b_2y$ werden berechnet, die Geradengleichungen aufgeschrieben und beide Regressionsgeraden (mit Abstandsmessung parallel zur Y-Achse bzw. zur X-Achse) in das Koordinatensystem eingezeichnet.

Mit weiteren Klicks können die Messwerte, die Geraden und der Tabelleninhalt einzeln gelöscht werden, so dass auch hier für das „Experimentieren“ Gelegenheit gegeben ist.

Anregungen für Letzteres gibt das Projekt vor:

1. Füllen Sie die Tabelle mit Werten, wobei Sie entweder Y oder X konstant halten.
   Warum entsteht nur eine Geradengleichung, während die andere nicht erzeugt werden kann?

2. Füllen Sie die Tabelle mit den folgenden Werten: **X** 3 6 3 6
   Y 3 6 6 3
   und sagen die die Lage der beiden Geraden voraus.

3. Denken Sie sich Möglichkeiten dafür aus, dass beide Geraden zusammen fallen und überprüfen Sie Ihre Überlegungen.

Fünfte Datei *flaechen_unter_normalkurve.xls*

Das Verständnis der operativen Statistik wird in meinem Seminar auf den Vergleich der empirisch erhobenen und der zugehörigen theoretischen Verteilung gegründet. Im Vorfeld dieses Verständnisses spielt die Interpretation der Flächen unter der aus der Gauß'schen Normalverteilung $N(\mu,\sigma^2)$ durch z-Standardisierung gewonnenen Funktion eine entscheidende Rolle, indem diese Überlegungen bis hin zur Herleitung des Kriteriums für den z-Test ($DF = \infty$) und den unabhängigen t-Test ($DF < \infty$) führen. Die für die Studierenden m.E. ungewöhnliche Interpretation der Fläche eines bestimmten Integrals als Wahrscheinlichkeit soll mit dieser vierten Arbeit unterstützt werden.

Auch in diesem Excel-Tabellenblatt können durch Doppelklicken Werte gewählt werden: Rationale Zahlen mit Zehntelgenauigkeit zwischen –3 und +3.

Wird nur eine Zahl z_1 markiert, berechnet das Programm die Wahrscheinlichkeit dafür, dass die Variable z im Intervall $-\infty < z < z_1$ liegt. Im Falle, dass zwei Zahlen z_1 und z_2 markiert wurden, wird die Wahrscheinlichkeit für das Intervall $z_1 < z < z_2$ berechnet. Sehr anschaulich lassen sich mit diesem Programm die Wahrscheinlichkeiten dafür angeben, dass ein Wert in der einfachen, doppelten oder dreifachen Streuungsbreite liegt – Grundlage für das Erfassen von z-basierten Normwertskalen der Testdiagnostik.[8]

Sechste Datei *histogramm_interaktiv.xls*

Mit dieser VBA-Arbeit soll das Verständnis für ein Histogramm vertieft werden. Ausgehend von einem Beispiel, das das in Jahren angegebene Lebensalter in beliebig[9] breiten Intervallen samt beliebiger[9] dazu gehöriger Fallanzahlen eingeben lässt, werden nach Betätigung einer Schaltfläche die Intervallbreiten sowie die Häufigkeitsdichten berechnet. Mit diesen Größen zeichnet sich das Histogramm.

Ersichtlich werden soll,, dass ein Histogramm die Verteilung einer intervallskalierten Variablen anders zeigt als ein Balkendiagramm. Mit ihm wird der visuellen Wahrnehmung des Menschen eher Rechnung getragen als mit dem Balkendiagramm, denn ebenso wie beim Kreisdiagramm ist wegen der Multiplikation von Intervallbreite und Häufigkeitsdichte die Rechteckfläche der Häufigkeit direkt proportional, beim Balkendiagramm ist es die Höhe des Balkens. Deutlich wird auch, dass die Balken eines Histogramms ohne Zwischenraum nebeneinander liegen. Im Sonderfall lässt sich zeigen, dass bei gleichen Intervallbreiten auch gleich breite Balken entstehen.

[8] vgl. dazu auch die Publikation „Normwerte der Testdiagnostik“ unter http://www.db-thueringen.de/servlets/DocumentServlet?id=10051

[9] Die Beliebigkeit musste insofern limitiert werden, als dass das vorgezeichnete Koordinatensystem nicht gesprengt werden darf: Auf der Ordinate darf die Häufigkeitsdichte nicht größere Werte annehmen als 20 und auf der Abszisse die Intervallbreite nicht größere Werte als 100.

Alle sechs Excel-Dateien werden als weitere Derivate diesem Aufsatz in der Digitalen Bibliothek Thüringen angehängt und sollen damit der interessierten Öffentlichkeit zur Verfügung stehen. Der Autor verbindet mit dieser Publikation auch die Hoffnung, dass die Arbeit von anderen Lehrenden der Disziplin als nützliche Hilfe angenommen wird.

Sollte der Nutzer Fehler entdecken oder sollte sich Inkompatibilität mit einer jüngeren Excel-Version als 2003 zeigen, bitte ich um Mitteilung unter shs@uni-jena.de.