Grundbegriffe der Beschreibenden Statistik

Dokument erstellt am: 03. Mai 2005
Letztmalig geändert am: 25. Mai 2005

1. Datenmatrix und Messniveaus

1.1. Merkmale, Datenmatrix, uni- und multivariate Analysen

Ziel empirischer Sozialforschung ist es, den Zusammenhang zwischen Untersuchungseinheiten (oder *Merkmalsträger*) und seinen Eigenschaften (oder *Merkmale*) zu ergründen. Sowohl Merkmalsträger und Merkmale können verschiedenartig sein. Merkmalsträger können z.B. Individuen, Gruppen, Regionen und Staaten sein. Jeder Merkmalsträger kann über ein oder mehrere Merkmale wie z.B. Alter, Geschlecht, Einkommen, Meinungen, Wahlentscheidungen, Kriminalitätsraten usw. verfügen. Natürlich variieren die individuellen Entscheidungen, d.h. ihre *Merkmalsausprägungen* oder *Werte*.

Die in einer Umfrage erhobenen Daten lassen sich tabellarisch organisieren. In der empirischen Sozialforschung repräsentieren in der so genannten *Datenmatrix* die Reihen die einzelnen *Fälle* (engl. *case*), das sind die Merkmalsträger, Untersuchungseinheiten, Objekte usw. Die Spalten enthalten die Merkmale oder *Variablen* (engl. *variable*), und zwar in jeder Spalte ein- und dasselbe Merkmal für jeden Fall. In das Tabellenfeld am Kreuzungspunkt von Fall (Merkmalsträger) und Variable (Merkmal) wird die Merkmalsausprägung, d.h. der *Wert* (engl. *value*) der Variable, eingeschrieben. Für die statistische Analyse ist es zwar wünschenswert, aber nicht notwendig, dass es zu jeder Variablen eines jeden Falls einen Wert gibt (die Datenmatrix wäre dann *vollständig*); ein Wert (eine Merkmalsausprägung) darf auch fehlen (sie darf dann aber nicht in die rechnerische Analyse mit einbezogen werden).

Werte können verschieden sein: Zahlen wie im Falle des Alters, oder Kategorien wie. z.B. männlich, weiblich oder Transvestit bzgl. des Geschlechts. Für Computer ist es in der Regel günstiger, auch letztgenannte Kategorien durch Zahlen auszudrücken, sie müssen dann mit einem Etikett (engl. *label*) versehen werden, um auch später die Merkmalsausprägung nachvollziehen zu können.

Die Auswertung erfolgt in aller Regel spaltenweise. Analysiert man nur eine Variable, d.h. eine Spalte, so spricht man von univariater Analyse. Wertet man hingegen die Abhängigkeit einer (abhängigen) Variablen von *einer* (unabhängigen) Variablen aus, so spricht man von *bi*variater Analyse. Sinngemäß zeigt eine *multi*variate Analyse den Zusammenhang einer (abhängigen) von mehr als einer unabhängigen Variable auf.

Eine reihenweise (fallweise) Analyse ist selten, weil sie dem Ziel entgegensteht, eine allgemeingültige Aussage für eine Gruppe von Merkmalsträgern zu erhalten.

1.2. Messniveaus (Skalentypen)

Werte (oder Merkmalsausprägungen) können sowohl qualitativer Natur (z.B. Geschlecht, Nationalität, Konfessionszugehörigkeit) als auch quantitativer Natur (z.B. Alter, Körpergröße, Anzahl Familienmitglieder, Einkommen etc.) sein. Dies spiegelt sich in der Charakterisierung der Variablen, seinem *Messniveau* oder *Skalentyp* wider. Insbesondere steht die Frage, ob Variablenwerte diskret oder kontinuierlich verteilt sein können, bzw. ob es im Falle von diskreten Variablen *Ordnungsregeln* gibt.

Folgende vier Messniveaus (Skalentypen) werden unterschieden:

Nominalskalen. Die einfachste Form besteht darin, die Ergebnisse eines Merkmals zu kategorisieren (oder zu klassifizieren). Jeder Wert, jeder Merkmalsausprägung, wird ein Wort oder ein Zahlenwert zugeordnet. Diese Kategorien können aber *nicht* rangmäßig geordnet werden. Hierunter fallen alle Alternativklassifikationen wie z.B. männlich/weiblich, berufstätig/nicht berufstätig, Jugendlicher/Erwachsener. Aber es können durchaus auch mehrere Kategorien nötig sein, wie z.B. im Falle des Familienstandes (ledig, verheiratet, getrennt lebend, geschieden und verwitwet) oder Rückennummern der Spieler einer Fußballmannschaft. Charakteristisch für Ausprägungen einer Nominalskale ist, dass sie sich gegenseitig ausschließen.[1]

Ordinalskalen. Ordinalskalen werden verwendet, wenn sie die Ausprägungen nicht nur klassifizieren, sondern auch rangmäßig ordnen lassen; die Größe der (numerischen) Differenz der einzelnen Kategorien spielt hierbei keine Rolle. Ein Beispiel hierfür sind Bewertungskategorien wie z.B. sehr gut, gut, genügend, ausreichend und ungenügend oder sehr oft, oft, selten und nie.

Nominal- und Ordinalskalen charakterisieren *qualitative* Merkmalsausprägungen.

Quantitative Merkmalsausprägungen werden mit folgenden **metrischen** Messniveaus charakterisiert: den Intervall- und Ratioskalen.

Intervall- und Ratioskalen. Die Ausprägungen eines Merkmals lassen sich nicht nur rangmäßig ordnen, sondern Differenzen lassen sich *exakt* (also nicht *willkürlich*) angeben und haben eine Bedeutung. Wenn man beim Metzger 100 g Fleisch mehr verlangt, so bleibt die Menge gleich, egal, ob auch der Waage bereits 100 g, 200 g Fleisch oder gar nichts liegt. Ebenso lässt sich z.B.

[1] Eine Variable mit zwei Ausprägungen heiß dichotom, sinngemäß eine Variable mit drei oder mehreren Ausprägungen trichotom bzw. polytom.

die Abweichung der Studiendauer eines Studenten von der Regelstudienzeit angeben. Je größer die Abweichung von der Regelstudienzeit, um so länger dauerte das Studium. Dies bedeutet aber nicht, dass ein Student, dessen Studium zwei Semester über der Regelstudienzeit dauert, doppelt solange studierte wie ein Student mit nur einem Semenster über der Regelstudienzeit.[2]

Eine wichtige Rolle spielt hier die Existenz eines *absoluten* (*invarianten*) Nullpunktes. Eine numerische Skale ist automatisch auch eine Intervallskale. Eine Ratioskale (oder Verhältnisskale) ist eine Intervallskale, bei der die verhältnismäßige Graduierung gewahrt ist, d.h., sie besitzt einen absoluten Nullpunkt. Eine Intervallskale muss keinen absoluten Nullpunkt besitzen.

In der Praxis wird bei der Variablencharakterisierung zumeist nit zwischen Ratio- oder Intervallskale differenziert.

[2] Das bedeutet natürlich nicht, dass der Autor eine Studienzeit jenseits der Regelstudienzeit gut heißt oder gar fördern würde.

2. Ausgewählte Verfahren der Beschreibenden Statistik

2.1. Vorbemerkungen

Dieses Skript soll Begleitmaterial für das Propädeutikum „Einführung in SPSS“ sein, dessen Ziel es ist, den Umgang mit dem Programm selbst zu erlernen. Aus diesem Grund erfolgt hier keine vollständige Beschreibung der Methoden der Beschreibenden Statistik. Hier möge der Leser auf im Kapitel 4 genannte weiterführende Literatur zurückgreifen.

Da man ohne die Durchführung statistischer Berechnungen kaum die Bedienung des Programms erlernen kann, folgt hiernach eine kurze Beschreibung einfacher statistischer Verfahren.

Es ist wenig sinnvoll, einem Interessenten alle einzelnen Umfrageergebnisse vorzulegen, aus der Vielzahl der Daten ist es kaum möglich, das Wesentliche zu erfassen. Hierfür werden statistische Verfahren herangezogen, die eine Entscheidungsfindung anhand weniger Maßzahlen ermöglichen. Es sei hier bereits darauf hingewiesen, dass die Auswahl geeigneter statistischer Verfahren in der Hand des Auswerters liegt, somit seine Kenntnisse und Erfahrungen voraussetz, ein immer geeignetes Verfahren gibt es nicht.

In den nachfolgenden Abschnitten werden Häufigkeitsverteilungen, Mittelwerte und deren Streuungswerte als auch Kreuztabellen beschrieben.

2.2. Häufigkeiten und Häufigkeitsverteilungen

Eine deutliche Datenreduktion erreicht man bereits, wenn man bestimmt, wie viele Falle (Merkmalsträger) auf jeden möglichen Wert einer Variablen (Merkmalsausprägung) fallen. Diese Anzahl für jeden einzelnen Variablenwert nennt man **Häufigkeit *H*** (engl. *frequency*). Die Häufigkeit lässt sich sowohl absolut als auch relativ bzw. prozentual bezogen auf die Summe alle möglichen Häufigkeiten angeben:

$$H_{j,\,rel} = \frac{H_{j,\,abs}}{\sum_{i=1}^{n} H_{i,abs}} \quad \text{mit} \quad j = [1,\, n]$$

Prozentuale Häufigkeiten erhält man, indem man relative Häufigkeiten mit 100 % multipliziert.

Ein bekanntes Anwendungsbeispiel für Häufigkeiten sind Wahlergebnisse bzw. deren Hochrechnungen.

Die Berechnung von Häufigkeiten lässt sich auch nominale, ordinale und metrische variablen anwenden.

Die Funktion, die den Zusammenhang zwischen Häufigkeiten (y-Werte) und ihren zugehörigen Variablenwerten (x-Werten) beschreibt, nennt man **Häufigkeitsverteilung** (engl. *frequency distribution*). Zur besseren Veranschaulichung kann diese Werte in einem Balkendiagramm, man nennt dieses Diagramm *Histogramm*, in ein Streifendiagramm oder in ein Tortendiagramm – letztes insbesondere im Falle von relativen oder prozentualen Häufigkeiten anwenden – einzeichnen.

Im Falle *metrischer* Daten kann man die Häufigkeitsverteilung als Funktion auffassen, es lassen sich die einzelnen Messwerte mit einem Kurvenzug verbinden. Aus dieser Darstellung lassen sich für die nachfolgende Auswertung wichtige Schlussfolgerungen ziehen:

1. Die Häufigkeitsverteilung ist symmetrisch oder asymmetrisch. Viele Verfahren setzen eigentlich eine symmetrische Verteilung voraus. Bei asymmetrischen Kurvenverläufen hat man es mit rechtsschiefen (*oder* linkssteilen) bzw. linksschiefen (*oder* rechtssteilen) Kurven zu tun.

2. Die Häufigkeitsverteilung besitzt ein, zwei oder mehrere Maxima. Man nennt sie dann unimodal, bimodal bzw. multimodal. Auch hier gilt, dass viele Analyseverfahren unimodale Kurvenverläufe voraussetzen.

Hier gilt: Erfüllen die Messdaten nicht die Anforderungen für das gewählte Verfahren, so werden die Ergebnisse unsinnig und verschleiern wichtige Aussagen.

2.3. Ausgewählte Mittelwerte und Streuungswerte

Die Messergebnisse lassen sich häufig noch weiter verdichten, wenn man geeignete Maßzahlen berechnet.

Die bekanntesten und meist eingesetzten Maßzahlen sind *Mittelwerte* und *Streuungswerte*. Sie lassen sich in der Regel nur auf metrische Variablen anwenden. Wenn sinnvoll, sollten zu Mittelwerten immer geeignete Streuungswerte angegeben werden, um die Häufigkeitsverteilung zu charakterisieren.

2.3.1. Mittelwerte

Ein sehr anschaulicher Mittelwert ist der **Modus**. Er benennt den am häufigsten vorkommenden Wert einer Häufigkeitsverteilung. Der Modus kann auf nominale, ordinale und metrische Variablen angewandt werden. Nehmen wir das Beispiel: 1, 1, 1, 2, 2, 3, 4, 4, 4, 4, 5. Der Wert 4 kommt am häufigsten vor, nämlich 4-mal; der Modus ist somit 4.

Der bekannteste Mittelwert ist das **arithmetische Mittel** $\bar{x}$ (sprich „x quer“, engl. *mean value*, umgangssprachlich einfach Mittelwert). Das arithmetische Mittel lässt sich nur für metrische Variablen einsetzen. Man erhält das arithmetische Mittel, indem man alle Messwerte aufsummiert und die Summe durch die Anzahl der Messwerte teilt, als gilt

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Die Berechnung des Mittelwertes setzt streng genommen monomodale und symmetrische Verteilungen voraus. Wenn ich beispielsweise das Einkommen von zehn Hartz-IV-Empfängen (Arbeitslosengeld-2-Empfängern) und einem Millionär bilde, so erhalte ich ein Durchschnittseinkommen, das überhaupt nicht realisiert wird, d.h., es ist ein Artefakt. (Bei einer schiefen Kurve sollte man wenigsten die durchschnittliche Abweichung angeben.)

2.3.2. Streuungswerte

Streuungswerte geben an, wie weit vom Mittelwert aus gesehen, die Häufigkeitsverteilung verteilt ist.

Einfache Streuungswerte sind das **Minimum** $x_{\min}$ (die kleinste vorkommende Merkmalsausprägung), das **Maximum** $x_{\max}$ (die größte vorkommende Merkmalsausprägung) und der **Range**

$$R = x_{\max} - x_{\min}.$$

Diese Streuungswerte lassen sich nur auf metrische Variablen anwenden (das Minimum bzw. Maximum ließen sich auch noch auf ordinale Variablen anwenden, wenn sie numerisch verkodet sind). Je weiter Minimum und Maximum vom Mittelwert abweichen bzw. je größer der Range ist, um so weiter sind die Häufigkeiten verteilt.

Das Streuungsmaß Range versagt aber, wenn die Verteilung wenige, aber stark vom häufigsten Wert oder dem Mittelwert abweichende Extrema besitzt. Hier hilft das Streuungsmaß Quartilabstand weiter. Zunächst teilen wir die gesamte Verteilung in vier Abschnitte – in die so genannten **Quartile** –, die jeweils 25 % aller Fälle beinhalten. Bezeichnet man mit Q_1 den Schnittpunkt zwischen erstem end zweitem Quartil, und sinngemäß Q_3 den Schnittpunkt zwischen drittem und viertem Quartil, so erhält man den **Quartilabstand** (engl. *interquartile range*) zu

$$\text{Quartilabstand} = Q_3 - Q_1 \qquad \text{bzw.}$$

$$\text{mittlerer Quartilabstand} = \frac{Q_3 - Q_1}{2}$$

Ein weiteres Streuungsmaß ist das **arithmetische Mittel der Abweichungen vom Mittelwert** gemäß

$$\overline{\Delta x} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n}$$

Dieses Streuungsmaß ist nur für metrische Variablen bestimmbar. Es beschreibt die Kurvensymmetrie: ist $\overline{\Delta x} = 0$, so ist die Häufigkeitsverteilung symmetrisch, ist dagegen $\overline{\Delta x} > 0$ bzw. $\overline{\Delta x} < 0$, so ist die Häufigkeitsverteilung linksschief bzw. rechtsschief.

Man muss letzteres Streuungsmaß von der **durchschnittlichen Abweichung AD** (engl. *average deviation* oder *mean deviation*) unterscheiden, die anders definiert ist, nämlich:

$$AD = \frac{\sum_{i=1}^{n} | x_i - \bar{x} |}{n}$$

Dahinter verbirgt sich das arithmetische Mittel der Abweichungs*beträge*, d.h., das Vorzeichen der Abweichung wird nicht berücksichtigt. Auch dieses Streuungsmaß ist nur für metrische Variablen bestimmbar. Dieses Maß gibt die Breite einer Verteilungsfunktion an. Es ist aber wenig gebräuchlich, statt seiner wird eher die **Standardabweichung** *s* (engl. *standard deviati-*

on) bzw. die Varianz s^2 (engl. *variance*), das ist das Quadrat der Standardabweichung, verwendet. Die Standardabweichung *s* ist wie folgt definiert:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$$

Die Standardabweichung[3] ist ebenfalls nur für metrische Variablen bestimmbar. Sie beschreibt die Breite der Häufigkeitsverteilung. Eine große (kleine) Standardabweichung bedeutet eine große (kleine) Häufigkeitsverteilung. Das Intervall $\bar{x} \pm s$ umfasst ca. 67 % alle Fälle.

2.3.3. Angabe von Mittel- und Streuungswerten

Ein Mittelwert hat allein zwar bereits eine Bedeutung, bei Streuungswerten ist dies nicht der Fall. Ein Mittelwert sollte aber niemals alleinig angegeben werden, sondern immer mit einem passenden Streuungswert versehen werden, um eine Vorstellung von der Breite der Häufigkeitsverteilung geben zu können.

2.4. Die Kreuztabelle

Um den Zusammenhang zwischen zwei und mehreren Variablen ergründen zu können, benötigt man in der Regel kompliziertere Verfahren (z.B. Regressionsanalyse), für deren Erläuterung der zeitliche Rahmen dieses Propädeutikums nicht ausreichen würde.

Eine einfache bivariate Analyse lässt sich mittels einer bivariate Tabelle bzw. bivariaten Kreuztabelle (engl. *contingency table*, *two-way frequency table*, *cross-table*). Dies ist eine Tabelle, die die Häufigkeiten zweier Variablen, eine spaltenmäßig, die andere reihenmäßig notiert, vergleicht. Die erste Reihe bzw. Spalte enthalten alle möglichen Werte (Merkmalsausprägungen). Am Kreuzungspunkt einer bestimmten Spalte mit einer bestimmten Spalte wird die Häufigkeit notiert, die beide Merkmalsausprägungen erfüllen.

[3] Die Berechnung der Standardabweichung setzt streng genommen eine (symmetrische) Gauß-verteile Häufigkeitsverteilung voraus.

3. Anhang: Das Summensymbol

Wenn man die Summation einer sehr großen oder (vorher) nicht genau bekannten Anzahl von Summanden mathematisch ökonomisch schreiben will, so benutzt man die Summensymbolik, erkennbar am griechischen Großbuchstaben Σ (Sigma). Nehmen wir an, wie wollten *n* Werte der Variablen *x* summieren, deren Werte wir noch nicht kennen, so müssen wir als erstes „Stellvertreter" einführen: x_1 für den ersten Wert, x_2 für den zweiten usw. bis x_n, dem n-ten Wert. Die Summe *y* ließe sich schreiben als

$$y = x_1 + x_2 + x_3 + \ldots + x_n$$

oder viel ökonomischer

$$y = \sum_{i=1}^{n} x_i$$

Man liest dies als: *y* ist gleich der Summe von *i* gleich 1 bis n über alle Elemente x_i. *i* ist eine Lauf- oder Zählvariable und drückt aus, dass alle *ganzzahligen* Zahlenwerte vom ersten Wert, in unserem Fall 1, bis zum letzten Wert *n* genau einmal als Index einer Summandenvariablen in aufsteigender Reihenfolge auftreten. D.h., die Angabe unter dem Summenzeichen legt den Namen dieser Laufvariablen und deren Startwert, die Angabe oberhalb des Summenzeichens deren Endwert fest. Der Endwert muss größer oder gleich dem Startwert sein. Ist der Endwert gleich dem Startwert, so besteht die Summe nur aus einem einzigen Summanden.

Natürlich dürfen sowohl die Laufvariable und der Endwert, soweit sinnvoll, auch „hinter" dem Summenzeichen verwendet werden.

So bedeuten z.B.

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

d.h., bei z.B. 20 Elementen (*i* läuft von 1 bis 20) wird die Summe durch 20 geteilt. Andererseits bedeutet

$$\sum_{i=1}^{n} \frac{x_i}{i} = \frac{x_1}{1} + \frac{x_2}{2} + \frac{x_3}{3} + \ldots + \frac{x_n}{n}$$

Weitere Beispiele:

$$\sum_{i=1}^{n} 1 = 1 + 1 + 1 + \ldots + 1 = n$$

$$\sum_{j=3}^{m} j = 3 + 4 + 5 + \ldots + m$$

Es hat sich eingebürgert, als Bezeichnung für die Laufvariable den Kleinbuchstaben *i* und folgende (bei verschachtelten Summationen) und *n* und folgende Kleinbuchstaben als Symbol für eine beliebig große ganze Zahl zu verwenden.

So bedeutet am nachfolgenden Beispiel einer Doppelsumme

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} = {} & x_{11} + x_{12} + x_{13} + \ldots + x_{1m} + \\ & x_{21} + x_{22} + x_{23} + \ldots + x_{2m} + \\ & x_{31} + x_{32} + x_{33} + \ldots + x_{3m} + \\ & \ldots + \\ & x_{n1} + x_{n2} + x_{n3} + \ldots + x_{nm}\end{aligned}$$

Einen Spezialfall der Summe mit unendlich vielen Summanden (das Symbol ∞ bedeutet unendlich) stellt die unendliche Reihe dar; je nach der Wahl der Summanden kann die Summe endlich (d.h. konvergent), unendlich oder unbestimmt sein. Zum Beispiel ergibt

$$\sum_{i=0}^{\infty} \frac{1}{i!} = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots = e = 2{,}718281828459045\ldots$$

Derartige Summen spielen eine wichtige Rolle bei der Berechnung der Funktionswerte transzendenter Funktionen (Potenzreihen, Taylorreihen), so ist z.B.

$$e^{x} = \exp(x) = \sum_{i=0}^{\infty} \frac{x^{i}}{i!}$$

4. Weiterführende Literatur

- Benninghaus, Hans: Einführung in die sozialwissenschaftliche Datenanalyse. – München: Oldenbourg, [5]1998.
- Benninghaus, Hans: Deskriptive Statistik. – Stuttgart: Teubner, [7]1992. – (Statistik für Soziologen; 1). – (Teubner-Studienskripten; Studienskripten zur Soziologie).
- Benninghaus, Hans: Deskriptive Statistik. – Wiesbaden: VS Verlag für Sozialwissenschaften, [9]2002. – (Studienskripten zur Soziologie).
- Bortz; Jürgen: Lehrbuch der empirischen Forschung für Sozialwissenschaftler. – Berlin: Springer, 1984.
- Diekmann, Andreas: Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. – Reinbek: Rowohlt, [5]1999. – (Rowohlts Enzyklopädie; 55551).
- Sahner, Heinz: Schließende Statistik. – Stuttgart: Teubner, [4]1997. – (Statistik für Soziologen; 2). – (Teubner-Studienskripten; Studienskripten zur Soziologie).
- Sahner, Heinz: Schließende Statistik. – Wiesbaden: VS Verlag für Sozialwissenschaften, [5]2002. –(Teubner-Studienskripten; Studienskripten zur Soziologie).