Bernward Tewes
unter Mitarbeit von
Hans-Joachim Mittag und
Hans-Georg Sonnenberg

Einführung in SPSS

mit Ausblicken auf die freie Statistiksoftware R

kultur- und
sozialwissenschaften

FernUniversität in Hagen

Inhaltsverzeichnis

1 Einleitung

Die Berechnung einer einzelnen statistischen Kennzahl ist bei überschaubaren Datenmengen wie man sie in den Beispielen des Kurses *Statistik* (33209) findet mit keinem großen Aufwand verbunden und kann problemlos im Kopf oder mit Hilfe kleiner Tabellen durchgeführt werden. Ein Taschenrechner ist dabei recht hilfreich.

Aber solche Beispiele sind aus didaktischen Gründen gewählt worden und spiegeln letztlich nicht die Datensätze wider, auf die Sie in der empirischen Praxis treffen werden. Sie werden i.d.R. mehr als ein Merkmal zu untersuchen haben und mehr als eine Handvoll Beobachtungen/Fälle vorliegen haben. Hier ist dann der Computer gefragt.

Wenn man fit ist im Umgang mit einem Tabellenkalkulationsprogramm wie Microsoft Excel, ist ein solches Programm zumindest für die einfacheren statistischen Berechnungen eine Alternative. Es werden einige Funktionen zur Berechnung statistischer Kennzahlen sowie zur grafischen Darstellung angeboten. Sollten allerdings die Kenntnisse im Umgang mit einem solchen Programm nicht so gut sein und/oder die statistischen Analysen zahlreicher und komplexer werden, so ist der Griff zu einem speziellen Statistiksoftwarepaket sinnvoll.

SAS und SPSS

Programme zur statistischen Analyse von Daten sind etwa ab den 60er Jahren des letzten Jahrhunderts entwickelt worden. Von den ersten Vertretern, die in den USA entstanden und (zunächst) für Großrechner entwickelt wurden, besitzen heute nur noch *SAS* (ursprünglich ***S****tatistical* ***A****nalysis* ***S****ystem*) und *SPSS* (ursprünglich ***S****tatistical* ***P****ackage for the* ***S****ocial* ***S****ciences*) eine Marktbedeutung.

Das Programm *SAS* ist damals zunächst zur Analyse von Daten aus der landwirtschaftlichen Forschung an der NC State University, USA, entwickelt worden, das *SAS Institute* wurde 1976 gegründet. Zwar bietet *SAS* auch heute noch eine komplette Umgebung, um statistische Analysen durchzuführen, aber *SAS Institute* sieht sein Produkt eher als umfassendes Instrument für Unternehmensentscheidungen.

Der Ursprung von *SPSS* geht auf Nie, Hull und Bent (Stanford University, USA) zurück, die es ab 1965 entwickelten. 1968 wurde die *SPSS Inc.* gegründet und damit das Produkt kommerziell vertrieben. Mit dem Aufkommen leistungsfähiger Arbeitsplatzrechner sind neben der Großrechnerimplementation eine MS-DOS- und später eine Windows-Version entstanden. Die Windows-Variante bildete dann über viele Jahre den Entwicklungsschwerpunkt. 2008 wurde *SPSS* auf der Basis von Java neu programmiert, was dazu führte, dass nun sowohl für Windows als auch für Mac OS, Linux und weitere Unix-Versionen aktuelle Versionen zur Verfügung stehen. Da man sich ähnlich wie *SAS* auch als Software für betriebswirtschaftliche Entscheidungen im Unternehmen platzieren woll-

te, erhielt das Produkt Anfang 2009 einen neuen Namen: *PASW Statistics* (**P**redictive **A**nalytics **S**oft**W**are). Jedoch blieb dies nur eine kurze Episode, denn mit der Übernahme der *SPSS Inc.* durch *IBM* erhielt das Produkt den Namen *SPSS* zurück und heißt heute *IBM SPSS Statistics*.

S bzw. S-Plus

In den 70er Jahren wurde in den AT&T Bell Laboratories eine Sprache namens *S* explizit für statistische und grafische Anwendungen entwickelt. Ab 1988 erfolgte dann ein kommerzieller Vertrieb dieser Entwicklung mit grafischer Benutzeroberfläche (GUI, **G**raphical **U**ser **I**nterface) als *S-Plus* durch *Statistical Sciences Inc.*, heute wird dieser von *TIBCO Software Inc.* durchgeführt.

Statistica und Stata

In den 80er Jahren sind mit *STATISTICA* und *Stata* zwei weitere Software-Pakete zur Auswertung von Daten mit statistischen Methoden entstanden. *STATISTICA* wurde von der 1984 gegründeten *StatSoft* entwickelt. Es stellt heute eine übersichtliche Benutzeroberfläche (Graphical User Interface, kurz GUI) zur Verfügung und wird ausschließlich für die Windows-Plattform angeboten. *Stata* gehört der 1985 gegründeten *StataCorp* und versteht sich eher als eine Art „statistisches Betriebssystem“, d.h. es bietet eine einfache Programmiersprache. Zwar ist es heute auch mit einer als GUI zu bezeichnenden Menüstruktur zu bedienen, der Schwerpunkt insbesondere auch von Nutzerseite liegt jedoch offenbar immer noch auf der Kommandozeile.

Open-Source-Produkt R

Während alle bisher angeführten Programme zumindest heute kommerziell vertrieben werden, ist in den 90er Jahren in Neuseeland ein Programm entstanden, das unter der GNU General Public License zur Verfügung gestellt wird und somit kostenlos zur Verfügung steht. Unter dem Namen *R* wurde ab 1992 eine Sprache entwickelt, die man als Dialekt von *S* bezeichnen kann (hoher Kompatibilitätsgrad). Bei der Implementation der Sprache in einer Entwicklungsumgebung sind die Entwickler Robert Gentleman und Ross Ihaka (Universität Auckland) aber andere Wege gegangen. Namensgebend für R war der erste Buchstabe der Vornamen der ersten Entwickler. Das Projekt ist schnell gewachsen, seit 1997 gibt es ein internationales Kernentwicklerteam. Als Open-Source-Produkt erhält es zudem über Erweiterungspakete verschiedenster Entwickler zu fast jedem Bereich der Statistik eine passende Prozedur.

Der Abriss zur Entwicklung von Statistiksoftware ist natürlich bei Weitem nicht vollständig. Er beschränkt sich auf mehr oder weniger universell einsetzbare Produkte, die auch heute noch eine Marktbedeutung besitzen. Warum ist nun die Wahl in diesem Kurs auf *SPSS* bzw. *R* als Ergänzung gefallen?

Nun, *SPSS* ist ein umfassendes Statistikpaket mit einer relativ einfach zu bedienenden Oberfläche, ermöglicht aber auch die Nutzung per Syntaxsprache. Ferner bietet es zahlreiche Module zur Erweiterung der schon recht umfangreichen Möglichkeiten der Basiskomponente. Ein neues Modul, *IBM SPSS Statistics Developer*, ermöglicht es nun sogar, *R*-Algorithmen in der *SPSS*-Programmumgebung ausführen zu lassen. Insgesamt ist es dem Programm nicht zu Unrecht gelungen, eine Art Marktführerschaft zu erringen.

R kann man wie *S* als eine statistisch orientierte Programmiersprache mit guten Grafikfunktion ansehen. Der Erfolg von *R* manifestiert sich u.a. auch in der bereits angesprochenen Zugriffsmöglickeit innerhalb von *SPSS* auf *R*-Routinen sowie in der Verwendung als Softwarebasis für das Statistiklabor der Freien Universität Berlin, das sich als didaktische Statistik-Software versteht (www.statistiklabor.de). Da man mit *R* i.d.R. auch dann noch eine Lösung finden kann, wenn z.B. SPSS keine (offensichtliche) Möglichkeit mehr bereitstellt, sei es über die Erweiterungspakete oder selbst programmiert, ist es häufig mehr als nur eine Alternative. Die kostenlose Bereitstellung sowie die Tatsache, dass *R* zunehmend an Boden gewinnt und *R*-Kenntnisse heute häufiger schon als Merkmal in Stellenprofilen gewünscht werden, sind weitere Gründe, hier darauf einzugehen.