Statistik-Schulung für Azubis unter Verwendung des SAS-Research Analyst 3.0

Silke Wurzinger

Knoll-AG Ludwigshafen
Telefon: 0621 / 589 1948
eMail: silke.wurzinger@knoll-ag.de

Abstract

Die Knoll-AG bildet ihre Lehrlinge in der 3 1/2 jährigen Ausbildung zum Biologielaboranten unter anderem in EDV und Biometrie aus. Den zukünftigen Biologielaboranten soll in drei Wochenkursen das Handwerkszeug vermittelt werden, um im späteren Laboralltag die anfallenden Daten elektronisch zu verarbeiten und den Biometriker bei der statistischen Auswertung zu unterstützen.
Im ersten Wochenkurs werden den Azubis Grundlagen von MS Office beigebracht. Ziel des zweiten und dritten Kurses ist es, Zusammenhänge der Versuchsplanung zu begreifen und Beschreibende und Teststatistik nicht nur theoretisch, sondern auch praktisch durchzuführen. Eine realitätsnahe Ausbildung erhöht erfahrungsgemäß die Akzeptanz und den Nutzen im späteren Arbeitsumfeld. Deswegen erfolgt die Datenerfassung während des Kurses hauptsächlich in Excel. Obwohl eine Anzahl von Statistiken ebenfalls mit Excel durchgeführt werden könnte, legt die Knoll-AG Wert auf eine Ausbildung mit einem von der FDA (Food and Drug Administration) anerkannten System.
Trotz Berührungsängsten mit der englischen Sprache und dem Windows-ähnlichem aber nicht völlig identischen Aufbau, bietet die AF-Oberfläche von SAS-RA einen möglichst einfachen Einstieg in die Welt von SAS. Eine menügesteuerte Auswahl von Analysen, die Schnittstelle zu Excel, die thematische Gruppierung der Statistikbereiche, die Interpretationshilfen der analytischen Statistik, die Dokumentation des Programmablaufes und der gewählten Parametern, sowie ein ausführlicher Output sind einige der Gründe, die für die Ausbildung mit SAS-RA sprechen. Anhand von Musterbeispielen soll die Applikation und die damit mögliche Ausbildung dargestellt werden.

Warum überhaupt Statistik für Azubis?

In Forschung und Entwicklung wird man zwangsläufig mit Statistik konfrontiert, auch wenn man kein Statistiker ist. Nun kann es natürlich nicht Sinn einer modernen Ausbildung sein, z.B. dem Tierpfleger die Vorzüge der logistischen Regression nahe zu bringen. Dennoch fordert die einerseits immer größere Spezialisierung, andererseits die immer komplexeren Abläufe ein Verständnis für das Gesamtkonzept von der Versuchsplanung bis zur Auswertung. Bei BASF Pharma werden Biologielaboranten in ihrer vierjährigen Ausbildung an die verschiedensten Arbeitsgebiete herangeführt. Sie werden in ihrem Beruf schwerpunktmäßig Versuche nach Anleitung vorbereiten, durchführen und protokollieren. Ziel der Ausbildung in EDV & Biometrie ist es, den Umgang mit EDV-Systemen zu erlernen, um die Daten später in geeigneter Form richtig zu erfassen, so dass sie direkt ausgewertet, bzw. an weiterführende Stellen portiert werden können.

1. Ausbildungsinhalte

Der EDV-Unterricht findet in drei Kursen mit je einer Woche Vollzeit statt.
Kurs 1 - Einführung in MS-Office:
Betriebssystem, Textverarbeitung, Tabellenkalkulation, Mailsysteme/Web werden den Azubis in Grundzügen beigebracht.

Kurs 2 - Biometrie I mit Excel und SAS-RA:
Deskriptive Statistik, Dateneingabe, Import, Export sind die Schwerpunkte im ersten Biometriekurs.
Kurs 3 - Biometrie II mit Excel und SAS-RA:
Schließt mit Deskriptiver und Analytischer Statistik die Ausbildung in diesem Bereich ab.

1. Ziele des Statistikunterrichts

Die Hauptziele des Kurses sind, dem angehenden Biologielaboranten statistische Verfahren verständlich zu machen damit er diese korrekt durchführen kann. Die Azubis lernen zunächst, die richtige Auswertung zu wählen, die notwendigen Rechenoperationen per Hand (und Taschenrechner) durchzuführen sowie auch mit speziellen Programmen umzugehen und diese richtig zu bedienen. Als Programme wurde einmal Excel gewählt, weil dies das Produkt ist, das an den meisten Laborarbeitsplätzen installiert ist. Um auch ein von der FDA anerkanntes Programm in den Unterricht zu integrieren, fiel die Wahl außerdem auf SAS. Um die Bedienung für Gelegenheitsbenutzer zu ermöglichen, entschied sich die Knoll AG für den Research Analyst.

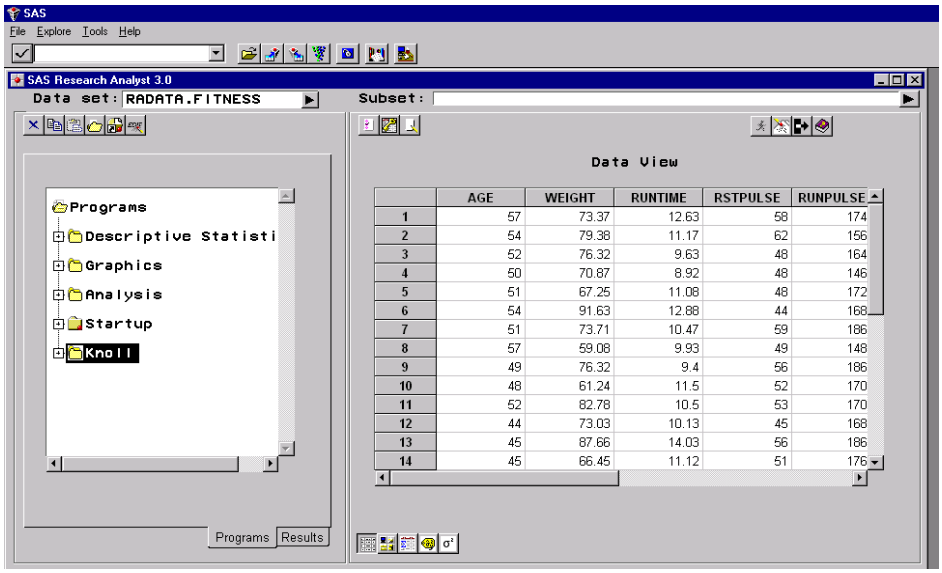
2. Ablauf des Statistikunterrichts

Aller Anfang ist die graue Theorie, sowie die Berechnung mit Taschenrechner. Dies benötigen die Azubis zum einen, um Zugang zu den Formeln zu erhalten, aber auch als prüfungsrelevante Fertigkeiten für das Abschlussexamen.
Je nach Stoffgebiet steigen wir in die Auswertung am PC mit Excel oder SAS ein. Während eine Histogramm einer Häufigkeitsverteilung mit Excel einfach und schön zu erstellen ist, erhält man z.B. den χ^2-Test mit SAS-RA um ein vielfaches leichter. Meist werden beide Möglichkeiten durchgeführt, verglichen und interpretiert. Normalverteilungsüberprüfung und einige Tests, wie z.B. Varianzanalyse werden nur mit SAS durchgeführt. Ein wichtiger Bestandteil des Unterrichts ist der Import und Export. Denn selbst wenn die Biologielaboranten später Daten nicht selbst auswerten, so sollen sie zumindest einige Grundzüge der Dateneingabe kennenlernen, um später bei der Übergabe ihrer Daten an Dritte einen möglichst verlust- und reibungsfreien Arbeitsablauf zu ermöglichen.

2. SAS-Research Analyst (RA)

1. Programmaufbau

Das **Eingangsbild** des präsentiert sich wie in Abbildung 1 gezeigt:



In der linken Hälfte des RA-Fensters kann die Auswertedatei selektiert werden, sofern sie bereits als SAS-Datei vorliegt. Die Wahl der gewünschten Auswerteprozedur – in RA als *Programs* bezeichnet - findet man im darunterliegenden Menübaum.
Im Karteiblatt *Programs* sind die Prozeduren logisch in *Descriptive Statistics*, *Graphics* und *Analysis* zusammengefasst. Die Ergebnisse können später jederzeit unter dem Karteiblatt *Results* abgerufen werden.
Die rechte Hälfte des Fensters enthält zunächst einen *View* auf die aktuellen Auswertungsdaten. Hier können Daten auch eingegeben oder verändert werden. Eine Datensatzselektion wird durch *Subset* ermöglicht.
Wird ein *Program* ausgewählt, wird im rechten Teil des Fensters eine Maske sichtbar, die je nach *Program* unterschiedliche Eingaben erfordert. Nach Auswahl der Variablen und Optionen wird mit dem *Run* Button das Programm abgeschickt und RA zeigt automatisch das Ausgabefenster an. Tritt während des Laufes ein ERROR oder eine WARNING auf, wird der Blick in das *Log* Fenster gewährt. Dies ist allerdings für den SAS-Neuling eher verwirrend als eine Hilfe, da er mit den Statements und der damit verbundenen Fehlermeldung in den seltensten Fällen etwas anfangen kann.
Abgerundet wird die Oberfläche durch Icons wie z.B. für den Import / Export. Über allem steht eine SAS-Menüleiste, die je nach gewähltem Programm unterschiedliche Auswahlmöglichkeiten bietet.

2. Datenselektion

SAS-RA verarbeitet externe Daten mit spezieller Dateistruktur, wie Excel, ASCII, dBASE. Er kann SAS-Dateien einlesen, oder aber die Daten können direkt in SAS-RA eingegeben werden. Während der Import mit dem Import Wizard problemlos geht, sofern die externe Datei portierfähig gestaltet ist, ist die Dateneingabe das schwächste Glied des RA. Zwar ist die Eingabe - ähnlich wie in Excel - in einem Spreadsheet möglich, jedoch die Variablendefinition ist relativ unflexibel. Einmal angelegte Variablen können nicht gelöscht werden, bereits als numerisch gespeicherte Variablen nicht mehr in character umgewandelt werden. Das Löschen von Datensätzen ist nur Datensatz für Datensatz und nicht im Block möglich. All dies sorgt dafür, dass für das Eingeben eine ausführliche Einweisung erforderlich ist. Abgeleitete Variablen, wie die Differenz zweier Daten können im Spread-Sheet nicht berechnet werden.

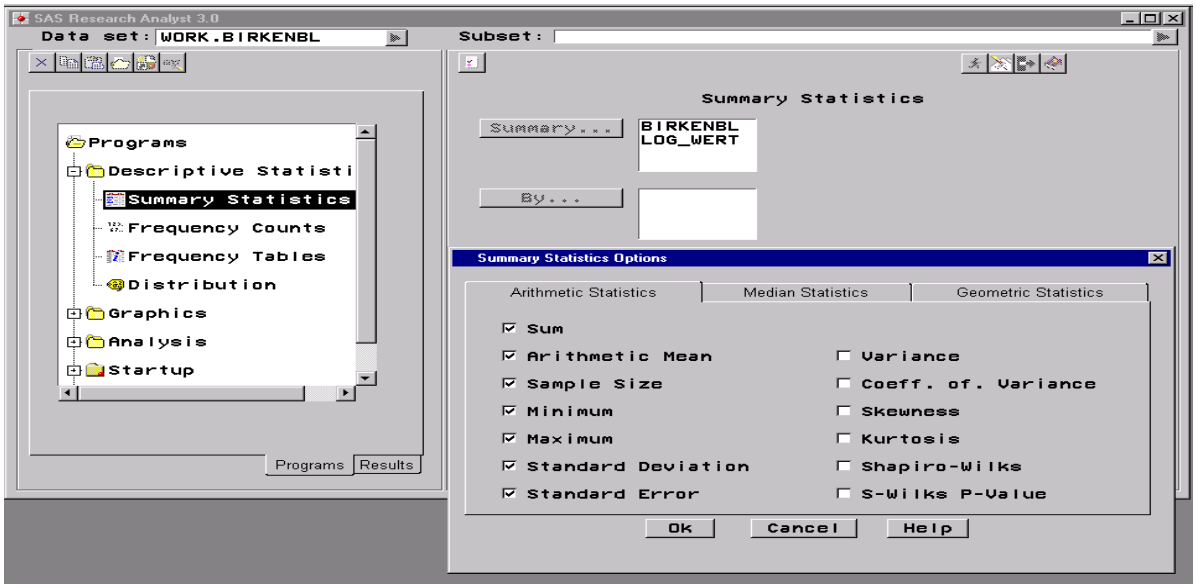
Ist die SAS-Datei angelegt, können mit der Option *Subset* Untergruppen gebildet werden. Das *Subset* entspricht dem WHERE-Statement. Die Formulierung des Statements kann sowohl manuell als auch menügesteuert erfolgen. Bei der manuellen Selektion heißt es zunächst Abfrageregeln zu vermitteln. Wenn geschl=m eine Fehlermeldung produziert (ganz klein und harmlos in der Statusleiste) und geschl='m' andere Ergebnisse liefert, als geschl='M', ist das für einen blutigen Anfänger die erste Herausforderung. Die Menüsteuerung ist da eine gute Hilfe, wenngleich die Selektion der einzelnen Abfragebestandteile etwas langwieriger ist.

3. Programm-Module

Im RA enthalten sind eine Reihe von aufbereiteten Prozeduren aus den SAS-Modulen SAS-BASE, SAS-GRAPH, SAS-STAT. Jedem *Program* ist ein Icon zugeordnet. Bei der Fülle an Prozeduren bleibt es leider nicht aus, dass nicht alle Icons unbedingt selbsterklärend sind. So ist es gut, dass SAS immer auch eine Wahl über den Menübaum ermöglicht.

Der Menüpunkt *Descriptive Analysis* setzt sich aus den Programs *Summary* (PROC UNIVARIATE mit OUT-Option), *Frequency Counts* (PROC FREQ mit nur einer Analysevariablen), *Frequency Tables* (PROC FREQ mit zwei gekreuzten Analysevariablen und der Möglichkeit des χ^2-Testes), *Distribution* (PROC UNIVARIATE ohne PLOT-Statement) zusammen.

Exemplarisch ist im folgenden das Auswahlformular der *Summary* mit Optionsfenster in Abbildung 2 angezeigt:



Die Auswahl der notwendigen Analysevariablen und Optionen kann größtenteils intuitiv erfolgen, sobald die Grundidee dem Benutzer transparent gemacht worden ist. Einen Vorteil bietet SAS dadurch an, dass nur die Variablen zur Auswahl stehen, die per Definition für die Analyse geeignet sind, z.B. nur numerische Variablen bei Summary. Dies reduziert die Fehlerquote erheblich.
Die Ausgabe der Berechnungen erfolgt in ein separates Fenster. Pro Programmlauf erstellt SAS eine *Code*-Datei mit dem Programm-Code, eine *Log*-Datei, sowie das automatisch angezeigte *Output* Fenster. Diese Fenster werden temporär unter *Results* in chronologischer Reihenfolge gespeichert und können über Prozedurname und Zeitpunkt des Laufes immer wieder angewählt und angesehen werden. Für versierte SAS-Benutzer sind *Log* und *Code* recht nützliche Beigaben. Auf SAS-Einsteiger ohne Programmiererfahrung haben diese Fenster eher eine abschreckende Wirkung. Der Inhalt der Fenster kann ausgedruckt, oder aber in einer permanenten Datei gespeichert werden.
Das *Output* Fenster enthält neben den puren Auswertungsdaten alles, was eine gut dokumentierte Ausgabe benötigt: Den Namen des ausgewählten *Programs*, Datum und Uhrzeit, Dateiname und sofern verwendet, einen Hinweis auf Datensatzselektion.

Analog sind die graphischen Möglichkeiten in folgenden Programmpunkten verwirklicht: *Scatter Plot*, *Bar Chart*, *Range Plot*, *Regression Scatter Plot*, *Surface Plot, Contour Plot, Normal Probability Plot*. Anders als bei der deskriptiven Statistik haben hier SAS-Einsteiger eher Schwierigkeiten die Formulare auszufüllen. So ist z.B. beim Bar Chart möglich *Independent, Summary, Subgroup, Group* und *By* auszufüllen. Erst wenn alle gewünschten Einstellungen vorgenommen sind und der *Run* Button getätigt wurde, hat der Benutzer einen Einblick, welche Art von Graphik er soeben erstellt hat. Dies ist einer der Gründe, warum Azubis Graphiken made by Excel den Vorzug geben.

Die Analytische Statistik wird von SAS-RA in *Regression Analysis*, *General Linear Models*, der Palette von *t-Tests*, sowie der Auswahl einiger *Non-parametric Tests* aufgeteilt. Das besondere an den Tests in RA ist es, dass bereits in der Eingabemaske Null- und Alternativhypothese formuliert werden. Je nach Seitigkeit der Fragestellung berechnet SAS anhand eines 5% Signifikanzniveaus, ob der Test signifikant geworden ist oder nicht und gibt neben dem p-Wert das Ergebnis in Worten wieder. Aussagen wie *'highly signifcantly'* bei $p < 0.001$, oder bei p-Werten die knapp über der 0.05-Grenze liegen: *'Note that p= 0.051 is close to the 5% significance level.*', oder sogar der Hinweis darauf, dass der t-Test nicht korrekt gewählt wurde: *'Test for normality of sample: p= 0.008. The data is not normally distributed. A Wilcoxon Test may be more reliable'* runden das Ganze ab.
Ein Problem stellt die Auswertung dann dar, wenn auf 1% Signifikanzniveau verglichen werden soll. SAS hat hierfür keine Option – die Interpretationshilfe ist dann sinnlos.
Neben dem aufbereiteten Ausgabefenster *Summary* bietet SAS noch ein weiteres Ausgabefenster *Output* an, in dem die Prozedur in der Form ausgegeben wird, die man beim 'normalen' Programmieren bekommen würde.

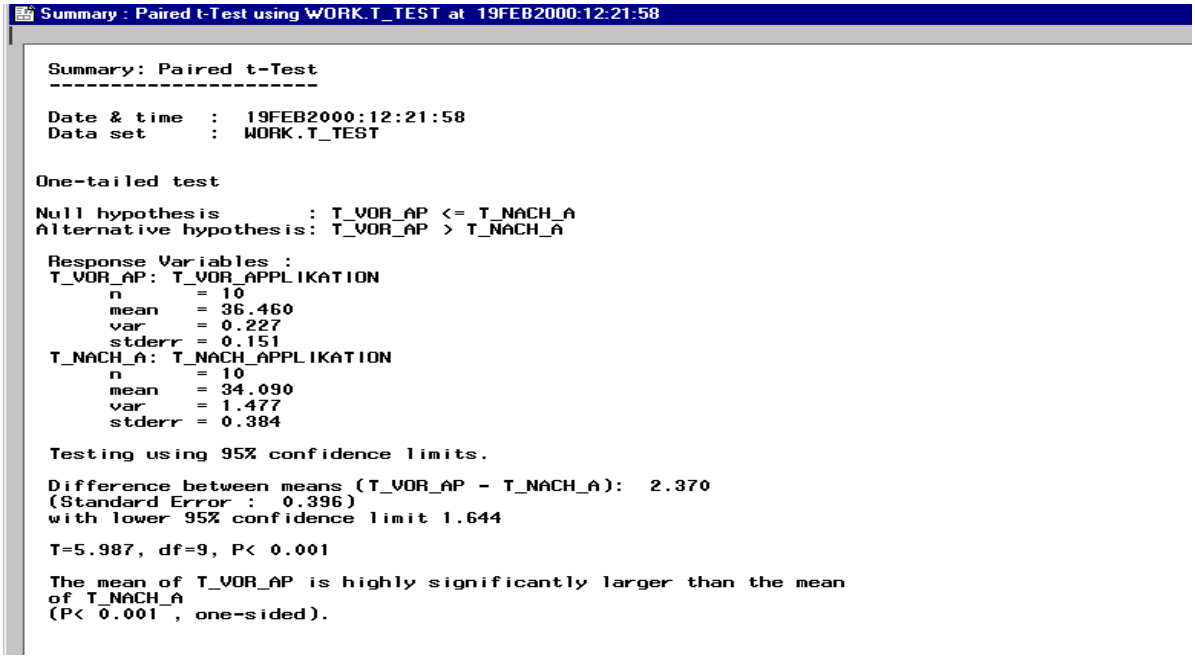
4. Beispiel einer Unterrichtslektion

Exemplarisch für die verschiedenen Kapitel des Statistikunterrichts soll im folgenden die Auswertung eines t-Tests mittels SAS-RA dargestellt werden.
Im Rahmen des Lehr-Kapitels 'Parametrische Tests' ist folgende Fragestellung zu beantworten: Aufgrund theoretischer Überlegungen vermutet man, dass bei einer Langzeitnarkose bei Ratten die Körpertemperatur sinkt. An 10 Ratten wird vor der Applikation und 7 Stunden später die Temperatur gemessen (nach Keller, S. 206, Übung 50).
Die Daten liegen bereits in geeigneter Form in Excel vor. Es wurde sowohl die Temperatur vor *t_vor_ap* (vorher), und die Temperatur am Ende der Betäubung *t_nach_a* (nachher) gemessen, sowie die Differenz *differen* (nachher-vorher) gebildet. Im ersten Schritt werden die Daten importiert und per *Data View* betrachtet. Beim Vergleich zweier Messwerte am gleichen Individuum handelt es sich um eine verbundene Versuchsanordnung. Da bereits eine Absenkung der Temperatur vermutet werden kann, wird einseitig mit einem Signifikanz-niveau von 5% geprüft.

Einzige Schwierigkeit ist, aus den drei t-Tests den richtigen zum vorhandenen Datenmaterial auszuwählen. Während der *Unpaired t-Test* ziemlich schnell als unverbundener t-Test identifiziert wird und damit ausscheidet, bleibt die Qual der Wahl zwischen *Paired t-Test* und *Single-sample t-Test*.

Arbeitet man mit den Datenpaaren vorher - nachher ist die Eingabemaske des *Paired t-Test* geeignet. Hier wird der Benutzer um die Eingabe der zwei *Response*-Variablen gebeten. Weiterhin ist das Hypothesenpaar zu formulieren. SAS gibt automatisch die Nullhypothese vor: *vorher = nachher*. Erst nachdem die Alternativhypothese einseitig formuliert worden ist: *vorher > nachher* und das Programm abgelaufen ist, wird die Nullhypothese auf eine einseitige Fragestellung umgestellt. Dies ist sicherlich nicht ganz optimal programmiert. Als Option kann gewählt werden zwischen dem 95% oder 99% Konfidenzintervall.
SAS präsentiert als Ergebnis deskriptive Statistik der beiden Variablen, mit Mittelwert, Varianz und Standardfehler, sowie das Ergebnis des t-Testes mit Testgröße, Freiheitsgraden und p-Wert. Als Interpretationshilfe ist die Signifikanz-Aussage angefügt. *The mean of vorher is highly significantly larger than the mean of nachher (p<0.001, one-sided).*

Wie die Ausgabe der RA aussieht, ist in Abbildung 3 dargestellt:



Soll direkt die Differenz-Variable ausgewertet werden, wird mit Hilfe *des Single-sample t-Test* die Differenz-Variable als *Response* angegeben und mit einer Konstanten - hier in diesem Fall dem Wert 0 - verglichen. Die Nullhypothese wird wiederum automatisch von SAS als *Differenz = 0* vorgegeben. Da die Alternativhypothese einseitig formuliert wird, passt SAS-RA nach Lauf des Programmes die Nullypothese an. Die Ausgabe ist analog dem *Paired t-Test*.

Zum Abschluss werden die Daten vergleichsweise mit Excel analysiert. Anders als bei SAS muss das Hypothesenpaar nicht formuliert werden. Als Ergebnis liefert Excel den p-Wert bei einseitiger und zweiseitiger Fragestellung. Der Benutzer kann sich anschließend den - hoffentlich - richtigen heraussuchen. Eine Interpretationshilfe besitzt Excel nicht.

3. Beurteilung des SAS-RA

SAS RA bietet einiges zur Unterstützung des Unterrichts in Statistik. Es ermöglicht auch relativ unbedarften Benutzern die Auswertung ihrer Daten. Trotz Menüführung bleibt es nicht aus, Grundzüge von SAS zu vermitteln. Ob es die 8-Zeichen Einschränkung bei Variablennamen ist, oder das plötzliche Auftauchen einer Fehlermeldung im Log Fenster: Nicht immer gelingt es, alle Klippen im Umgang mit SAS zu umschiffen. Als negativ wird von den Azubis vor allem die englische Oberfläche, und die nüchterne Ausgabe empfunden. Insgesamt erleichtert SAS-RA jedoch den Einstieg in die Welt von SAS enorm und lässt auch in relativ kurzer Zeit Erfolge bei der Auswertung von Daten zu.

Literatur

Keller F, Statistik für naturwissenschaftliche Berufe, 4. Auflage, pmi Verlagsgruppe GmbH, Frankfurt 1993

Paulus M, Der SAS Research Analyst – Ein Grundkurs, 1. Ausgabe, SAS Institute GmbH, 1994