Lehr- und Handbücher der Statistik

Herausgegeben von
Universitätsprofessor Dr. Rainer Schlittgen

Bisher erschienene Werke:

Caspary/Wichmann, Lineare Modelle
Chatterjee/Price (Übers. Lorenzen), Praxis der Regressionsanalyse, 2. Auflage
Degen/Lorscheid, Statistik-Aufgabensammlung
Harvey (Übers. Untiedt), Ökonometrische Analyse von Zeitreihen, 2. Auflage
Harvey (Übers. Untiedt), Zeitreihenmodelle, 2. Auflage
Heiler/Michels, Deskriptive und Explorative Datenanalyse
Naeve, Stochastik für Informatik
Oerthel/Tuschl, Statistische Datenanalyse mit dem Programmpaket SAS
Pokropp, Lineare Regression und Varianzanalyse
Rinne, Wirtschafts- und Bevölkerungsstatistik
Schlittgen, Statistik, 5. Auflage
Schlittgen/Streitberg, Zeitreihenanalyse, 5. Auflage

Stochastik für Informatik

Von
Universitätsprofessor
Dr. Peter Naeve
Statistik und Informatik
Fakultät für Wirtschaftswissenschaften
Universität Bielefeld

R. Oldenbourg Verlag München Wien

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Naeve, Peter:
Stochastik für Informatik / von Peter Naeve. - München ;
Wien : Oldenbourg, 1995
(Lehr- und Handbücher der Statistik)
ISBN 3-486-23243-6

Gesamtherstellung: R. Oldenbourg Graphische Betriebe GmbH, München

ISBN 3-486-23243-6

Vorwort

Eigentlich wollte ich diesem Werk den Titel **Statistik für Informatik** geben. Doch nach Fertigstellung der vorliegenden Kapitel mußte ich feststellen, daß es mir nicht gelungen ist, empirische Daten und deren Analyse aufzunehmen, nachdem zuvor Analyseverfahren ausreichend diskutiert wurden. Diesen Mangel bedaure ich sehr — vielleicht findet dieses Buch trotzdem genügeng freundliche Leser, so daß eine künftige Auflage Gelegenheit gibt, das Fehlende nachzuholen. So wie das Manuskript jetzt vorliegt, hat es eindeutig einen Schwerpunkt in der Modellierung stochastischer Phänomene in der Informatik. Dieser Schwerpunktsetzung wird der jetzt gewählte Titel sicher besser gerecht.

Ein Blick auf Vorlesungsverzeichnisse, Lehrbücher und viele Informatikjournale kann den Eindruck erwecken, als ob Stochastik und Statistik eher unbedeutende Randgebiete in der Informatik wären. Ich hoffe, daß der vorurteilslose Leser nach kritischer Lektüre des Buches für sich diesen Eindruck korrigiert.

Dieses Buch ist für Studenten geschrieben. Allerdings möchte ich nicht verhehlen, daß ich nach dem Besuch der einen und anderen Informatikertagung vorsichtig anmerken würde, daß ich dort so manch einen (gestandenen) Informatike getroffen habe, dem die Lektüre auch etwas bringen könnte.

Das Manuskript ist aus einer Vorlesung hervorgegangen, die ich unter dem Titel **Statistik für Betriebsinformatik** seit mehreren Jahren an der Fakultät für Wirtschaftswissenschaften der Universität Bielefeld gehalten habe. Der Titel der Lehrveranstaltung spiegelt die Bezeichnung des Faches im Hauptstudium wider. Die Inhalte sind aber auf Informatik und nicht auf Betriebswirtschaftslehre ausgerichtet.

Vor der Veröffentlichung des vorliegenden Buches im Oldenbourg Verlag — Herrn Kollegen Schlittgen sei für die auffordernde Ermunterung zum Eintritt in die Reihe „Lehr- und Handbücher der Statistik" herzlich gedankt —, gab es eine Reihe handschriftlicher Fassungen des Manuskriptes, ehe ich mich zu der von den Studenten geforderten Skriptform durchringen konnte. Den Grund meines Zögerns mag man dem folgenden Zitat aus dem Vorwort zur Version 2.0 des Skriptes entnehmen.

> Ich habe lange gezögert, ein „Skript" zur Statistik in Betriebsinformatik zu veröffentlichen. Zuerst war es vor allem die Unsicherheit bezüglich Inhalt, Reihenfolge und Stil, die mich zurückhielt.

> Die Kopfzeilen in meinem handschriftlichen Manuskript zeugen von fortwährender Umstellung im Verlauf der Semester. Unterschiedliche Schreibwerkzeuge (rote, blaue und grüne Tinte, diverse Kugelschreiber und Bleistifte) deuten auf häufige Umarbeiten und Ergänzungen hin. Langsam aber schien sich die Unsicherheit zu geben.
>
> Nun aber war das Argument stark, ob man denn überhaupt den Hörer zum Leser machen solle. Würde nicht der Student das Skript mißverstehen als die endgültige Statistik für Betriebsinformatik? Was als wirkliche „Einführung" gemeint wäre, drohe als „Einengung", als „Eingrenzung" zu verkommen. So schien es mir.
>
> Im letzten Sommersemester aber änderte sich meine Ansicht plötzlich total. Da war das Verfassen eines Skriptes die „Droge", die mir den Streß eines Semesters mit vielen Verpflichtungen überstehen half. Die Version 1.0 habe ich nur für mich geschrieben.
>
> Zu Beginn dieses Jahres schien sie mir dann doch nicht so schlecht geraten zu sein. Eine Überarbeitung allerdings war angebracht, damit aus *meinem* Skript auch ein Skript für *Studenten* werden kann.

Inzwischen bin ich bei meiner internen Zählung bei der Version 2.575. Diese auf den ersten Blick etwas merkwürdige Zählung wird durch ein Zitat aus dem Vorwort der Version 2.5 deutlich.

> Die Versionsnummer 2.5 dieser neuen Fassung läßt vermuten, daß der Übergang von der Version 2.0 zur Version 2.5 nur eine kleine Änderung gegenüber der vorhergehenden Fassung bedeutet. Oberflächlich betrachtet, ist dies auch richtig. Im wesentlichen die Beseitigung einiger (auch peinlicher) Fehler, was ist das schon groß. Vielleicht ist es ja nur der nun vorhandene *Index*, der für den Leser das Vergeben einer neuen (kleinen) Versionsnummer rechtfertigt. ...
>
> Immerhin, wenige Seiten zusätzlichen Text gibt es im Kapitel 7. Ein erster Versuch, der bislang fehlenden Numerik einen Platz einzuräumen. Aber natürlich völlig unzureichend, denn wer sich der Verteilungsfunktion der Normalverteilung annimmt, der darf auch die Inverse nicht vergessen.
>
> Alles wird reumütig eingestanden. Leider haben *ventis adversis* mehr verhindert. Alle in der vorigen Fassung eingeräumten Defizite bestehen weiter. Neue sind hinzugekommen, Stichwort „total quality". Sie werden auch beseitigt werden. Ehrenwort[1]! Ein Kapitel über Simulation regt sich

[1] Vielleicht graut es ja mit Hinblick auf die Klausuren — siehe Bemerkung in Kapitel 10 auch manchem Leser vor der Einlösung des Versprechens.

> bereits im Kopf des Verfassers. Aber es wollte noch nicht heraus[2]. Der ungeduldige Leser muß sich auf die nächste Version vertrösten lassen. Der Verfasser ist guter Hoffnung, daß das Werk in der Abfolge der Versionen sich seinem Ideal mit 99.5 % nähert. Fürwahr eine hehres Ziel.
>
> Um sich immer daran zu erinnern, hat der Verfasser beschlossen, an der Versionsnummer den Grad der Zielerreichung ablesbar zu machen. Wie?? Natürlich standesgemäß über die Vorgabe der Versionsnummer λ des idealen Werkes, die sich als Lösung der Gleichung[3]
>
> $$\lambda = \Phi^{-1}(.995)$$
>
> ergibt.

Wie schon gesagt, das Buch ist in erster Linie für Studenten geschrieben. Sie sollten sich daher auch besonders zu Kritik und Kommentar, Wunsch und Anregung aufgefordert fühlen. Wie die nachstehenden Zitate aus diversen Vorworten zeigen, machten die Studenten in Bielefeld leider keinen sehr regen Gebrauch von dieser Einladung.

> Alle Fehler sind nur dem Autor anzulasten.
>
> Damit ihre Anzahl im Laufe der Versionen abnimmt, bitte ich alle Studenten, mir freudig – wann darf man schon einmal einen Professor korrigieren – alle gefundenen Fehler mitzuteilen. Leider folgten dieser Aufforderung bei der Version 1.0 nur 5 von 120 Studenten. Dabei gab es Fehler beinahe satt, denn der Autor ist ein lausiger Tipper.
>
> Aber nicht nur kleinliche Fehler-Beckmesserei ist gewünscht. Anregungen und Verbesserungsvorschläge werden späteren Studentenjahrgängen zugute kommen. Dem Autor aber geben sie die Gewißheit, auch gelesen worden zu sein. Und das tut gut, denn etwas eitel darf man ja wohl sein! (aus Version 2.0.)
>
> Es ist keine Phrase, wenn auch in dieser Fassung der Leser herzlich zu Kritik und Anregungen eingeladen wird. Leider ließ sich die Zahl der Personen, die sich als kritische Leser offenbarten, noch fast an einer Hand abzählen. Der Verfasser mag nicht glauben, daß ihm sein Werk so unübertroffen gut gelungen ist, daß sich nichts mehr dazu anmerken ließe. Aber auch das Gegenteil erscheint ihm nicht richtig zu sein. Es kann doch nicht so schlecht sein, daß man besser darüber den Mantel des

---

[2] Hephaistos hilf!

[3] Weiß da etwa jemand nicht, was Φ bedeutet? Wie wäres es mit einem Blick ins Kapitel 6 des Skriptes?

> Schweigens ausbreitet. Also Kritiker aller Semester stürzt euch auf das Werk. (aus Version 2.5.)
>
> Der Verfasser hat sich bemüht kleine Fehler, gegenüber der vorigen Version zu beheben. Es werden sicher noch viele im Text verborgen sein. Leider ist die Rückmeldung durch die Studenten immer noch fast gleich Null. Bitte, bitte, meldet die Fehler und Unklarheiten. Der Autor verspricht auch feierlich, nicht beleidigt zu sein, wenn das Aufzeigen eines Fehlers mit einem spöttischen oder tadelnden „na, na" begleitet wird. (aus Version 2.57.)

Ich bedanke mich schon jetzt für alle eingehenden Bemerkungen und Stellungnahmen der Leser. Aber es ist mir auch ein Bedürfnis, eine Anerkennung für bereits erhaltenen Unterstützung auszusprechen. Meine Mitarbeiter Bernhard Strohmeier und Peter Wolf verdienen meinen besonderen Dank. Vom Inhalt bis zum Layout haben sie meine Arbeit kritisch unterstützt. Rückschauend kann ich nur immer wieder feststellen, es war ein Genuß, von ihnen kritisiert zu werden. Wenn ich auch gestehen muß, daß es für einen Autor oft schwer auszuhalten ist, ein gerade frisch aufgeschriebenes Stück „zerpflückt" zu sehen. Ich hoffe, der Leser wird durch eine positive Aufnahme des Buches auch die segensreiche Mitwirkung der beiden indirekt anerkennen. Andreas Handl hat durch die Übernahme der Lehrveranstaltung mir nicht nur die Möglichkeit eines Freisemesters eröffnet, sondern auch zahlreiche Verbesserungsvorschläge und Erweiterungswünsche gemacht. Nicht allen habe ich mich angeschlossen. Vermutlich wird mich der eine oder andere Leser dafür tadeln. Richtig so, denn wie heißt es korrekt: *Alle Fehler sind dem Autor anzulasten.*
Der Kenner hat schon lange gesehen, daß dieses Buch mit Hilfe des TEX-Systems erstellt wurde. Genauer habe ich aus dem Public Domain Softwaresystem emtex die Komponente LATEX2e benutzt. Die Packages babel, makeidx, amstex, amsthm, amssymb, amsxtra, minitoc, euler und beton halfen bei der Erstellung eines dem Autor gefallenden Layouts. Es ist hier aber auch der Platz einmal kritisch auf die Veränderungen im Buchdruck hinzuweisen. Die Verweise auf LATEX etc. bedeuten auch, es hat kein Setzer und auch kein klassischer Drucker mehr an diesem Buch mitgewirkt. Und es sind sicher auch noch weitere Arbeitsplätze durch diese Verlagerung der technischen Bucherzeugung zum Autor hin eingespart worden. Ich hoffe, der Leser wird es wenigstens an einem akzeptablen Buchpreis merken.
Habe ich die Version 1.0 in nächtlichen Aktionen in den Rechner ge- und vertippt, konnte ich mich bei der Überarbeitung und Erweiterung auf meine Sekretärin, Frau Karin Wüstenbecker, stützen. Es ist dem Output in jeder Hinsicht gut bekommen.
Ich habe außerdem noch eine Reihe weiterer Softwarehilfen benutzt. Einige Passagen im Literate-Programming-Stil wurden mit Hilfe von *APL*2WEB erstellt. *APL*2WEB wurde von Christoph von Basum im Rahmen seiner Dissertation erstellt. Zur Druck-

legung wurde das von Bernhard Strohmeier erstellte Package apl2itan verwandt. Das benutzte *APL*-System ist APL*PLUS ver. 7.1 von Manugistics Inc.(vormals STSC). Bernhard Strohmeier pflegt an dieser Stelle immer seinen Spott über mich auszugießen, ob meiner Vorliebe für alte Versionen. Meine Antwort: *never change a running system.*
Andere Passagen im Literate-Programming-Stil und S-Plus wurden mit einem im Rahmen einer Diplomarbeit von Wilhelm Kuhlmann erstellten C-Programm oder einer von Lorenz Lang angepaßten Spidery WEB Version angefertigt. S-Plus ist ein Produkt von StatSci Inc.
Einige Zeichnungen wurden mit GNUPLOT erstellt. GNUPLOT wird angeboten unter GNU General Public License.
Nach diesen vielen technischen Anmerkungen und Hervorhebungen möchte ich am Schluß meine Frau um Entschuldigung für die vielen Nächte bitten, in denen ich sie mit einem Rechner betrog. Ihr Verständnis hat erst dieses Buch geschaffen.

Peter Naeve

Inhaltsverzeichnis

Kapitel 1

Einleitung

Inhaltsangabe

1.1 Warum Statistik für Informatik?

Die Antwort auf die in der Überschrift gestellte Frage wird hier in Form vieler Beispiele gegeben. Allen diesen Beispielen ist gemein, daß sie Situationen aus dem Bereich der (angewandten) Informatik beschreiben, die man ohne Statistik nicht meistern kann.

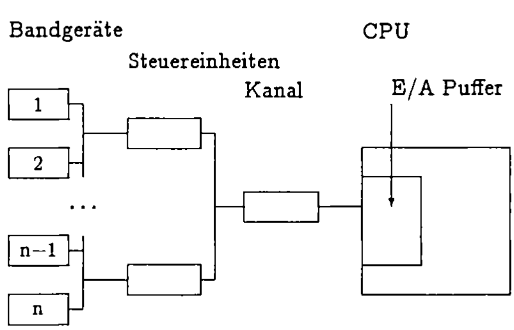
Beispiel 1: *Analyse von Algorithmen*

Der folgende Ausschnitt aus einem Pascal-Programm löst das Problem, das (ein) maximale(s) Element in einem Vektor x zu suchen.

```
          ...
          j := n;
          xmax := x[n];
          k := n-1;
          while k <> 0 do
            begin
*             if x[k] > xmax then
                begin
**                xmax := x[k];
**                j := k;
                end;
              k:= k-1;
            end;
          ...
```

Wie häufig werden die beiden mit ** markierten Anweisungen ausgeführt? Das hängt offensichtlich ab vom Ausgang der mit * markierten Anweisung. Und dort kommt mehr von der Datensituation ins Spiel als nur der Wert der Variablen (Länge) n. Wichtig ist, wo sich im Vektor x das maximale Element befindet. Wir benötigen also Konzepte, die uns mehr sagen lassen, als daß das maximale Element auf jedem Platz stehen kann, — und damit wohl jede Antwort zwischen 0 und n−1 richtig ist.
Die Lösung dieses Problems greift auf so bekannte[1] Begriffe wie Zufallsvariable, Verteilungen und Momente zurück. Wir benötigen aber auch neue Werkzeuge wie „Erzeugende Funktionen“ aller Art. ◇

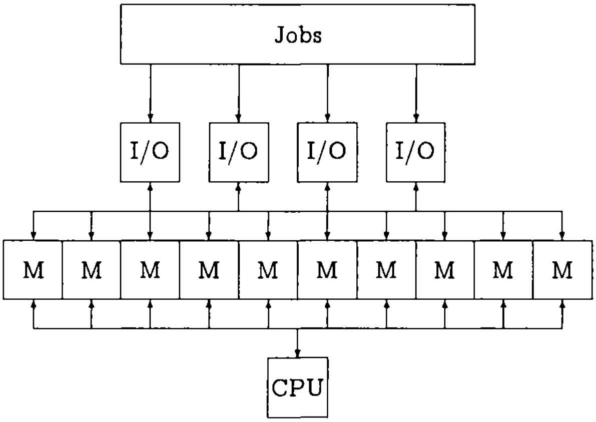
Beispiel 2: *Analyse von Hardware-Architekturen*
Betrachten wir die folgende technische Konfiguration:



[1]Vorausgesetzt wird beim Leser ein Statistikwissen, wie man es üblicherweise in einer statistischen Grundausbildung erwerben kann.

Wie kann man das Verhalten eines solchen Systems (oder ähnlicher Systeme wie z.B. Terminals und Konzentratoren) beschreiben und mit Hilfe der Beschreibung (eines Modells) Hinweise für die Auslegung des Systems (hier Zahl der Steuereinheiten, physikalische Blocklänge usw.) erhalten?
Zur Lösung dieser Problemlage greift man auf „Dinge“ zurück wie Warteschlangentheorie, Exponentialverteilung, Poissonverteilung, kurz: auf Statistik. Schon recht einfache Architekturen erfordern bereits größere Anstrengungen, wie die Untersuchung eines aus einem Kanal und m I/O Einheiten bestehenden Systems durch Wedekind [52] zeigt. ◇

Beispiel 3: *Analyse von Multiprogrammingsystemen*
Mit dem Übergang zur Multiprogrammingnutzung von Computersystemen rückten Probleme der Resourcenverteilung (CPU, Kanäle usw.) und der Jobplanung in den Vordergrund. Man versuchte bald, diese Probleme mit Hilfe wahrscheinlichkeitstheoretischer Modelle zu lösen. Dies lag nahe, da die Anforderungen an Resourcen durch die einzelnen Jobs vorher nicht bekannt sind. Als erstes benötigt man ein Modell des Multiprogrammingbetriebes. Gaver [17] schlug Modelle der nachstehenden Form vor:



Es gibt also ein unerschöpfliches Reservoir von Jobs. Jobs werden in Segmente zerlegt. Ein Segment benötigt einen Platz (M) im Arbeitsspeicher, I/O-Einheiten und CPU. Wenn man sich überlegt, daß alle diese Anforderungen mit stochastischen Größen verbunden sein können (Platzbedarf, I/O-Zeit, CPU-Zeit), wird einsichtig, daß schon sehr einfache Modelle größeren Aufwand beim Einsatz des auch hierbei erfolgversprechenden Instrumentariums aus der Warteschlangentheorie erfordert. Einen ersten Eindruck vermitteln die Arbeiten von Hannsmann [21] und Diruf [14]. ◇

Beispiel 4: *Zuverlässigkeit eines Computers*
Ein Computer besteht aus vielen Einzelteilen, die in einer Fülle von Verknüpfungen miteinander zusammenarbeiten. Leider gilt für alle diese Teile, daß sie nicht 100 %-ig zuverlässig sind. Wie kann man

a) die Zuverlässigkeit einer einzelnen Komponente beschreiben,

b) aus der Zuverlässigkeit der einzelnen Komponenten die Zuverlässigkeit des ganzen Computers berechnen oder abschätzen?

Die Statistik hält mit der sogenannten Zuverlässigkeitstheorie die Antwort auf diese Fragen bereit. ◇

Beispiel 5: *Informationsübertragung bei gestörten Kanälen*
Viele Vorgänge in der Informatik lassen sich auf das folgende Schema zurückführen:

Sender → [ENCODE] → [CHANNEL] → [DECODE] → Empfänger

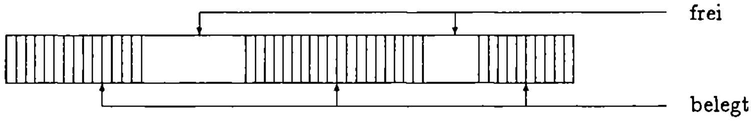
Wie läßt sich ein solches System

a) beschreiben und eventuell so

b) bauen, daß trotz Störungen bei der Übertragung (CHANNEL) möglichst viele Informationen übertragen werden?

Die Antwort auf diese Fragen findet man in der auf der Statistik aufbauenden Informationstheorie. Eine wichtige Rolle spielt dabei das auch sonst in der Statistik nützliche Konzept der Entropie. ◇

Beispiel 6: *Analyse von Magnetplatten*
Dateien werden oft auf einer Magnetplatte gehalten. Dateien entstehen, wachsen, schrumpfen und vergehen. Dies führt dazu, da nicht ständig „aufgeräumt" wird, daß die Platte nach einiger Zeit wie folgt „genutzt" wird:



Es drängen sich schnell Fragen der Art auf: Wie ist der freie Platz verteilt? Wie ist die Verteilung der Größe des freien Platzes? Was kostet mich der freie Platz an

überflüssigen Bewegungen des Schreib-/Lesekopfes und damit an Zeit? Wie wirken sich verschiedene Belegungsstrategien (first fit, best fit usw.) aus?
Die Antwort ist nicht ohne Statistikkenntnisse zu geben. Diskrete Verteilungen und die Theorie der „Rekurrenten Ereignisse" werden benötigt. ◇

Beispiel 7: *Simulation*
Simulation ist ein Werkzeug, das mit dem Aufkommen von Computern eine weite Verbreitung gefunden hat. Insbesondere für die Lösung von Problemen, die durch Unsicherheit in den Daten gekennzeichnet sind, scheint Simulation oft das einzige erfolgversprechende Hilfsmittel zu sein. Was ist dabei zu tun. Schaut man genauer hin, so hat man es mit drei Problemen zu tun.

1. Man muß Stichproben aus beliebig verteilten Grundgesamtheiten ziehen können. Dazu benötigt man ein Verständnis von Zufallszahlengeneratoren (random number generators).

2. Simulationen durchzuführen heißt, ein Experiment zu machen. Experimente müssen aber geplant werden (welche Parameter sollen wie gesetzt werden). Dazu benötigt man ein Verständnis von Versuchsplanung (experimental design).

3. Jeder Versuch kostet, das ist auch bei Simulationsexperimenten nicht anders. Es gibt Techniken, die Kosten klein zu halten. Das Schlagwort, das man verstehen muß, heißt varianzreduzierende Verfahren (variance reduction). ◇

Beispiel 8: *Analyse von Software-Projekt-Kosten*
Es gibt viele Vorschläge, Projektkosten abzuschätzen (vorherzusagen). Einer dieser Vorschläge ist das sogenannte Putman-Modell, das sich am Software Life Cycle orientiert.
Putman [40] unterstellt (aufgrund empirischer Analysen — er wendete also Statistik an) für die zeitliche Verteilung der benötigten „Manpower"

$$Y = K/t_d^2 \cdot t \cdot e^{-t^2/(2t_d^2)}\,.$$

K ist dabei der totale Aufwand, t_d die Entwicklungszeit. Putman schlägt nun vor, durch Aufteilung des Systems in Module — die Zahl der betrachteten Module nimmt beim Durchlaufen des Software Life Cycles zu — und durch Schätzung deren Umfangs in „Source code statements" sich einen Wert für K zu verschaffen. Dabei versucht er, den erwarteten Umfang E und dessen Standardabweichung σ zu schätzen. Er verwendet dazu zwei Formelpaare[2].

$$
\begin{aligned}
\hat{E} &= (a+b)/2 \\
\hat{\sigma} &= |b-a|/6 \\
\hat{E} &= (a+4m+b)/6 \\
\hat{\sigma} &= |b-a|/6\,.
\end{aligned}
$$

Dabei steht a für den kleinsten vermuteten Umfang, b für den größten vermuteten Umfang und m für den wahrscheinlichsten Umfang. Neben den bereits explizit erwähnten Bezügen zur Statistik, ohne Kenntnis der Beta-Verteilung läßt sich der Formelapparat von Putman nicht rechtfertigen und nicht verstehen. ◇

1.2 Informatik für Statistik

Nach den vielen Beispielen, die die Nützlichkeit der Statistik für die Informatik zeigen, wollen wir eilen zu versichern, daß die Informatik für Statistiker von Nutzen ist. Dies werden wir an vielen Stellen des folgendes Textes auf verschiedene Weise demonstrieren.

Zum einen sind nicht nur Tabellen und Abbildungen mit Hilfe eines Computers erstellt worden, sondern der ganze Text wurde auf dem Computer „gesetzt“. Zum anderen werden wir uns das Konzept des Algorithmus zu Nutze machen und viele Verfahren explizit als Algorithmus formulieren. Wir werden dazu eine Notation verwenden, die sich an Vorschläge von Knuth [27] anlehnt.

An einigen Punkten wollen wir aber auch „lauffähige“ Programme angeben. Aber keine Angst! Es wird nicht seitenlang Code auf den Leser herniederrieseln. Wir haben uns die Mahnung von Knuth:

> Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.

zu Herzen genommen. Wir werden daher die entsprechenden Passagen in dem durch Knuth geprägten Literate-Programming-Stil[3] präsentieren.

[2] Ähnliche Formeln finden sich auch bei der Aufwands- und Zeitschätzung im Zusammenhang mit PERT.

[3] Wir verwenden das System *APL2WEB*[7] und Spidery-WEB [43].

Die Beispiele werden wir in den „Sprachen" *APL*[4] und S-Plus[5] formulieren. Wir hoffen, damit auch die üblichen Vorurteile gegen die Sprache (das System) unseres Herzens, *APL*, überwinden zu helfen.
Zum Einlesen in diesen neuen Stil möge das nachstehende kleine Beispiel dienen:
Web 1:

1. α-trimmed mean.

Der Mittelwert $\bar{x}$ ist, wie man weiß, recht sensitiv gegenüber „Ausreißern". Beobachtungen x_ν, die in ihrem absoluten Wert weit über denen der übrigen Beobachtungen liegen, können die naheliegende und vielleicht auch vernünftige Interpretation von $\bar{x}$ als Maß für die Lokation der meisten Beobachtungswerte unmöglich machen.
Will man diese Interpretation aufrecht erhalten, muß man diese „Ausreißer" vor der Berechnung von $\bar{x}$ aus dem Beobachtungsvektor entfernen.
Der sogenannte α-trimmed mean *bietet eine Möglichkeit, dies zu tun. Die Idee ist einfach. Entferne die α-Prozent kleinsten und die α-Prozent größten Beobachtungen aus dem Beobachtungsmaterial und berechne aus den übrigen Daten nach herkömmlicher Weise den Mittelwert.*
Eine APL-Funktion, die dies leistet, soll hier entwickelt werden. Als linkes Argument wird ihr der Wert für α mitgegeben, während das rechte Argument der Vektor der Beobachtungen ist. Als explizites Resultat erhält man den errechneten α-trimmed mean.

```
∇ r←alpha trimmean x    1
```

Eine weitere Definition in Sektion 2.

2. *Der Rumpf der Funktion ist rasch skizziert.*

1. *sortiere x*

2. *entferne α-Prozent Beobachtungen an jedem Ende*

3. *berechne vom Rest den Mittelwert*

Daraus ergibt sich der folgende Funktionsrumpf.

```
∇+ r←alpha trimmean x    1
⟨sortiere x    3⟩
⟨trimme x      4⟩
⟨mittele x     5⟩
```

3. *Die Umsetzung in APL ist dank der Mächtigkeit dieser Sprache rasch geleistet. Schreiben wir zuerst das Sortieren in APL hin,*

```
⟨sortiere x    3⟩ ≡
x←x[⍋x]
```

Dieses Programmstück wird in Sektion 2 verwendet.

[4] Wir verwenden APL∗PLUS von STSC.
[5] S-Plus ist ein Product von StatSci.

4. *und nun das Trimmen,*
⟨trimme x 4⟩ ≡
```
x←n↓(-n←⌊alpha×⍴x)↓x
```
Dieses Programmstück wird in Sektion 2 verwendet.

5. *zum Schluß das eigentliche Mitteln.*
⟨mittele x 5⟩ ≡
```
r←(+/x)÷⍴x
```
Dieses Programmstück wird in Sektion 2 verwendet.

6. *Wir haben jegliche Spitzfindigkeit bezüglich der Struktur von x außer Betracht gelassen. Aufräumen wollen wir aber unsere APL-Umgebung, daher machen wir lokal, was man lokal machen kann.*
⟨Local Variables of *trimmean*(1) 6⟩ ≡
n

In S-Plus, einem speziell auf die Arbeit eines Statitikers ausgerichteten System ist bereits die Lösung der gestellten Aufgabe im Sprachumfang[6] enthalten. Die Funktion `mean` kann durch eine entsprechende Setzung des Parameters `trim` die Arbeit von *trimmean* verrichten.

1.3 Ziele

Was sollen nun mit dieser Abhandlung (und einer entsprechenden Veranstaltung) für Ziele erreicht werden?
Ein wesentlicher Bestandteil statistischen Arbeitens ist die Bildung geeigneter Modelle. Ein Statistiker sucht als erstes nach einem statistischen Modell für die gegebene Problemlage. Diese Suchprozesse werden in dieser Abhandlung an zahlreichen Beispielsituationen demonstriert. Durch nachvollziehendes „learning by doing" wird so dem Leser klar, wie man unter statistischen Gesichtspunkten — und dies sind nach Meinung des Verfassers häufig sehr relevante Gesichtspunkte — mit dem Problem umgeht.
Die statistische Modellierung findet natürlich nicht im leeren Raum statt. Ein Fundus an statistischen Konzepten und Begriffen ist unabweisbare Voraussetzung für erfolgreiches Tun. Es wird hier einmal auf die Statistikkenntnisse zurückgegriffen, die in einer statistischen Grundausbildung vermittelt werden. Es werden aber auch zahlreiche neue Konzepte, Begriffe und Methoden eingeführt.
Die mathematischen Voraussetzungen sollten in den einschlägigen Veranstaltungen des Grundstudiums gelegt worden sein. Als voraussetzbar sieht der Verfasser den Stoff an, wie er beispielsweise durch Wetzel u.a. [53] beschrieben ist. Wenn der

[6] Siehe [8].

Verfasser meint, aus für den Leser unbekannten, unvertrauten Bereichen schöpfen zu müssen, wird ein entsprechender Exkurs eingeschoben.
Im folgenden wird häufig die Rede sein von dem Statistik-Skript oder auch nur dem Skript. Gemeint ist damit das von einem Autorenkollektiv erstellte Skript: „Grundausbildung in Statistik für Wirtschaftswissenschaftler" [4]. Durch dieses Skript ist gleichsam die Basis an Statistikkenntnissen definiert, von der aus diese Abhandlung entwickelt wird. Da das Buch [5] von Günter Bamberg und Franz Baur bei Studenten der Wirtschaftswissenschaften recht bekannt ist, werden wir es ebenso für Verweise nutzen.
Die nur knappe Zeit, die für eine Veranstaltung „Statistik für Informatik" in der Regel zur Verfügung steht, und auch der begrenzte Umfang, den ein Buch nur haben sollte, legen eine exemplarische Darstellung des Gegenstandes nahe. Dies ist auch angesagt angesichts der Tatsache, daß das Ziel auf Anwendung von Statistik auf Informatik ausgerichtet ist. Es wird sich aber herausstellen, daß die theoretische Geschlossenheit nur vermeintlich fehlt. Sie wurde nur nicht als vorrangiges Gliederungskonzept verwandt.
Um die vielleicht etwas abstrakten Ausführungen zu verdeutlichen, sei an einem kleinen Problem — es könnte Gegenstand einer Diplomprüfung sein (war es auch) — einmal demonstriert, welchen Umgang mit Statistik der Verfasser sich nach der Lektüre dieses Buches beim Leser erhofft.

Aufgabe 1:

Zwei Prozesse konkurrieren um eine Leitung (BUS). Diese Leitung wird jeweils im Vielfachen einer Zeiteinheit einem Prozeß zugeordnet. Die Zuordnung geschieht dabei auf folgende Weise. Wenn der Prozeß, dem die Leitung zugeteilt ist, in einem Intervall der Länge von 3 Zeiteinheiten keine Nachricht über die Leitung abgesetzt hat, geht die Kontrolle über die Leitung an den anderen Prozeß über.
Jeder Prozeß sendet mit der Wahrscheinlichkeit p eine Nachricht in einer Zeiteinheit und mit der Wahrscheinlichkeit $q = 1 - p$ keine Nachricht. Das Geschehen in aufeinanderfolgenden Zeiteinheiten sei unabhängig voneinander.
Nehmen Sie an, zur Zeit habe der Prozeß I die Leitung zugeteilt bekommen und er habe in der augenblicklichen Zeiteinheit gerade eine Nachricht abgesetzt. Berechnen Sie die Wahrscheinlichkeit dafür, daß der Prozeß II $k = 0, 1, 2, 3, 4$ Zeiteinheiten mehr als die obligatorischen 3 Zeiteinheiten warten muß, ehe die Kontrolle wieder an ihn übergeht. ♡

Lösung:

1. Beschreibung der Situation.

   Machen wir einen Versuch, das Problem etwas zu formalisieren. Aus der Sicht von Prozeß II betrachtet läßt sich die Situation wie folgt zusammenfassen. Machen wir die Identifikation $A ::=$ 'Prozeß I sendet Nachricht' und $\bar{A} ::=$ 'Prozeß I sendet keine Nachricht', dann ergibt die Beschreibung der Aufgabenstellung, daß $P(A) = p$ und $P(\bar{A}) = 1 - p$ gilt, außerdem ist die Unabhängigkeit des Geschehens in verschiedenen Zeitintervallen gegeben.

Dies erinnert doch sehr an die Beschreibung eines Bernoulli-Prozesses. Zur Erinnerung[7], ein Bernoulli-Prozeß ist bestimmt durch ein

> ... Zufallsexperiment mit zwei möglichen Ausgängen A und $\bar{A}$. Es gelte $P(A) = p$, $P(\bar{A}) = 1 - p$.
>
> Denken wir uns nun dieses Zufallsexperiment mehrmals hintereinander ausgeführt. Dabei sei vorausgesetzt, daß
>
> – die aufeinanderfolgenden Experimente sich nicht gegenseitig beeinflussen, d.h. daß der Ausgang der i-ten Durchführung unabhängig vom Ausgang der j-ten Durchführung ist,
>
> – es bei jeder Durchführung des Experimentes nur die beiden Ausgänge A und $\bar{A}$ gibt,
>
> – die Wahrscheinlichkeit p sich im Verlauf der Experimente nicht ändert.

In der Tat, das Warten auf die Kontrolle über den Bus gleicht der Beobachtung eines Bernoulli-Prozesses.

2. Überprüfung der rund um den Bernoulli-Prozeß bekannten Verteilungen.

   Welche Verteilungen sind uns bisher im Zusammenhang mit einem Bernoulli-Prozeß begegnet?

   i) Binomialverteilung[8]

      Sie entfällt, da sich die Binomialverteilung bei der Betrachtung eines festen n und der Frage nach der Anzahl der beobachteten A unter n Beobachtungen des Bernoulli-Prozesses ergibt.

   ii) Geometrische Verteilung[9]

      Sie entfällt leider auch. Sie ist zwar schon „einschlägig“, aber es wird nur die Wartezeit bis zu einem $\bar{A}$ betrachtet, wir benötigen aber drei $\bar{A}$.

   iii) Pascal-Verteilung (Negative Binomialverteilung) [10]

      Leider ist diese Verteilung auch nicht die gesuchte Verteilung. Es wird zwar die Wartezeit für das $r = 3$-fache Auftreten von $\bar{A}$ gemessen, aber

[7] Siehe auch Statistik-Skript S. 195f.
[8] Statistik-Skript S. 196f., Bamberg/Baur S. 99.
[9] Statistik-Skript S. 198.
[10] Statistik-Skript S. 199.

die $\bar{A}$ müssen nicht — wie in der Aufgabenstellung verlangt — benachbart auftreten.

3. Vollständige Enumeration der Fälle.

Was bleibt zu tun? Wir müssen jeweils zu gegebenen k alle Fälle ermitteln, um daraus $P(X = k)$ zu ermitteln. Wir fassen diese in den nachstehenden Tabellen zusammen. Auf den ersten Blick fehlen auch denkbaren Fälle, wie z.B. für $k = 3$ der Fall $AA\bar{A}\bar{A}\bar{A}\bar{A}$. Dieser ist aber bereits bei $k = 2$ mit abgehandelt worden, da die Teilfolge $AA\bar{A}\bar{A}\bar{A}$ schon zum Übergang der Kontrolle an den Prozeß II führt. Entsprechende Argumentationen gelten in den übrigen (Pseudo-)Fällen.

k	Fall	P(Fall)
0	$\bar{A}\bar{A}\bar{A}$	q^3
$P(X = 0) = q^3$		
1	$A\bar{A}\bar{A}\bar{A}$	pq^3
$P(X = 1) = pq^3$		
2	$AA\bar{A}\bar{A}\bar{A}$	p^2q^3
2	$\bar{A}A\bar{A}\bar{A}\bar{A}$	pq^4
$P(X = 2) = p^2q^3 + pq^4$		
3	$AAA\bar{A}\bar{A}\bar{A}$	p^3q^3
3	$\bar{A}AA\bar{A}\bar{A}\bar{A}$	p^2q^4
3	$A\bar{A}A\bar{A}\bar{A}\bar{A}$	p^2q^4

k	Fall	P(Fall)
3	$\bar{A}\bar{A}A\bar{A}\bar{A}\bar{A}$	pq^5
$P(X = 3) = p^3q^3 + 2p^2q^4 + pq^5$		
4	$AAAA\bar{A}\bar{A}\bar{A}$	p^4q^3
4	$\bar{A}AAA\bar{A}\bar{A}\bar{A}$	p^3q^4
4	$A\bar{A}AA\bar{A}\bar{A}\bar{A}$	p^3q^4
4	$AA\bar{A}A\bar{A}\bar{A}\bar{A}$	p^3q^4
4	$\bar{A}\bar{A}AA\bar{A}\bar{A}\bar{A}$	p^2q^5
4	$\bar{A}A\bar{A}A\bar{A}\bar{A}\bar{A}$	p^2q^5
4	$A\bar{A}\bar{A}A\bar{A}\bar{A}\bar{A}$	p^2q^5
$P(X = 4) = p^4q^3 + 3p^3q^4 + 3p^2q^5$		

Die gefundenen Wahrscheinlichkeiten lassen sich noch vereinfachen.

$$
\begin{array}{lrl}
k = 2 & p^2q^3 + pq^4 & = pq^3(p + q) \\
 & & = pq^3 \\
k = 3 & p^3q^3 + 2p^2q^4 + pq^5 & = pq^3(p^2 + 2pq + q^2) \\
 & & = pq^3(p + q)^2 \\
 & & = pq^3 \\
k = 4 & p^4q^3 + 3p^3q^4 + 3p^2q^5 & = pq^3(p^3 + 3p^2q + 3pq^2) \\
 & & = pq^3(p^3 + 3p^2q + 3pq^2 + q^3) - pq^6 \\
 & & = pq^3(p + q)^3 - pq^6 \\
 & & = pq^3 - pq^6 .
\end{array}
$$

Damit haben wir für die gesuchten Wahrscheinlichkeiten berechnet:

$$\begin{aligned}
P(X=0) &= q^3 \\
P(X=1) &= pq^3 \\
P(X=2) &= pq^3 \\
P(X=3) &= pq^3 \\
P(X=4) &= pq^3 - pq^6 .
\end{aligned}$$

4. Einbettung in eine größere Problematik:

   siehe den Abschnitt über die Theorie der „Rekurrenten Ereignisse" im Kapitel 4.

1.4 Wie sollte man dieses Buch lesen?

Natürlich von der ersten bis zur letzten Seite linear weg, ist eine erste schnelle Antwort. So sehr es den Autor freuen würde, seinen Leser so handeln zu sehen, verlangen will er es nicht von ihm. Dazu quält er sich noch immer zu sehr mit dem Reihenfolgeproblem der Kapitel. Ist es wirklich zweckmäßig, mit diesem Kapitel 1 zu beginnen? Sollte nicht Kapitel 5 weiter nach vorne rücken? Oder vielleicht doch besser ganz an das Ende?
Die hier getroffene Anordnung ist nur eine von mehreren möglichen. In seinen Vorlesungen hat der Autor bereits andere versucht – von denen er damals sehr überzeugt war. Vermutlich wird er bei künftigen Veranstaltungen von der hier gefundenen Reihenfolge der Kapitel abweichen.
Also sollte sich der Leser nicht von der getroffenen Reihenfolge bevormundet fühlen. Will er mit Kapitel X beginnen, bitte sehr. Die einzelnen Kapitel sind jeweils auf einen Gegenstand, eine Sicht o.ä. zentriert und daher in gewisser Weise in sich selbst ruhend. Natürlich gibt es Bezüge zu anderen Kapiteln. Diese sind hoffentlich deutlich genug hervorgehoben.
Nun gibt es Leser, die wollen mehr die „praktische"[11] Seite unseres Gegenstandes: Statistik für Informatik sehen (lesen), während andere ihren „theoretischen" Neigungen nachgehen möchten. Damit der jeweilige Lesertyp etwas einfacher das ihn Interessierende findet – das andere gleichsam ausblenden kann –, werden neben den üblichen Überschriften ausgiebig weitere Strukturierungshilfen[12] angewandt. Sie sind

[11] Die Wörter praktisch und theoretisch wurden in Anführungszeichen gesetzt, da sicherlich ein Praktiker dieses Buch für Theorie und ein Theoretiker es für Praxis halten wird.

[12] So etwas wie Hypertext für den/die arme(n) Mann/Frau. Diese Fußnote ist der einzige Ausrutscher in die augenblicklich übliche Beleidigung des Auges durch die bi-geschlechtliche /-Schreibweise.

in der nachstehenden Tabelle zusammengefaßt:

		Layout	
Anfang	Ende	Schrifttyp	Größe
Algorithmus		*italic*	normal
Aufgabe	$\heartsuit$	roman	normal
Beispiel	$\diamondsuit$	roman	normal
Beweis	qed	roman	small
Definition	$\triangle$	*italic*	normal
Exkurs	$\bowtie$	roman	small
Lösung	$\heartsuit\!\!\heartsuit$	roman	normal
Satz	$\square$	*italic*	normal
WEB		*italic*	small

Der Autor hofft, daß der Leser diese Hilfen auch als solche empfindet. Noch mehr hofft er aber insgeheim doch, daß sich der Leser verhält, wie es die eingangs formulierte Antwort nahelegt.

1.5 Angaben zur Literatur

In der Einleitung wurden die folgenden Monographien und Artikel angesprochen und verwandt:

Autor	Verw.	Titel
Autorenkollektiv	[4]	Grundausbildung in Statistik
Bamberg	[5]	Statistik
Basum	[7]	Making *APL* Readable: A New Direction for Design
Becker	[8]	The New S Language
Diruf	[14]	A Model Structure for the Optimization of Microprocessing Systems
Hannsmann	[21]	Long-Range Planning of Minimum-Cost Systems Based on the Gaver Model
Knuth	[27]	Literate Programming
Putman	[40]	Estimating Software Costs
Wedekind	[52]	Application of Queueing Theory to the Planning of I/O Configurations
Wetzel	[53]	Mathematische Propädeutik

Die Nummern beziehen sich auf die entsprechenden Angaben im Literaturverzeichnis.