Christian-Dietrich Schönwiese

Praktische Statistik

für Meteorologen und Geowissenschaftler

5. vollständig überarbeitete und erweiterte Auflage

Borntraeger

Christian-Dietrich Schönwiese

Praktische Statistik für Meteorologen und Geowissenschaftler

5. vollständig überarbeitete und erweiterte Auflage

Mit 80 Abbildungen, 66 Tabellen im Text und 11 Tabellen im Anhang

Gebr. Borntraeger · Stuttgart · 2013

Schönwiese, C.-D., Praktische Statistik für Meteorologen und Geowissenschaftler

Anschrift des Autors:
Prof. Dr. C.-D. Schönwiese
Institut für Atmosphäre und Umwelt
Goethe-Universität Frankfurt am Main
Postfach 11 19 32
60054 Frankfurt am Main

Gerne nehmen wir Hinweise zum Inhalt und Bemerkungen zu diesem Buch entgegen:
schoenwiese@borntraeger-cramer.de

5. vollständig überarbeitete und erw. Auflage 2013
4. verb. und erw. Auflage 2006
3. verb. und erw. Auflage 2000
2. verb. Auflage 1992
1. Auflage 1985

ISBN ebook (pdf) 978-3-443-01121-5
ISBN 978-3-443-01069-0
Informationen zu diesem Titel: **www.borntraeger-cramer.de/9783443010690**

Verlag: Gebr. Borntraeger Verlagsbuchhandlung
Johannesstraße 3A, 70176 Stuttgart, Germany
mail@borntraeger-cramer.de, www.borntraeger-cramer.de

∞ Gedruckt auf alterungsbeständigem Papier nach ISO 9706-1994
Printed in Germany by Strauss Offsetdruck GmbH, Mörlenbach

Vorwort zur 5. Auflage

Es freut mich sehr, dass das Interesse an diesem Buch nach der 4. Auflage (2006) nun schon zur 5. Auflage geführt hat. So konnte ich wiederum eine intensive Überarbeitung vornehmen. Dabei hat sich das Grundkonzept gegenüber der 3. Auflage allerdings kaum verändert, da nur an wenigen Stellen Erweiterungen vorgenommen worden sind. Das war in der 4. Auflage das neue Kap. 14.9 (Extremwertanalyse); hier ist es das neue Kap. 6.5 (Nachweisgrenze). Dazu kommen aber diverse kleinere Ergänzungen und Hinweise in den bisherigen Kapiteln, hier u.a. zur Bayes'schen Statistik, Bootstrapping und Jacknife. Das Problem der Autokorrelation bzw. Persistenz ist nun ausführlicher als zuvor behandelt. Schließlich habe ich erneut viel „Kleinarbeit" investiert, insbesondere alle Formeln und Beispiele überprüft – was noch zu einigen Korrekturen geführt hat – sowie gelegentlich nochmals Umformulierungen mit dem Ziel der Präzisierung und noch besseren Verständlichkeit vorgenommen.

Zur EDV-Software gibt es keine grundsätzlich neuen Aspekte, außer dass der Nutzer sich laufend an den dafür zuständigen Stellen (i.a. Rechenzentren, z.B. Hochschulrechenzentrum) darüber informieren sollte, welche aktuellen Programmpakete zu möglichst günstigen Konditionen verfügbar sind. Gerade an den Hochschulen besteht aber auch die sehr begrüßenswerte Tendenz, für die jeweils anvisierten Analysemethoden die Programme selbst zu schreiben, was aufgrund der in diesem Buch gegebenen Informationen weitgehend möglich ist. Dies kommt natürlich sehr der Transparenz zugute und erlaubt, auch die bei der Standard-Software oft vernachlässigten aber enorm wichtigen Testverfahren mit einzubeziehen.

Mein Dank geht auch diesmal zunächst an meine Leser und vor allem die Kollegen aus der Meteorologie, Physik und Chemie, die mich mit fachlichen und Literaturhinweisen unterstützt haben. Auch in der Geographie wird mein Buch offenbar intensiv genutzt, wobei in diesem Fall auch das im gleichen Verlag erschienene zweibändige Werk „Statistische Methoden in der Geographie" von G. Bahrenberg et al. als Parallellektüre sehr empfehlenswert ist. Nicht zuletzt geht mein Dank an den Verlag, nunmehr an Herrn Dr. A. Nägele, Herrn Dr. W. Obermiller und das gesamte Verlagsteam, das mich in gewohnt guter Kooperation mit weiteren Hinweisen und formaler Durchsicht unterstützt hat. Ganz besonders dankbar bin ich dabei Frau Dipl. Ing. J. Obermiller für die intensive Durchsicht der Formeln. Das Ergebnis ist somit ein inhaltlich deutlich verbessertes Buch, das in sorgfältiger und ansprechender Druckausführung hoher Qualität bereitsteht.

Bleibt wiederum die Hoffnung, dass die vorliegende kurze und einfache Einführung in die „Praktische Statistik" auch und gerade in dieser neuen 5. Auflage für alle Interessenten und Leser möglichst hilfreich sein möge. Und natürlich sind kritisch-konstruktive Hinweise nach wie vor willkommen.

Frankfurt a.M., im Winter 2012 Christian-Dietrich Schönwiese

Vorwort zur 3. Auflage

Während ich mich bei der 1992 erschienenen 2. Auflage im wesentlichen auf Druckfehlerberichtigungen und einige Präzisierungen beschränkt habe (für entsprechende umfangreiche Hinweise sei an dieser Stelle insbesondere Herrn Dipl.-Met. K. Breitmeier bestens gedankt), liegt nun eine wesentlich überarbeitete, ergänzte und umfassende Neubearbeitung vor, die auch die

neueren Entwicklungen der Statistik und ihre Anwendung in den Geowissenschaften berücksichtigt. Dies hat, trotz einiger Kürzungen bei den grundlegenden Aspekten, den Umfang von zuletzt 231 Seiten nunmehr auf fast 300 Seiten anwachsen lassen. Gänzlich neu sind die Kapitel Clusteranalyse, EOF-, Hauptkomponenten- und Faktorenanalyse, Neuronale Netze sowie einige Unterkapitel (wie z.B. Kanonische Korrelationsanalyse, Zeitreihen-homogenität /-inhomogenität und Trendanalyse).

Das Grundkonzept, eine möglichst einfache (auch ohne vertiefte mathematische Grundkenntnisse lesbare) und übersichtliche erste Einführung in die „Praktische Statistik“ bereitzustellen, habe ich jedoch beibehalten, einschließlich der – extrem kurzen – Beispiele, die nach wie vor keine wissenschaftlichen Analysen darstellen, sondern i.a. per Taschenrechner nachvollziehbar sein und somit möglichst rasch die jeweilige Methodik demonstrieren sollen. Um so wichtiger sind die Hinweise auf weiterführende Literatur und auch einige wenige Beispiele, die über die genannte „Taschenrechner-Statistik“ hinausgehen; denn alle fortgeschrittenen statistischen Methoden sind hier nur angerissen, zum Teil auch nur erwähnt.

Für den Praktiker wird inzwischen eine fast unübersehbare Menge an EDV-Software angeboten, welche die Anwendung so ziemlich aller hier behandelten statistischen Methoden auf dem PC, der Workstation bzw. dem Großrechner erlaubt (jedoch nicht immer mit den notwendigen statistischen Tests). Das Literaturverzeichnis enthält nun auch einige Hinweise zur anspruchsvolleren Software. Ich empfehle aber dringend, diese nicht ohne weiteres als „Black-Box“-Verfahren einzusetzen, sondern sich zuerst genau zu informieren, wie die jeweilige Methode funktioniert. Genau in dieser Richtung möchte das vorliegende Buch hilfreich sein, und dabei auch einiges von den Möglichkeiten und Grenzen der Statistik, sozusagen deren „Philosophie“, vermitteln.

Meine derzeitigen Mitarbeiter Dr. M. Denhard, Dr. J. Grieser, Dr. J. Rapp und Dipl.-Met. T. Staeger haben mich mit Hinweisen und Beispielbeiträgen unterstützt; dafür danke ich sehr. Für das Schreiben auf EDV-Datenträger, einschließlich des Einbaus der Formeln, bin ich Frau P. Sutor sehr dankbar. Weiterhin geht mein Dank wiederum an den Verlag, insbesondere an Herrn Dr. Nägele und sein Team, für die Aufgeschlossenheit, die technischen Hinweise und die so sorgfältige und ansprechende Druckausführung. Nicht zuletzt danke ich meinen Lesern für ihr Interesse; ohne sie wäre diese 3. Auflage nicht möglich gewesen. Kritische Hinweise sind weiterhin willkommen.

Frankfurt a.M., im März 2000 Christian-Dietrich Schönwiese

Aus dem Vorwort zur 1. Auflage

Naturwissenschaftliche Fachgebiete, in denen viele Messdaten anfallen, können auf eine eingehende, korrekte und sinnvolle Anwendung statistischer Methoden nicht verzichten. So werden inzwischen auch im Rahmen des Meteorologie-Studiums Statistik-Vorlesungen angeboten. Meines Erachtens gehört die Statistik in allen geowissenschaftlichen Fächern, und nicht nur dort, zu den notwendigen mathematisch-methodischen Grundlagen.

Das vorliegende Buch – aus meiner Vorlesung „Praktische Statistik“ am Fachbereich Geowissenschaften der Universität Frankfurt a.M. entstanden – wendet sich vorwiegend an diesen Leserkreis, kann aber sicherlich auch in den Biowissenschaften und – mittels der Geographie – bis in die Wirtschaftswissenschaften hinein genutzt werden. Die Kurzbeispiele stammen zwar zum

größten Teil aus meinem eigenen Arbeitsgebiet, der Klimatologie; sie sollen aber lediglich das numerische Vorgehen rasch und übersichtlich demonstrieren und dürfen nicht als abgeschlossene Arbeitsergebnisse angesehen werden. Das Buch ist weiterhin gezielt für den Praktiker geschrieben; d.h. nicht die mathematische Herleitung der Methoden, sondern deren Anwendung in der Praxis steht im Vordergrund. Schließlich ist es so einfach gehalten, dass es als erste Einführung in die Statistik eingesetzt werden kann.

Der vorgegebene Rahmen des Umfangs konnte natürlich keinesfalls auch nur eine annähernd erschöpfende Behandlung der Thematik erlauben. Insbesondere Geophysiker und Meteorologen werden auf weiterführende Literatur nicht verzichten können. Dennoch hoffe ich, dass ich mein besonderes Anliegen verwirklichen konnte: das erste exakte Heranführen an die wesentlichen statistischen Begriffe, Denkweisen und Methoden. Nichts ist ja gefährlicher, als an irgendeiner speziellen und relativ fortgeschrittenen Stelle in die Statistik einzusteigen und auf das Fundament, bewusst oder unbewusst, zu verzichten.

Außer meinen Studenten, die mich mit ihren Fragen auf manche Verständnisprobleme aufmerksam gemacht haben, möchte ich für die Durchsicht des Manuskripts bzw. Hinweise danken: Prof. Dr. H. Berckhemer, Dr. W. Birrong, Prof. Dr. H. Fleer, Dr. V. Kroesch und Dr. J. Malcher. Von Herrn E.O. Bühler stammen die meisten Abbildungen und Frau Dipl.-Met. F. Taghavi-Talab übernahm einen Teil der Reinschrift.

Um einen günstigen Ladenverkaufspreis zu erzielen, habe ich die Anfertigung einer reproduktionsreifen Druckvorlage auf mich genommen. Unter dem Zeitaufwand hierfür sowie für die Arbeit am Manuskript hat meine Familie erheblich gelitten und ich möchte meiner Frau und meinen Kindern für ihr Verständnis herzlich danken.

Schließlich danke ich dem Verlag, insbesondere Herrn Dr. E. Nägele und seinen Mitarbeitern/innen für die freundliche Unterstützung bei der Druckvorbereitung.

Frankfurt a.M., im Mai 1985 Christian-Dietrich Schönwiese

Inhaltsverzeichnis

1 Grundlagen

1.1 Einführung

Über Definition und Zielsetzung der Statistik bestehen in der Öffentlichkeit häufig unklare oder sogar falsche Vorstellungen. So wird beispielsweise eine reine Zusammenstellung von Wirtschaftsdaten oder Wetterrekorden bereits als „Statistik“ bezeichnet. Schlimmer noch sind Fehlinterpretationen vermeintlich statistischer Ergebnisse. Ein immer wieder zitiertes Beispiel dafür ist die „Korrelation“ – gemeint ist dabei der gleichsinnige zeitliche Trend – von Bevölkerungs- und Storchenzahl in Mitteleuropa in der ersten Hälfte des letzten Jahrhunderts. Beide Trends sind zwar abwärtsgerichtet, aber aus ganz unterschiedlichen Gründen. Die rein formale (und unvollständige) Korrelationsrechnung (zur quantitativen Schätzung von Zusammenhängen, näheres in Kap. 11) kommt zwar immer zu einem bestimmten Wert des „Korrelationskoeffizienten“. Doch beschreibt diese Maßzahl hier lediglich den zufällig gleichen Trend. Es handelt sich um eine „Scheinkorrelation“, und nur der böswillige Kritiker wird unterstellen, die Statistik behaupte hier einen ursächlichen Zusammenhang – was sie im Übrigen prinzipiell *nicht* kann.

Sogar der Naturwissenschaftler, insbesondere wenn er an Formulierungen exakter und sicherer Zusammenhänge gewöhnt ist, steht der Statistik nicht selten skeptisch gegenüber. Dies ist auf die Ungewissheit zurückzuführen, die – wenn auch möglicherweise sehr klein – den statistischen Analyseergebnissen stets innewohnt. Diese Unsicherheit kommt beispielsweise in der häufig als „klassisch“ angesehenen Definition der Statistik von A. Wald (1902–1950) zum Ausdruck: „Statistik ist eine Zusammenfassung von Methoden, die uns erlauben, vernünftige optimale Entscheidungen im Fall von Ungewissheiten zu treffen“ (zitiert nach Sachs, 2004).

Tatsächlich ist eine Vielzahl solcher Methoden entwickelt worden, wobei sich rasch die mathematisch eindeutige Ausdrucksweise durchgesetzt hat („mathematische Statistik“). Leider werden statistische Definitionen nicht in derartig einheitlicher Weise verwendet wie beispielsweise in der Physik. Dies, die Literaturfülle – meist im Blickwinkel bestimmter Anwendungsgebiete – und das ungewohnte Umgehen mit Ungewissheiten bzw. Wahrscheinlichkeiten können den Zugang zur Statistik sehr erschweren.

Sollen Definition und Zielsetzung der Statistik von Anfang an möglichst klar und überzeugend sein, so lässt sich dies am besten mit Hilfe des Determinismus-Begriffs erreichen, etwa in der Weise, wie das Bendat und Piersol (1966) getan haben. Dazu spalten wir irgendeinen Vorgang – ob naturwissenschaftlich oder nicht spielt hier keine Rolle – auf in

- Einfluss- (oder Eingangs-) größe(n)
- Wirkungsmechanismus (oder Zusammenhang, Funktion) und
- Wirkungsgröße(n) oder kurz Wirkung

Als primitives aber anschauliches Beispiel kann ein Fahrkartenautomat dienen: Eingeworfenes Geld und gedrückte Wähltaste sind die Einflussgrößen, die ausgegebene Fahrkarte ist die Wirkungsgröße. Der Steuermechanismus des Automaten stellt den Zusammenhang zwischen Einfluss- und Wirkungsgröße(n) her. Das gesamte in sich abgeschlossene Geschehen nennt der Statistiker (und nicht nur dieser) einen Vorgang (Prozess, Mechanismus). Die Tatsache, dass als Folge dieses Vorgangs eine Fahrkarte ausgeworfen wird, heißt Ereignis.

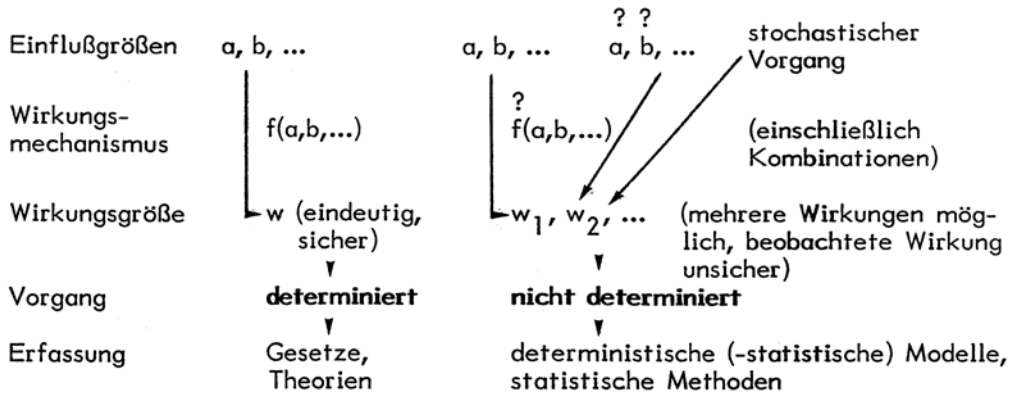

Abbildung 1: Schema eines determinierten und nicht determinierten Vorgangs. Die Fragezeichen bedeuten, dass die Größe(n) bzw. Funktion(en) unbekannt sind.

Und nun definieren wir mit Hilfe der Abb. 1 wie folgt: Ein determinierter (bestimmter, sicherer) Vorgang führt von vollständig, sicher und hinreichend genau bestimmbaren Einflussgrößen über einen ebensolchen Wirkungsmechanismus zu einer ebensolchen Wirkung. Enthält der Wirkungsmechanismus eine Aussage über die Zeit (im mathematischen Sinn eine prognostische Gleichung), so ist die Wirkung sicher und eindeutig prognostizierbar (mit einer Genauigkeit, die durch die Einflussgrößen, insbesondere deren Messgenauigkeit, festgelegt ist; näheres in Kap. 6). Determinierter und „vollständig determinierter" Vorgang werden somit hier synonym definiert.

In allen anderen Fällen handelt es sich um einen nicht determinierten Vorgang. Ein solcher Vorgang lässt sich am einfachsten durch seine Wirkung kennzeichnen: Diese ist nämlich unsicher und stellt nur eine von mehreren Möglichkeiten (Ergebnissen) dar, die trotz gleicher Rand-(Rahmen-) Bedingungen hätten ebenfalls eintreten können. Daher sind nicht determinierte Vorgänge niemals sicher prognostizierbar. Im günstigsten Fall lassen sich Wahrscheinlichkeitsaussagen über die künftige Wirkung abschätzen.

Beispiele:

1. Die Bahndaten eines Satelliten, d.h. dessen Bewegung in Raum und Zeit, beruhen auf einem determinierten Vorgang. Einflussgrößen sind die Massen der Erde und des Satelliten sowie dessen Erdabstand und Umlaufgeschwindigkeit. Wirkungsmechanismen sind das Gravitations- und Zentrifugalgesetz. Wirkungsgröße ist die Bahnkurve, die sich sicher und exakt berechnen und somit auch prognostizieren lässt.
2. Auch das Auftreten einer Sonnenfinsternis ist ein determinierter Vorgang und auf Grund der Keplerschen Gesetze der Himmelskörperbewegungen exakt prognostizierbar.

3. Die Zugbahn einer Gewitterwolke (Cumulonimbus) ist von den atmosphärischen Strömungsgegebenheiten und den Vorgängen in der Wolke selbst abhängig. Da der Wirkungsmechanismus nicht vollständig erfassbar ist, lassen sich Position und Vertikalerstreckung dieser Wolke zu einer späteren Zeit nicht sicher prognostizieren. Es handelt sich daher um einen nicht determinierten Vorgang. (Auch künftige Erdbeben sowie Bevölkerungszahl und Energiebedarf einer bestimmten Stadt im Jahr 2050 sind offenbar „Wirkungsgrößen" nicht determinierter Vorgänge.)

Es gibt nun unterschiedliche Gründe dafür, dass ein Vorgang in nicht determinierter Weise in Erscheinung tritt; vgl. Abb. 1. So kann es sein, dass die Einflussgrößen und/oder der Wirkungsmechanismus nur zum Teil bekannt sind. Die Frage, ob ein Vorgang determiniert auftritt oder nicht, ist somit auch eine Frage der wissenschaftlichen Erkenntnis und Evolution; das heißt, durch die Entdeckung von Gesetzmäßigkeiten können aus zuvor nicht determinierten nun determinierte Vorgänge werden. Es kann sich aber auch um einen Vorgang handeln, der prinzipiell rein zufallsgesteuert abläuft, so dass sich Determinismen niemals werden aufdecken lassen. Schließlich ist es möglich, dass es sich um einen komplexen (zusammengesetzten, vielschichtigen) Vorgang handelt, der determinierte und nicht determinierte Anteile enthält (sog. nicht vollständig determinierter Vorgang). Dies ist in den Geowissenschaften besonders häufig. Nicht selten sind in der Praxis auch Genauigkeit oder Größenordnung der Betrachtung maßgebend dafür, ob ein Vorgang als determiniert angesehen werden kann oder nicht.

Beispiel:

4. Die eine Masse m bestimmter spezifischer Wärmekapazität c beeinflussende Wärmebilanz dQ (Einflussgrößen) führt über das physikalische Gesetz dT = dQ/cm zu einer eindeutigen und sicher prognostizierbaren Temperaturänderung dT. Dies setzt jedoch die Prognostizierbarkeit von dQ sowie makroskopische Betrachtung voraus. Lässt man unkontrollierte Wärmeleitungseffekte zu oder geht man gar zur molekularen Betrachtung über, so wird aus dem zunächst determinierten Vorgang offensichtlich ein nicht determinierter Vorgang.

Dem rein zufallsgesteuerten Vorgang ist in der statistischen Theorie seit jeher besondere Aufmerksamkeit gewidmet worden. Leider hat dies aber nicht zu einheitlich anerkannten Definitionen von Zufall und Zufallsartigkeit geführt. Viele Autoren bezeichnen schlicht alle nicht determinierten Vorgänge als zufällig. Andere verweisen darauf, dass sich selbst reine Zufallsvorgänge in durchaus determinierte Anteile auflösen lassen und sich nur wegen des vielfältigen und insgesamt unüberschaubaren Zusammenwirkens eine zufallsartige Wirkung einstellt. Hier soll, ohne Anspruch auf eine endgültige Einsicht, folgende Festlegung gelten: Ein elementarer zufälliger Vorgang besitzt (bei gleichen Randbedingungen) stets mehrere gleich wahrscheinliche Wirkungen. Bei zufallsartigen Vorgängen kann diese Wahrscheinlichkeit variieren.

Beispiele:

5. Die mit einem regelmäßigen nicht manipulierten Würfel gewürfelte „Augenzahl" (sechs mögliche Wirkungen bzw. Variationsmöglichkeiten) stellt in einem bestimmten Fall (Ereignis) offenbar eine von mehreren gleich wahrscheinlichen Möglichkeiten (Ereignissen) dar. Es handelt sich somit um einen elementaren zufälligen Vorgang.

6. Die genaue Minimumtemperatur der kommenden Nacht an einem bestimmten Ort (Bodennähe) ist nicht sicher vorhersagbar. Es gibt somit mehrere Möglichkeiten, die eintreten können. Wählt man ein hinreichend großes Temperaturintervall, in dem dieser Wert mit großer Wahrscheinlichkeit liegen wird, so sind die einzelnen Temperaturwerte dieses Intervalls in ihrem möglichen Auftreten keineswegs gleich wahrscheinlich. Es handelt sich somit um einen zufallsartigen Vorgang.

Die Stochastik soll hier als die Theorie zufälliger Vorgänge definiert und daher als Teilbereich der Statistik aufgefasst werden, wobei die Statistik auch die Theorie (und Empirik) zufallsartiger Vorgänge mit einschließt (während viele Autoren Stochastik und Statistik synonym definieren). Es kommt im Übrigen häufig vor, dass ein zufallsartiger Vorgang aus der Überlagerung eines zufälligen (stochastischen) Vorgangs mit einem determinierten Vorgang hervorgeht (nicht vollständig determinierter Vorgang). Es lässt sich dann ein stochastischer Anteil zumindest theoretisch abspalten, beispielsweise die Belastung einer physikalischen Messung mit „zufälligen Fehlern“ (näheres in Kap. 6 „Fehlerrechnung“).

Obwohl der Begriff der Wahrscheinlichkeit erst späterer Stelle (Kap. 1.7) eingehend definiert werden soll, können einfache Festlegungen schon jetzt getroffen werden: Wir sprechen von einem sicheren Ereignis und ordnen ihm die „Ereigniswahrscheinlichkeit“ eins zu, in Symbolik.

$$p(E) = 1 \mathrel{\widehat{=}} 100\% \tag{1-1}$$

(engl. Wahrscheinlichkeit → probability, Ereignis → event), wenn andere Ereignisse nicht auftreten können. In formaler Analogie ist dann ein Ereignis, das nie eintritt, das unmögliche Ereignis mit der Wahrscheinlichkeit

$$p(E) = 0 \mathrel{\widehat{=}} 0\,\% \tag{1-2}$$

Unmögliche Ereignisse werden in der Praxis kaum zur Debatte stehen. Aber es ist kennzeichnend für die Statistik, dass sie sich mit unsicheren Ereignissen beschäftigt (vgl. S. 1, Definition nach Wald). Somit gilt für die statistische Ereigniswahrscheinlichkeit

$$0 < p(E) < 1 \tag{1-3}$$

Nach diesen Definitionen muss für die stochastische Wahrscheinlichkeit eine engere Festlegung gelten. Um dies zu erkennen, folgen wir einem Gedankenexperiment von Polya (1963): Es herrsche leichter Regen, bei dem in langsamer Folge nur ab und zu ein Tropfen fällt. Wir betrachten zwei vor uns liegende Steine und fragen: „Auf welchen Stein fällt der nächste Tropfen?“ – Auch ohne Kenntnis der stochastischen Theorie ist leicht einzusehen, dass diese individuelle Frage nicht eindeutig und sicher beantwortet werden kann. Betrachtet man aber eine längere Zeitspanne, in der viele Tropfen fallen, so erscheint die Aussage wahrscheinlich, dass auf jeden der beiden Steine ungefähr gleich viele Tropfen fallen werden. Man spricht in diesem zweiten Fall von einer Massenerscheinung (kollektiver Vorgang), im ersten Fall von einer Einzelerscheinung (individueller Vorgang).

Es lässt sich nun empirisch wie theoretisch zeigen, dass bei stochastischen Vorgängen stets

$$\lim_{n\to\infty} p(E) = c = \text{konstant} \tag{1-4}$$

gilt; d.h. mit steigender Anzahl n beobachteter Ereignisse strebt die Ereigniswahrscheinlichkeit $p(E)$ der Massenerscheinung einem konstanten Festwert c zu. Die folgende Tabelle (Tab. 1) nach Hengst (1967) verdeutlicht dies an Hand des ganz ähnlich gearteten Beispiels von Münzenwürfen, die offensichtlich nur zwei Variationsmöglichkeiten zulassen: K = Kopf, Z = Zahl. Bei gleicher Eintrittswahrscheinlichkeit dieser beiden Möglichkeiten ist im Grenzfall ($n = \infty$) die beobachte Häufigkeit von K und Z identisch. Ist die Gesamthäufigkeit gleich eins, so muss folglich die relative Häufigkeit

$$H_r(K) = H_r(Z) = 0.5 \quad \text{für } n \to \infty \tag{1-5}$$

sein. Es ist ebenfalls offensichtlich, dass dieser Wert auf empirischem Weg nie exakt (d. h. nie mit beliebiger Genauigkeit) gefunden werden kann, weil stets nur endlich viele Versuche möglich sind. (Zur Schreibweise von Dezimalbrüchen siehe Tab. 4, S. 20)

Tabelle 1: Absolute Häufigkeit H und relative Häufigkeit H_r von „Kopf" (K) bei n Münzenwürfen; nach den angegebenen Autoren (zitiert nach Hengst, 1967)

Autor	n	$H(K)$	H_r (K)
Buffon	4 040	2 048	0.5080
Pearson	12 000	6 019	0.5016
Pearson	24 000	12 012	0.5005

Schon hier kommt die enge Verbindung zwischen relativer Häufigkeit und Ereigniswahrscheinlichkeit zum Ausdruck (näheres in Kap. 1.7). Beide Größen haben in Gleichung (1-5) bzw. Tabelle 1 offenbar den gleichen Zahlenwert (bei empirischer Bestimmung allerdings nur annähernd). Bei nicht stochastischen Vorgängen darf die Bedingung (1-4) jedoch nicht a priori vorausgesetzt werden, sondern ist von Fall zu Fall zu prüfen. Ist sie erfüllt, so spricht man auch von stationären Vorgängen, da stochastische Vorgänge stets auch stationär sind, während zufallsartige Vorgänge auch nicht-stationär sein können (näheres in Kap. 2.6 und 14.1).

Wird schließlich berücksichtigt, dass die Statistik eine Methodik anbietet, die in vielen wissenschaftlichen Disziplinen potentiell anwendbar und somit allgemein formulierbar ist, so lässt sich die Statistik kurz wie folgt definieren:

Die Statistik ist die methodische Wissenschaft zur Erfassung zufälliger und zufallsartiger Massenerscheinungen.

Der Begriff "Erfassung" beinhaltet nach Sachs (2004) „Beschreiben, Schätzen und Entscheiden"; man könnte auch in Beschreibung (deskriptive Statistik), Analyse und Prognose gliedern. Manchmal wird von „deduktiver Statistik" gesprochen, was die Herleitung statistischer Methoden (theoretische oder mathematische Statistik) ansprechen soll. Im vorliegenden Buch steht dagegen die Methodik selbst (ohne Herleitung) und deren praktische Anwendung im Mittelpunkt: *praktische Statistik*

Dabei lassen sich unterscheiden:

- Deskriptive Methoden (Stichprobenbeschreibung) → Kap. 2 und 3
- Verteilungstheorie (von Grundgesamtheiten = Populationen) → Kap. 4

- Schätzverfahren (von Ereigniswahrscheinlichkeiten, im weiteren Sinn) → Kap. 5–7
- Testverfahren (Entscheidung über Hypothesen) → Kap. 8 und 9
- Ähnlichkeitsanalyse (Clusteranalyse) → Kap. 10
- Analyse von Zusammenhängen (Korrelation und Regression u.ä.) → Kap. 11–13
- Spezielle Methoden der Zeitreihenanalyse → Kap. 14

Typische Kennzeichen statistischer Arbeitsweise sind allgemein:

- Zahlen und Skalen als Ausgangsmaterial → numerisches Vorgehen.
- Betrachtung zusammenfassender Größen und Funktionen → massenhaftes Vorgehen.
- Bezug auf definitive Wahrscheinlichkeit (im Gegensatz zu reinen Empirik) → probabilistisches Vorgehen.
- Nach Möglichkeit Treffen von Entscheidungen (ggf. auch Prognosen) → definitives Vorgehen.

Am Ende dieser Einführung darf nicht unerwähnt bleiben, dass die Statistik nicht die einzige Methode zur Behandlung nicht determinierter Vorgänge ist. In manchen Fällen (aber nur in diesen), in denen ein komplexer Vorgang einen dominierenden determinierten Anteil enthält, kann versucht werden, den Gesamtvorgang durch ein deterministisches Modell zu simulieren und zu approximieren (z.B. numerische Zirkulationsmodelle der Wettervorhersage, engl. general circulation model GCM). Nicht selten wird zur Verbesserung solcher Modelle auf zusätzliche statistische Beziehungen zurückgegriffen, wodurch ein gemischt deterministisch-statistisches Modell entsteht. Auch die sog. „Model Output Statistics“ (MOS), bei der determinierte Modellergebnisse mittels statistischer Beziehungen mit bestimmten interessierenden Größen – häufig Vorhersagegrößen – verknüpft werden, ist hier einzuordnen.

Prinzipiell müssen derartige Modelle mit Hilfe deskriptiver Statistiken auf ihre Verlässlichkeit hin geprüft (verifiziert) werden. Zudem ist auch in der reinen Statistik der Modell-Begriff üblich: So kann beispielsweise eine theoretische Häufigkeitsverteilung, die nach statistischen Methoden einer empirischen angepasst ist, als statistisches (ggf. auch stochastisches) Modell dienen, um daraus Wahrscheinlichkeiten künftiger Ereignisse herzuleiten (näheres in Kap. 4 und 5).

Es bleibt festzustellen, dass die Statistik im Gegensatz zum deterministischen Modell stets von den beobachteten Wirkungen (Ereignissen) ausgeht. Diese Vorgehensweise kann als real bezeichnet werden. Allerdings bleiben die möglichen Ursachen dabei zunächst außer Betracht. Dagegen geht ein deterministisches Modell von den Ursachen aus, ist somit kausal, wird die Wirkung der modellierten Vorgänge aber immer nur annähernd angegeben können. Es ist selbstverständlich, dass sich in einer modernen geowissenschaftlichen Arbeitsweise (und nicht nur dort) deterministische Betrachtungen (ggf. einschließlich Modellen) und Statistik sinnvoll ergänzen müssen.

1.2 Grundbegriffe

Nach der Definition der Statistik und Erläuterung ihrer Zielsetzung müssen nun noch einige weitere Grundbegriffe eingeführt werden. In der Tabelle 2 sind zur Veranschaulichung dieser Grundbegriffe zwei Beispiele aufgelistet: ein einfaches stochastisches (Würfelspiel), bei dem

diskrete Zahlenwerte auftreten, und ein komplizierteres nicht stochastisches (Temperaturmessung) mit kontinuierlicher Werteskala (stetige Variable). Es muss dabei nicht besonders betont werden, dass die nun folgenden Begriffe in der Statistik eine andere Bedeutung haben können als in anderen wissenschaftlichen Disziplinen (z.B. in der Mathematik oder Chemie).

Wir betrachten zunächst einen Vorgang (Prozess, Mechanismus), der in irgendeiner Weise der statistischen Betrachtung zugänglich ist. Es kann sich dabei um einzelne Gegenstände (z.B. Münze, Würfel), Systeme (Kombinationen) von Gegenständen (z.B. Roulette, Kartenspiel), Größen im physikalischen Sinn (z.B. Lufttemperatur, erdmagnetische Feldstärke), Indizes, Parameter bzw. Variable im mathematischen Sinn (skalar und vektoriell) handeln. Im Folgenden soll der Begriff „Variable“ bevorzugt werden.

Die notwendige Voraussetzung für die statistische Betrachtung von Variablen ist deren quantitativer Bezug, der ja schon im Namen „Variable“ zum Ausdruck kommt. So lässt sich ein Würfel (vgl. Tabelle 2) nur dann statistisch (stochastisch) untersuchen, wenn er auf seinen sechs Seiten auch Zahlenangaben enthält. Ein anderes Beispiel für diskrete (vgl. auch Kap. 1.3) Variable sind die Bevölkerungszahlen von Städten, während Lufttemperatur und erdmagnetische Feldstärke in kontinuierlicher Weise Zahlenwerten annehmen können (stetige Variable). In der Praxis hat man es aber dann doch, wegen der Begrenzung der Messgenauigkeit (näheres in Kap. 6), auch in solchen Fällen mit diskreten Zahlenwerten (genauer: Zahlenwertintervallen) zu tun.

Wichtig ist bei diesen Definitionen, dass es sich beim quantitativen Bezug der Variablen zunächst um potentielle Zahlenwerte handelt; d.h. der betreffende Vorgang ist noch nicht in Gang gekommen (z.B. die oben genannten Bevölkerungszahlen sind noch nicht bestimmt). Es werden somit vorweg das mögliche Zahlenintervall und die zugehörige Skala (vgl. Kap. 1.3) festgelegt. In besonderen Fällen kann es sich um Elemente handeln, denen in phänomenologischer Weise nur die Merkmale „ja“ (Auftreten $\rightarrow$ 1) und „nein“ (Nicht-Auftreten $\rightarrow$ 0) zugeordnet werden können (z.B. Gewitter oder Erdbeben ohne Berücksichtigung einer Stärke-Skala). Diese potentiellen Zahlenwerte einer Variablen heißen Merkmale.

Kommt nun ein (zufälliger oder zufallsartiger) Vorgang (vgl. Kap. 1.1) in Gang, dann treten Wirkungen auf, die in der Statistik Ereignisse genannt werden. Die dabei in Erscheinung tretenden Zahlenwerte heißen Merkmalswerte (Merkmalsausprägungen) oder Daten. Dabei ist darauf zu achten, dass die Ereignisse bzw. Merkmalswerte einem Vorgang mit konstanten Randbedingungen (Rahmenbedingungen) entstammen. Beispielsweise ist es wenig sinnvoll, bei einer Messreihe das Messgerät zu wechseln, es sei denn, es soll durch überlappende Messungen ein systematischer Messfehler aufgedeckt werden (näheres in Kap. 6).

Offensichtlich ist die Anzahl der Merkmale und Merkmalswerte im Allgemeinen unterschiedlich. (In Tabelle 2 ist beim Würfelspiel die Anzahl der Merkmale N=6 und die Anzahl der numerisch aufgelisteten Merkmalswerte n=10; bei der Temperaturmessung ist dort entsprechend n=6 und bei beliebig großer Messgenauigkeit theoretisch N=∞). In Erinnerung an die im Kap. 1.1 gegebenen Definitionen ist festzuhalten, dass das Auftreten numerisch unterschiedlicher Merkmalswerte für nicht determinierte Vorgänge typisch ist. Ein determinierter Vorgang muss dagegen bei gleichen Randbedingungen (im Rahmen der Messgenauigkeit) stets zum gleichen „Merkmalswert“ führen.

Von ganz wesentlicher Bedeutung ist nun weiterhin die Tatsache, dass Merkmalswerte in typischer Häufigkeit auftreten können. Die Feststellung dieser Tatsache geschieht in der Weise, dass die in Frage kommenden Merkmale aufgelistet und die aufgetretenen Merkmalswerte diesen zu-

geordnet werden. Dann lässt sich leicht übersehen, wie oft die einzelnen Merkmale durch die entsprechenden Merkmalswerte im Rahmen der Ereignisse realisiert worden sind. Mit anderen Worten: Die Häufigkeit ist die Anzahl numerisch gleicher Zahlenwerte. Bei der Zuordnung zu den Merkmalen spricht man von einer Häufigkeitsverteilung. (Diese lautet z.B. in Tab. 2, bei „Würfelspiel“: $A_1 = 1 \rightarrow H_1 = 0$; $A_2 = 2 \rightarrow H_2 = 3$; $A_3 = 3 \rightarrow H_3 = 1$; $A_4 = 4 \rightarrow H_4 = 2$; $A_5 = 5 \rightarrow H_5 = 2$; $A_6 = 6 \rightarrow H_6 = 2$; dabei sind A_j die Merkmale und H_i die Häufigkeiten. Im Kap. 1.1 ist bereits die fundamentale Bedeutung der Häufigkeitsverteilung für statistische Untersuchungen erwähnt worden (und zwar wegen ihres Bezugs zur Wahrscheinlichkeit; näheres in Kap. 1.6 und 1.7).

Begriff	**Symbol**	**Bsp. 1: Würfelspiel**	**Bsp. 2: Temperaturmessung**
Variable (Größen)	a, b, ...	Würfel	Lufttemperatur (z.B. am festen Ort zu variablen Zeiten)
Merkmale	A_j, B_j, ... (j = 1,2,...,J)	„Augenangaben“ 1,2,3,4,5,6	Skala, z.B. in °C (äußere Grenzen klimatologisch festgelegt, z.B. -30 °C $\leftrightarrow$ + 40°C)
Ereignisse	E_i	Tatsache, dass bei jedem Würfelvorgang eine bestimmte Augenzahl auftritt	Tatsache, dass bei jeder Messung ein bestimmter Temperaturwert auftritt
Merkmalswerte (Daten)	a_i, b_i, ... (i=1,2,...,n)	z.B. 2,6,3,2,5,6,4,4,2,5, ... (a_i, $n = 10$)	z.B. 15.1, 16.7, 14.3, 17.5, 16.8, 15.4, ... °C (a_i, $n = 6$)

Tabelle 2: Erläuterung einiger statistischer Grundbegriffe anhand von zwei Beispielen (Würfelspiel und Temperaturmessung; Definitionen der Grundbegriffe siehe Text).

Die Zusammenstellung von statistisch zu untersuchenden Merkmalswerten (Daten) nennt man ein Kollektiv (auch statistische Masse), häufig in Form eines Protokolls (Urliste), das in übersichtlicher Form die Daten, Häufigkeiten, Maßeinheiten und ggf. statistischen Analyse-Ergebnisse enthält (Beispiele folgen in Kap. 1.6). Wichtig ist nun die Frage, ob ein solches Kollektiv alle möglichen Ereignisse eines definierten Vorgangs enthält oder nicht. Im Allgemeinen wird das nicht der Fall sein; dann stellt das betreffende Kollektiv eine Stichprobe SP dar. Andernfalls sprechen wir von der Grundgesamtheit GG oder Population. Die Beispiele in Tab. 3 zeigen, dass Grundgesamtheiten einen endlichen (finiten) oder unendlichen (infiniten) Umfang aufweisen können. Im ersten Fall sind sie zumindest prinzipiell vollständig zugänglich, im zweiten Fall prinzipiell nicht.

Vorgang	Stichprobe (SP)	Grundgesamtheit (GG)
Würfeln	Gewürfelte Zahlenwert-Reihe a_i (z.B. Zahlenwerte aus Tab. 2: 2,6,3,2,5,6,4,4,2,5)	Gewürfelte Zahlen bei unendlich vielen Würfen (GG infinit, somit prinzipiell nicht zugänglich).
Politische Wahl	Nach bestimmten Kriterien teilweise ausgezählte Wählerstimmen in ebenfalls ausgewählten Wahlkreisen.	Vollständig ausgezählte Wählerstimmen aller Wahlkreise (GG finit, somit prinzipiell zugänglich).
Temperaturmessung	Tagesmittelwerte der Temperatur (z.B. Zahlenwerte aus Tab. 2) für einen bestimmten Ort und ein bestimmtes Zeitintervall.	Zeitlich kontinuierlich und unendlich lange gemessene Temperaturwerte an einem bestimmten Ort, genau genommen sogar kontinuierlich bezüglich der Raumkoordinaten.

Tabelle 3: Beispiele für Stichproben und Grundgesamtheiten (= Populationen) bei verschiedenen Vorgängen.

1.3 Zahl, Größe und Skala

Im Kap. 1.1 wurde unter anderem angemerkt, dass das numerische Vorgehen, d.h. die Beschäftigung mit Zahlen und Skalen, für die statistische Arbeitsweise kennzeichnend ist. In manchen Fällen sind die zu untersuchenden Daten reine Zahlen (z.B. Indexwerte wie der Kontinentalitätsindex der Klimatologie, Aktienindex, Erdbebenstärke nach der Richter-Skala). Häufiger treten in den Geowissenschaften aber Größen im physikalischen Sinn auf, die stets als Produkt aus einem Zahlenwert (reine Zahl) und einer Maßeinheit (z.B. $5.7 \cdot 1\text{kg m}^{-3} = 5.7\text{ kg m}^{-3}$; vgl. auch Kap. 6.1) aufzufassen sind. In der Praxis wird man, falls erforderlich, alle Daten in Zahlen gleicher Maßeinheit transformieren und mit diesen Zahlenwerten weiterrechnen.

Nur am Rande soll erwähnt sein, dass die in der Statistik anfallenden Zahlen im allgemeinen rational sind, d.h. sich als Quotient ganzer Zahlen ($z \subseteq \{-m, -m+1, \ldots, -1, 0, 1, \ldots, n\}$) bzw. in Form eines endlichen Dezimalbruchs ausdrücken lassen. (Dagegen umfasst die Indizierung der Merkmalswerte $i = 1, 2, \ldots, n$ nur natürliche Zahlen.) Ebenfalls im mathematischen Sinn lassen sich Größen in Skalare (jeweils durch einen Zahlenwert vollständig gekennzeichnet, z.B. Temperatur) und Vektoren untergliedern. Letztere müssen jeweils durch mindestens zwei Zahlenwerte gekennzeichnet werden, z.B. Wind durch Richtung und Geschwindigkeit. (Die Anzahl dieser Zahlen ist mit der Dimension der Vektoren identisch, näheres in Kap. 3).

Um mit Zahlen sinnvoll arbeiten zu können, müssen sie in einem bestimmten Bezug zueinander stehen; mit anderen Worten, sie müssen eine Skalierung aufweisen. Die einfachste Skalierung ist die Nummernskala, bei der die Daten einfach ohne weiteren quantitativen Bezug durchnumeriert werden (z.B. Lottokugeln). Der statistischen Bearbeitung besser zugänglich sind Daten, die nach einer Rangskala geordnet sind: Der höchste Zahlenwert erscheint zuerst (Rangplatz 1), dann der zweithöchste (Rangplatz 2) und so weiter bis zum Minimum (absteigende Rangfolge; ganz analog dazu kann auch nach einer aufsteigenden Rangfolge geordnet werden, beginnend mit dem Minimum).

Meist werden die Daten jedoch bezüglich einer Maßeinheit oder einem anderen Ordnungsprinzip auf eine lineare Intervallskala (Einheitsskala) bezogen sein; weist diese Skalierung einen absoluten Bezugspunkt (Nullpunkt) auf, so spricht man von einer Rationalskala (Verhältnisskala). In Tab. 4 sind die drei letztgenannten Skalierungen an Hand eines Beispiels erläutert. In manchen Fällen, insbesondere bei Regressionen (Kap. 11), kann es zweckmäßig sein, die in einer Intervall- oder Rationalskala linear geordneten Daten in eine nicht lineare Skala (z.B. logarithmische) zu transformieren. Im Allgemeinen aber sind zunächst Daten nach der Rationalskala zu bevorzugen.

Übrigens wird in diesem Buch (vgl. oben und auch schon Tab. 1 und 2) bei Dezimalbrüchen statt der deutschen Kommaschreibweise die englische Punktschreibweise verwendet, was bei Aufzählungen solcher Zahlenwerte, getrennt durch Kommas, von Vorteil ist.

i	Intervallskala	Rationalskala	Rangskala
1	15.1 °C	288.1 K	6
2	16.7	289.7	4
3	14.3	287.3	7
4	17.3	290.5	1
5	16.8	289.8	3
6	15.4	288.4	5
7	17.1	290.1	2

Hinweis:
Bei der Umrechnung der Temperaturwerte von der Celsius-Skala (°C) in die Kelvinskala (K) wurde die vereinfachte Formel $K = °C + 273$ benützt (genauer Wert: 273.15). In der Rangskala sind die Rangplätze R angegeben.

Tabelle 4: Beispiele einer Temperatur-Messreihe in verschiedenen Skalierungen. (Dabei ist *i* die Nummer der Messung bzw. der Index der Merkmalswerte.)

Recht häufig werden in der Statistik nun auch Zahlentransformationen vorgenommen, zum Teil aus Gründen der Vereinfachung. Ausgehend von reinen Zahlen z_i kann dies zu folgenden Zahlenarten führen:

Differenzzahlen	$d_i = z_i - c$	(1-6)
Verhältniszahlen	$v_i = z_i \,/\, c$	(1-7)
Prozentualzahlen	$p_i = (z_i \,/\, c) \cdot 100\,\%$ mit $\sum(z_i \,/\, c) = 1$	(1-8)
Normalzahlen	$n_i = z_i \,/\, \sum z_i$	(1-9)

Dabei ist c eine Konstante, z.B. der arithmetische Mittelwert (Definition folgt in Kap. 2.2). Aus Normalzahlen erhält man natürlich durch Multiplikation mit 100 % stets prozentuale Zahlen. Die folgende Tabelle bringt ein Beispiel zur Umrechnung in diese verschiedenen Zahlenarten.

z_i	d_i	v_i	n_i	p_i
15.1	−1.0	0.94	0.134	13.4 %
16.7	0.6	1.04	0.148	14.8 %
14.3	−1.8	0.89	0.127	12.7 %
17.3	1.2	1.07	0.153	15.3 %
16.8	0.7	1.04	0.149	14.9 %
15.4	−0.7	0.96	0.137	13.7 %
17.1	1.0	1.06	0.152	15.2 %

mit $\sum z_i = 112.7$ und
$c = 112.7/7 = 16.1$
(arithmetischer Mittelwert; vgl. Kap. 2.2)

Tabelle 5: Beispiel zur Zahlentransformation nach den Formeln (1-6) bis (1-9). Dabei sind die Zahlenwerte z_i aus Tab. 4 entnommen und v_i, n_i sowie p_i gerundet.

Bei Verwendung einer Rangskala (vgl. Beispiel in Tab. 4) und R = Rangplatz ist

$$\mathrm{PR} = \frac{R}{n} \cdot 100\,\% \tag{1-10}$$

(n = Kollektivumfang) der sog. Prozentrangplatz. Er gibt an, wie viel Prozent des betreffenden Kollektivs vor R liegt.

Beispiel:

7. Von 43 in einer bestimmten Region der Erde erfassten Vulkanausbrüchen steht der neueste auf Rangplatz 33 nach dem Kriterium des geschätzten Auswurfs vulkanischer Materie. Dann sind nach diesem Kriterium $\frac{33}{43} \cdot 100\,\% \approx 76.7\%$ als stärker (hinsichtlich ihrer Auswurfmasse) einzustufen.

1.4 Verschachtelung phänomenologischer Größenordnungen

Bei der statistischen Untersuchung nicht determinierter Vorgänge ist es wichtig, dass die Randbedingungen (Rahmenbedingungen) möglichst konstant bleiben; andernfalls treten Interpretationsschwierigkeiten auf. Insbesondere darf der jeweils betrachtete Vorgang nicht durch andere variierende Vorgänge beeinflusst sein. Eine besonders wichtige Rolle spielen diese Voraussetzungen bei der Korrelations- und Regressionsanalyse (Kap. 11) und bei der Zeitreihenanalyse (Kap. 14). Aber auch bei der Suche nach einem geeigneten Verteilungsmodell (hinsichtlich der Häufigkeit, vgl. Kap. 4 und 5) können dementsprechende Schwierigkeiten auftreten.

Man kann versuchen, solchen Schwierigkeiten von vornherein dadurch zu begegnen, dass die jeweils betrachtete räumliche und zeitliche Größenanordnung möglichst genau definiert und auf fachlicher Grundlage diskutiert und entschieden wird, inwieweit eine mehr oder weniger isolierte Betrachtung dieser Größenanordnung sinnvoll ist. Zwar kann die Statistik durchaus Entscheidungshilfen bereit stellen (z.B. Varianzspektren und gefilterte Daten im Rahmen der Zeitreihenanalyse, vgl. Kap. 14), doch muss die Entscheidung selbst wie gesagt auf fachlicher Grundlage fallen.

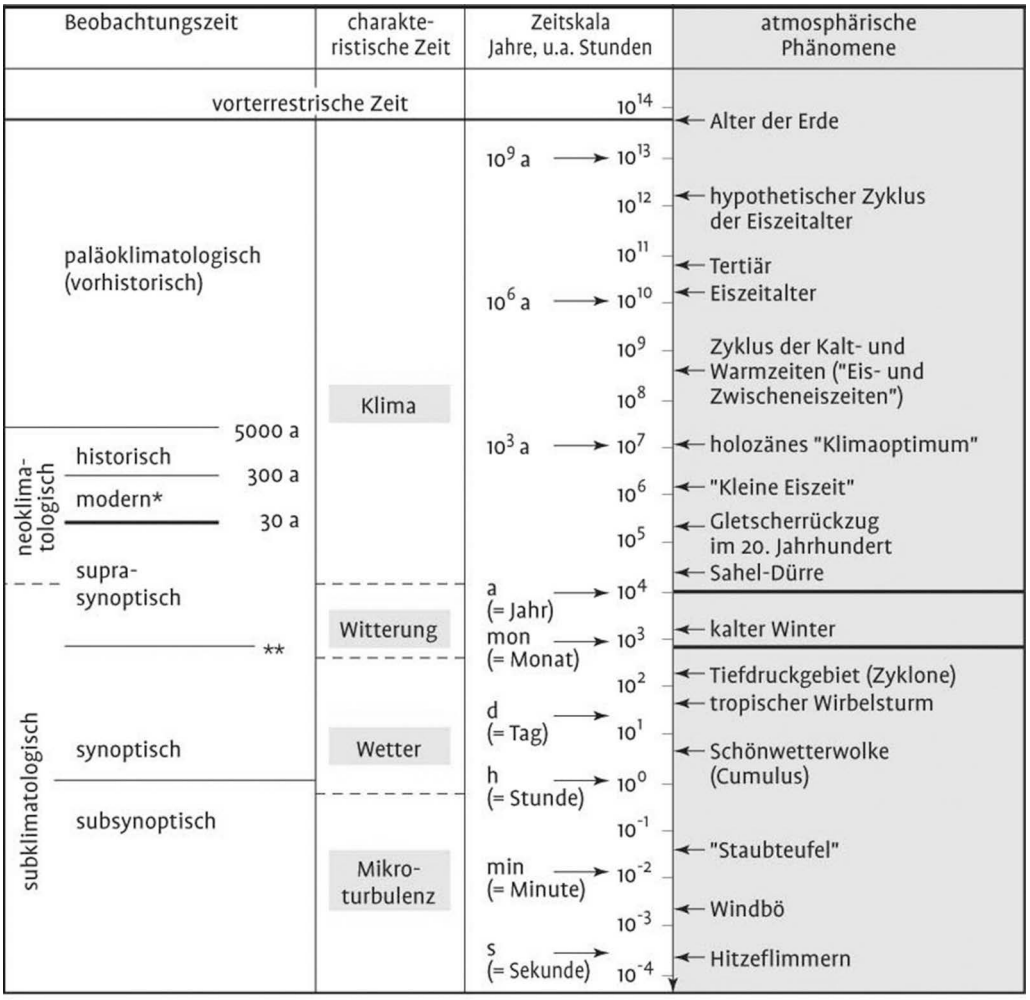

Abbildung 2: Skala zeitlicher Größenanordnungen in der Meteorologie/Klimatologie mit einigen Beispielen atmosphärischer Phänomene (nach Schönwiese, 2008). Dabei ist die charakteristische Zeit die mittlere Lebensdauer solcher Phänomene bzw. die Zykluszeit (vgl. Kap. 14). Die Beobachtungszeit muss gegenüber der charakteristischen Zeit relativ groß sein, um die statistischen Eigenschaften der Phänomene feststellen zu können.

Insbesondere für den meteorologisch bzw. klimatologisch ausgebildeten Anwender statistischer Methoden mag es daher hilfreich sein, sich diese Größenanordnungen vor Augen zu führen und bei der statistischen Analyse zu berücksichtigen. In der Abb. 2 ist daher vom phänomenologischen Standpunkt aus eine Skala zeitlicher Größenanordnungen mit einigen Beispielen atmosphärischer Phänomene wiedergegeben, wobei die gesamte Skala von Sekundenbruchteilen bis Jahrmilliarden abgedeckt ist. Für zeitliche Größenanordnungen bis zu einigen Monaten lassen

sich nach dem „Scale“-Diagramm der Abb. 3 entsprechende räumliche (horizontale) Größenanordnungen zuordnen; d.h. in einem solchen Raum-Zeit-Diagramm findet man die meteorologischen Phänomene vorwiegend in einem mehr oder weniger abgegrenzten diagonalen Bereich.

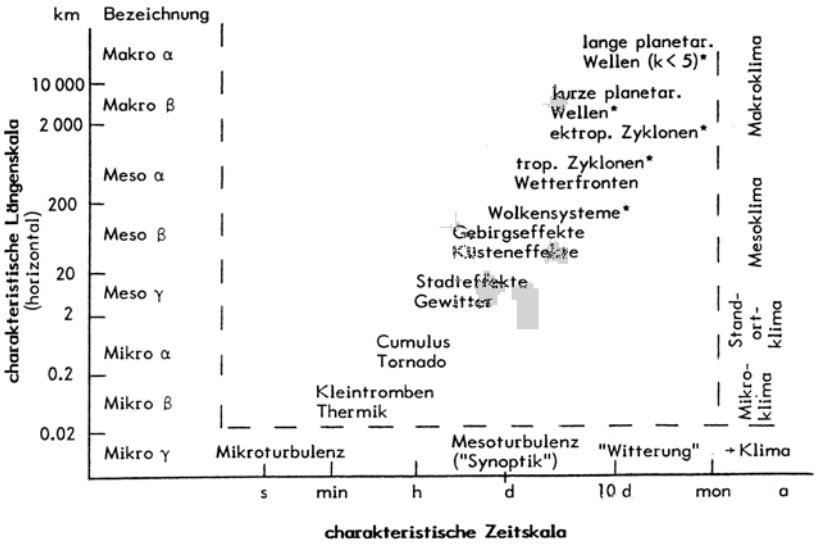

Abbildung 3: Skalenbegriff („Scale“) der Meteorologie, nach dem sich viele atmosphärische Phänomene räumlich-zeitlich zuordnen lassen; in Anlehnung an Fortak (1982) bzw. Orlanski (1975). *Anmerkungen: Die planetarischen Wellen dienen der Kennzeichnung des in Mäanderform um den Globus (i.a. mittlere geographische Breiten und ca. 5 km Höhe betrachtet) verlaufenden Strömungsmusters, wobei k die Wellenzahl ist. Die tropischen Zyklonen sind Hurrikane, Taifune u.ä., während die ektropischen (außertropischen) Zyklonen die wandernden Tiefdruckgebiete/Wetterfrontensysteme der mittleren Breiten sind. Mit Wolkensystemen sind hier die tropischen „cloud clusters“ gemeint.

Sollen nun bestimmte Vorgänge statistisch untersucht werden, so wird man sie im Allgemeinen definierten Phänomenen eines ebenfalls definierten Raum-Zeit-Bereichs zuordnen. Man spricht von skaligen Phänomenen, die jedoch meist mit subskaligen (kleinere Größenanordnung) und supraskaligen (größere Größenanordnung) Phänomenen verknüpft sind. In der deterministischen Modellierung müssen die subskaligen Phänomene parametrisiert, d.h. durch geeignete Rechengrößen bzw. Transformationen in die Berechnung eingebracht werden. Die supraskaligen können ggf. als Randbedingungen (auch variabler Art) berücksichtigt werden.

In der Statistik sind die subskaligen Phänomene meist durch Mittelbildung in den skaligen Merkmalswerten (Daten) enthalten und müssen daher i.a. nur bei der Interpretation der statistischen Ergebnisse berücksichtigt werden. Die supraskaligen Phänomene aber können sich gerade in der Statistik sehr störend auswirken, beispielsweise in der Art, dass sie Nicht-Stationarität (Definition in Kap. 2.6) erzeugen oder allmählich die Randbedingungen verändern.

Wären die zu beurteilenden Phänomene auf die jeweilige Größenordnung beschränkt und von Phänomenen anderer Größenordnungen unbeeinflusst, so würden sowohl bei der deterministischen als auch bei der statistischen Analyse und Modellbildung weit weniger Probleme auftreten. Die räumlich-zeitliche Verschachtelung der phänomenologischen Größenordnungen aber erfordert umsichtige Arbeitsweise und vorsichtige Interpretation, einschließlich geeigneter Repräsentanzüberlegungen (dazu näheres in Kap. 7).

1.5 Zeitreihen (Definition)

In den Geowissenschaften haben die betrachteten Vorgänge meist einen räumlichen und zeitlichen Bezug, der im Einzelnen variabel sein kann. Dann beziehen sich die betreffenden Merkmalswerte (Daten) a_i nicht auf feste Ortskoordinaten x_* , y_* , z_* und eine ebenfalls feste Zeitkoordination t_* , sondern es gilt:

$$a_i(x, y, z, t) \tag{1-11}$$

In der Geographie ist x meist mit der geographischer Länge λ und y mit der geographischen Breite φ identisch; z ist die vertikale Ortskoordinate (Höhe). Eine besondere Bedeutung aber haben, und das nicht nur in den Geowissenschaften, Zeitreihen gewonnen, so dass eine ganze Reihe von statistischen Methoden der speziellen Zeitreihenanalyse entwickelt worden sind. Hier soll zunächst nur ihre Definition erfolgen. Eine Zeitreihe hat die Form

$$a_i(t_i) \quad \text{mit } i = 1, \ldots, n \quad \text{und} \quad t_{i+1} - t_i = \Delta t = \text{ konstant} \tag{1-12}$$

Das heißt, es handelt sich um eine diskrete Datenreihe, deren Werte sich auf äquidistante Zeiten t_i beziehen. Damit können feste Zeitpunkte gemeint sein, aber auch Zeitintervalle (die mit jeweils gleicher Länge aufeinander folgen).

Beispiele:

8. Zeitreihen, die sich auf bestimmte Zeitpunkte beziehen, sind z.B. stündlich gemessene Temperaturwerte, täglich gemessene Komponenten des erdmagnetischen Feldes oder jährlich ermittelte Einwohnerzahlen bestimmter Städte.
9. Zeitreihen, die sich auf Zeitintervalle beziehen, sind z.B. Tageswerte der Verkehrsfrequenz in einer bestimmten Straße (gezählte Fahrzeuge jeweils im Laufe eines Tages), monatlicher Stromverbrauch einer Stadt, Jahressummen des Niederschlages, Jahresmittel der Sonnenflecken- Relativzahlen oder 30-jährige Mittelwerte des Luftdrucks.

In Beispiel 8 handelt es sich somit um Daten, die nur zu bestimmten Terminen vorliegen. Dies kann natürlich auch darin begründet sein, dass eine in Wahrheit stetige Variable (kontinuierlich variierende Größe, z.B. Lufttemperatur) aus technischen Gründen nur zu diesen Terminen gemessen wird. Der in Beispiel 9 gegebene Bezug auf Zeitintervalle kann in unterschiedlicher Weise zustande kommen: durch Akkumulation oder durch Mittelung über die betreffenden Zeitintervalle. Im letzteren Fall beziehen sich die Daten a_i stets auf die Mitten dieser Zeitintervalle Δt_i.

Durch äquidistante zeitliche Mitteilung von Zeitreihen $a_i(t_i)$ oder deren Akkumulation über jeweils gleich große Zeitintervalle entstehen neue Zeitreihen $b_i(z_i)$. Dies gilt aber nicht immer exakt, z.B. nicht für Monatssummen des Niederschlages, da die Monate des Jahres unterschiedlich lang sind. Zeitfunktionen

$$a(t) \tag{1-13}$$

(z.B. Analogregistrierungen des Luftdrucks mittels Barograph) müssen in diskrete Zeitreihen (möglicherweise sehr kleinen Zeitabstandes) umgewandelt werden, um statistisch auswertbar zu

sein. Im Folgenden sollen Zeitreihen, soweit sie in den Beispielen auftauchen, zunächst nicht anders wie sonstige Stichproben-Kollektive behandelt werden. Die speziellen Methoden der Zeitreihenanalysen folgen dann in Kap. 14.

1.6 Häufigkeitsverteilung und Klassenbildung

Nun soll wieder eine beliebige Stichprobe SP betrachtet werden, die bei konstanten Randbedingungen zustande gekommen ist. Wie bereits ausgeführt, ist es für statistisch zu untersuchende nicht determinierte Vorgänge typisch, dass Merkmalswerte (Daten) unterschiedlichen Zahlenwertes und im Allgemeinen auch unterschiedlicher Häufigkeit auftreten (vgl. Kap. 1.2, Tab. 2).

Der Einfachheit halber wollen wir uns zunächst wieder dem stochastischen Würfelbeispiel zuwenden. Um nicht von zu wenigen Daten auszugehen und um zu erreichen, dass alle Merkmale ($A_j = 1, \ldots, 6$) auch in den Merkmalswerten vertreten sind, soll in Erweiterung von Tab. 2 bzw. 3 weiter gewürfelt werden. Es könnte dann das im folgenden Beispiel aufgeführte Kollektiv zustande kommen.

Beispiel:

10. Folgende gewürfelte Zahlen sollen hinsichtlich ihrer Häufigkeit analysiert werden:
$\rightarrow a_i = 2, 6, 3, 2, 5, 6, 4, 4, 2, 5, 4, 5, 3, 1, 6, 2$; somit ist der Stichprobenumfang $n = 16$.

Wenn nun, wie bereits bei diesem einfachen Beispiel, nicht sofort zu überblicken ist, wie sich die Merkmalswerte auf die Merkmale verteilen, d.h. wie die Häufigkeitsverteilung aussieht, so ist es zweckdienlich, diese Verteilung zunächst in einer Strichliste festzuhalten; siehe Tab 6. Es wird somit ein Protokoll (Urliste) begonnen, das in der ersten Spalte die Merkmale A_j enthält; dann wird Datenwert für Datenwert in der Spalte der Strichliste SL bei dem betreffenden Merkmal ein Strich angebracht, bis der Stichprobenumfang n erschöpft ist.

A_j	SL	H_j	KH_j	RH_j	RKH_j	PKH_j
1	|	1	1	0.0625	0.0625	6.25 %
2	||||	4	5	0.2500	0.3125	31.25 %
3	||	2	7	0.1250	0.4375	43.75 %
4	|||	3	10	0.1875	0.6250	62.50 %
5	|||	3	13	0.1875	0.8125	81.25 %
6	|||	3	16	0.1875	1.0000	100.00 %
$\sum$		16		1.0000		

Die Abkürzungen bedeuten:
SL = Strichliste
H = Häufigkeit
KH = kumulative Häufigkeit
RH = relative Häufigkeit
RKH = relative kumulative Häufigkeit
PKH = prozentuale kumulative Häufigkeit

Tabelle 6: Protokoll (Urliste) zu Beispiel 10 (Merkmale A_j, $j = 1, \ldots, 6$) mit Errechnung der verschiedenen Arten von Häufigkeitsverteilungen; vgl. dazu auch Abb. 4.

Die Umsetzung dieser Strichliste in Zahlen ergibt die absoluten Häufigkeiten Hj und in Verteilung auf die Merkmale A_j die absolute Häufigkeitsverteilung

$$HV:\ H_j(A_j) \quad \text{mit} \quad j = 1, \ldots, J \tag{1-14}$$

mit J unterschiedlichen Merkmalen. Als Rechenkontrolle muss $\sum H_j = n$ gelten (in Tab. 6: 16). Für praktische Zwecke (die theoretische Begründung folgt in Kap. 1.7) kann es sinnvoll sein, die kumulativen (Summen-) Häufigkeiten zu bilden; dies führt zur kumulativen Häufigkeitsverteilung

$$KHV:\ KH_j = H_1, H_1 + H_2, \ldots, H_1 + H_2 + \cdots + H_n \tag{1-15}$$

Dabei muss als Rechenkontrolle $KH_{max} = n$ gelten (in Tab. 6 wiederum „16“). Greift man ein bestimmtes Merkmal (z.B. $A_j^* = A_{(j=3)}$ in Tab. 6) heraus, so gibt die zugehörige kumulative Häufigkeit KH_j an, wie viele Merkmalswerte im Bereich $A_j \leq A_j^*$ liegen (z.B. 7 Daten in Tab. 6).

Meist gewinnt man ein übersichtlicheres Bild, wenn man die Häufigkeitsverteilung normiert $\left(\text{d.h. } \sum H_j = 1\right)$. Dies führt zu den relativen Häufigkeiten RH_j bzw. zu relativen Häufigkeitsverteilung

$$RHV: RH_j = \frac{H_j}{\sum_{j=1}^{J} H_j} \quad \text{mit} \quad j = 1, \ldots, J \tag{1-16}$$

Durch Akkumulation lässt sich daraus wieder die kumulative Form, die relative kumulative Häufigkeitsverteilung, errechnen:

$$RKHV:\ RH_1,\ RH_1 + RH_2, \ldots,\ RH_1 + RH_2 + \cdots + RH_j \tag{1-17}$$

Schließlich können die RH_j als auch die RKH_j prozentual ausgedrückt werden, wobei in der Statistik besonders die kumulative Form, nämlich die prozentuale kumulative Häufigkeitsverteilung PKHV (PKH_j, vgl. Tab. 6) Bedeutung erlangt hat (vgl. auch Kap. 2.1 und 2.3). In der Abb. 4 sind für die in Tab. 6 angegebenen Zahlenwerte die absolute, relative, kumulative und prozentuale kumulative Häufigkeitsverteilung graphisch dargestellt. Im Fall der nicht kumulativen Form spricht man von einem Säulendiagramm (Histogramm), im Fall der kumulativen Form von einem Treppendiagramm (bzw. einer Treppenfunktion).

Schon bei diesen Beispielen wird man sich überlegen, inwieweit diese Verteilungen verallgemeinert werden dürfen, d.h. ob sie bereits den entsprechenden Prozess (Grundgesamtheit, Population) widerspiegeln. Da es sich beim Würfelspiel um einen elementaren stochastischen Vorgang handelt, ist auch ohne theoretischen Hintergrund der Schluss nahe liegend, dass die Verallgemeinerung dieses Vorgangs, genauer gesagt die Betrachtung der Grundgesamtheit (mit unendlich vielen Merkmalswerten), zu einer Verteilung führen müsste, bei der sich alle Merkmalswerte gleichmäßig (wegen gleicher Wahrscheinlichkeit ihres Eintretens) auf die Merkmale verteilen (Gleichverteilung, Definition folgt in Kap. 4.2) mit $RH_j = 1/J = 1/6$. In der Abb. 4 sind die sich aus diesen Überlegungen ergebenen theoretischen Verteilungen (mittlere Häufigkeitsfunktionen, horizontale bzw. im kumulativen Fall diagonale gestrichelte Linie) zusätzlich zu den empirischen Stichproben-Verteilungen vorläufig mit eingezeichnet.

In der geowissenschaftlichen Praxis treten aber meist empirische Verteilungen beliebiger Form auf, ohne dass sich sofort Hinweise auf die zugrunde liegenden Prozesse anbieten. Um nun trotz-

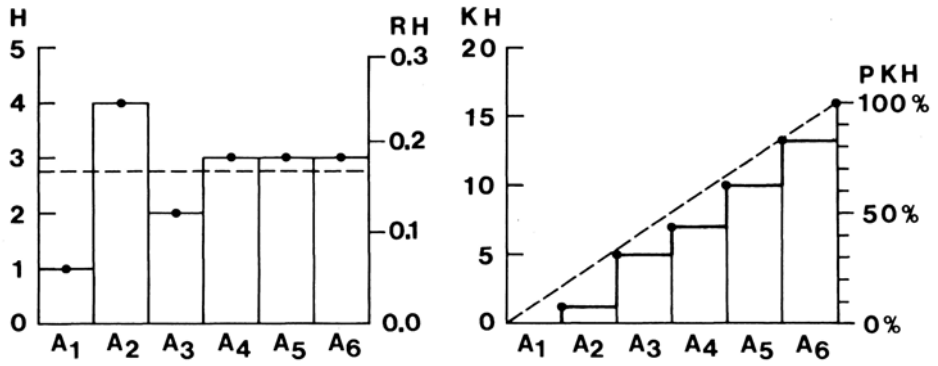

Abbildung 4: Häufigkeitsverteilung (links) und kumulative Häufigkeitsverteilung (rechts) zu Tab. 6 (Beispiel 10). Die Abszisse enthält jeweils die Merkmale A_j und die gestrichelten Linien sind (hier lineare) Schätzungen der zugehörigen (mittleren) Häufigkeitsfunktionen. In Vorausnahme späterer Intervall-Klasseneinteilungen ist der Datenbezug durch ausgefüllte Kreise gekennzeichnet (links "Intervallmitten"; rechts "Intervallobergrenzen" → "Treppenfunktion").

dem einen Schritt in Richtung Verallgemeinerung voranzukommen, zugleich um mögliche Zufälligkeiten der Stichproben-Verteilung abzuschwächen und vielleicht sogar schon zu unterdrücken, erweist es sich als zweckmäßig, sog. Klassen einzuführen. Darunter versteht man die Zusammenfassung von jeweils mehreren Merkmalen nach folgender Systematik:

- Empirische Schätzung der Klassenanzahl K auf Grund des Stichprobenumfangs n
- Im Zweifel Festlegung der geringeren Klassenzahl (mit geringerer Differenzierung), insbesondere falls „leere" Klassen entstehen sollten (d.h. Klassen ohne Merkmalswerte), die grundsätzlich zu vermeiden sind.
- Im Allgemeinen (wenn kein zwingender Grund zur Annahme nicht linearer Verteilungen besteht) Einteilung gleich großer Klassen ohne Lücken zwischen den Klassen.
- Die untere Intervallgrenze der ersten (unteren) Klasse und die obere Intervallgrenze der letzten (oberen) Klasse sollten in guter Näherung mit dem Minimum und Maximum des Datensatzes übereinstimmen bzw. gering darüber hinaus gehen.

Für die Schätzung der Klassenzahl K liegen mehrere empirische Formeln vor. Am meisten benutzt werden die folgenden (zu den ersteren vgl. Sachs, 2004; Hengst, 1967):

STURGES (1926)	$\rightarrow K = 1 + 3.32 \cdot \lg n$	(1-18)
STRAUCH ((1956)	$\rightarrow K = 1 + \lg n / \lg 2$	(1-19)
PANOFSKY UND BRIER (1958)	$\rightarrow K = 5 \cdot \lg n$	(1-20)

Dabei ist lg der dekadische Logarithmus und n wie bisher der Stichprobenumfang.
Da $1/\lg 2 \approx 3.3219$, sind (1-18) und (1-19) näherungsweise identisch. Die folgende Tabelle erlaubt eine Übersicht der Klassenzahlen K für bestimmte Stichprobenumfänge n. Beispiel 11 demonstriert den Effekt der Klassenbildung (anhand weniger Zahlenwerte).

Stichprobenumfang n	**10**	**20**	**30**	**40**	**50**	**100**	**200**	**500**	**1000**
Klassenzahl K nach (1-18)	4	5	6	6	7	8	9	10	11
Klassenzahl K nach (1-20)	5	7	7	8	8	10	12	13	15

Tabelle 7: Schätzung der Klassenzahl *K* nach den obigen empirischen Formeln für einige Werte des Stichprobenumfangs *n*

Beispiel:

11. In München wurden in den Jahren 1957 bis einschließlich 1968, jeweils im Monat April, folgende Anzahlen von Frosttagen gezählt: 9, 12, 4, 3, 0, 4, 2, 1, 4, 2, 9, 7 (n=7). Die folgende Tabelle (Tab. 8) enthält die zugehörige Häufigkeitsanalyse ohne und mit Klasseneinteilung.

Merkmalswerte A_j	0 1 2 3	4 5 6 7	8 9 10 11	12 (> 12)
Häufigkeiten H_j (ohne Klassen)	1 1 2 1	3 0 0 1	0 2 0 0	1 (0)
Klassen *KL*	0 – 3	4 – 7	8 – 11	12 – 15
Klassenbezogene Häufigkeiten H_k	5	4	2	1

Tabelle 8: Häufigkeitsverteilung ohne und mit Klassenbildung zu Beispiel 11

In Tab. 8 und der zugehörigen Abb. 5 ist nun gezeigt, was durch die Klassenbildung erreicht worden ist. Hier wie im Folgenden soll bei nicht sehr großen Stichproben (n < 100) die vorsichtigere Schätzformel für die Klassenzahl K nach Sturges bzw. Strauch benutzt werden. Es folgt dann nach (1-18) $K = K_{St} = 1 + 3.32 \cdot \lg 12 \approx 4.58$. Da im Zweifel die kleinere Klassenzahl gewählt werden soll, wird dies auf K = 4 abgerundet.
Das Beispiel 11 enthält J =13 (j = 0, 1, ... ,12) unterschiedliche Merkmale, die nach unten strikt abgeschlossen sind (da weniger als null Frosttage nicht möglich sind). Rein mathematisch ergibt sich daraus eine Klassenbreite von J/K = 13/4 = 3.25. Da es aber halbe und viertel Frosttage nicht gibt, folgt zwingend die in Tab. 8 vorgenommene Klassenteilung mit jeweils vier Merkmalen A_j pro Klasse KL. Während bei diesem Beispiel die nicht klassenbezogene Häufigkeitsverteilung, vgl. Abb. 5, ein sehr unregelmäßiges Bild ergibt, lässt sich aus der klassenbezogenen Häufigkeitsverteilung der Schluss ziehen, dass möglicherweise viele Frosttage systematisch (d.h. in Form einer auf die zugehörige Population bezogenen potentiellen Regel) weniger häufig auftreten als wenige Frosttage. Es wird somit empirisch eine in Richtung höherer Merkmale abfallende Häufigkeitsverteilung vermutet.

Methodik und Ergebnis dieses Beispiels lassen ein sehr wichtiges statistisches Prinzip erkennen, das insbesondere in Kap. 5 näher ausgeführt werden soll: Der Versuch, von einer mehr oder weniger mit Zufälligkeiten behafteten Stichprobe auf die zugehörige, allgemeiner gültige Population zu schließen, somit die Grundstruktur des betreffenden Vorgangs zu erkennen; dies ist im übrigen auch eine Voraussetzung für mögliche Prognosen (näheres in Kap. 5). Es deutet sich hier wiederum implizit an, dass es eine Beziehung zwischen Häufigkeitsverteilung und Ereigniswahrscheinlichkeit gibt. Explizit soll dies im folgenden Kap. 1.7 gezeigt werden.

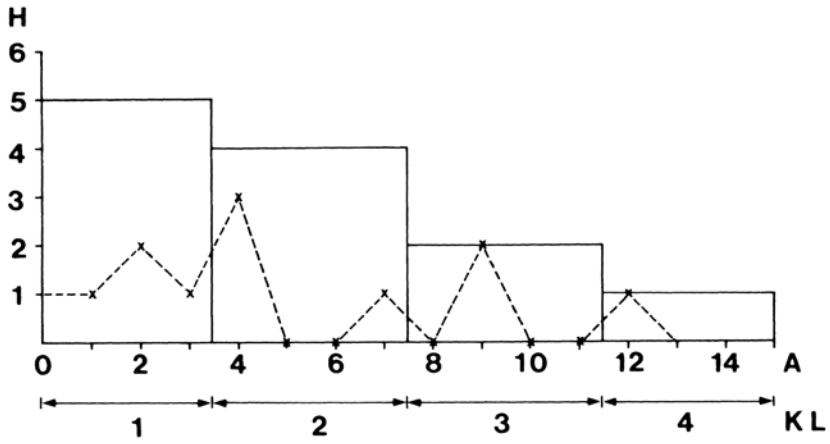

Abbildung 5: Häufigkeitsverteilung zu Tab. 8 (Beispiel 11) ohne (gestrichelt) und mit (ausgezogen) Klassenbildung. A sind die Merkmale, KL die Klassen.

Bezüglich der Klassenbildung bleibt anzumerken, dass die erreichten Vorteile natürlich mit einem Verlust an Differenzierung erkauft werden: Über die einzelnen Merkmale lässt sich dann nichts mehr aussagen. Formal wird dies dadurch berücksichtigt, dass sich alle Aussagen über eine klassenbezogene Häufigkeitsverteilung stets auf die Klassenmitten beziehen. Bei kumulativen klassenbezogenen Häufigkeitsverteilungen muss dieser Bezug jedoch definitionsgemäß für die Klassenobergrenzen gelten.

1.7 Wahrscheinlichkeit

In den vorangehenden Kapiteln ist schon einige Male der Wahrscheinlichkeitsbegriff benutzt worden, allerdings in noch recht vager Art und Weise. Doch hat sich schon abgezeichnet, dass der Weg zum Wahrscheinlichkeitsbegriff vom Begriff der Häufigkeitsverteilung und deren Verallgemeinerung ausgeht. Um dies möglichst anschaulich zu demonstrieren, werden im Folgenden drei Beispiele mit unterschiedlichem Verteilungstyp benutzt: Beispiel 10 (Würfeln, vgl. Tab. 6), Beispiel 11 (Frosttage im April, München, vgl. Tab. 8) und schließlich das nachfolgende Beispiel 12 (mit Tab. 9).

Beispiel:

12. In München hat die Mitteltemperatur des Monats Oktober in den Jahren 1911–1960 die in der folgenden Tabelle (Tab. 9) angegeben Werte angenommen. Tab. 12 listet dazu die klassenorientierte Häufigkeitsverteilung nach den in Kap. 1.6 behandelten verschiedenen Kriterien auf. Die Tabellen 10 und 11 enthalten die entsprechenden Ergebnisse für die zuvor behandelten Beispiele 10 (gewürfelte Zahlen) und 11 (Frosttage des Monats April in München).