

---

**Reinhard Göttner, Peter Fischer  
Reinhard Krieg**

**Was ist, was kann Statistik**

1975 Urania-Verlag Leipzig, Jena, Berlin

MSB: Nr. 86

Abschrift und LaTeX-Satz: 2023

<https://mathematikalpha.de>

# Inhaltsverzeichnis

<b>Vorwort</b>	<b>4</b>
<b>1 Von der Staatenbeschreibung zur modernen Statistik</b>	<b>5</b>
<b>2 Einige wichtige Grundbegriffe der Statistik</b>	<b>13</b>
<b>3 Von der Vorbereitung zur Analyse</b>	<b>19</b>
<b>4 Empirische Verteilungen und ihre Auswertung</b>	<b>22</b>
4.1 Fläche, steile, schiefe und mehrgipflige Verteilungen . . . . .	23
4.2 Zur Bildung von Gruppen oder Klassen . . . . .	35
4.3 Der Durchschnitt - ein charakteristischer Wert? . . . . .	38
4.4 Die Einzelwerte weichen vom Durchschnitt ab . . . . .	49
<b>5 Der Zusammenhang zwischen Ursache und Wirkung wird zahlenmäßig bestimmt</b>	<b>57</b>
5.1 Was bedeuten „Korrelation“ und „Regression“? . . . . .	57
5.2 Kausale Zusammenhänge . . . . .	60
5.3 Zur Regressionsanalyse . . . . .	63
5.4 Einfache, multiple und partielle Korrelation . . . . .	75
5.5 Rangkorrelation . . . . .	78
5.6 Korrelation qualitativer Merkmale . . . . .	82
<b>6 Was sind theoretische Verteilungen?</b>	<b>84</b>
6.1 Einige Aspekte der Wahrscheinlichkeitsrechnung . . . . .	85
6.2 Die Binomialverteilung - eine Verteilung für alternative Merkmalsvariationen . . . . .	90
6.3 Die Poisson-Verteilung und ihre praktische Anwendung . . . . .	95
6.4 Inhalt und Zweck der Normalverteilung . . . . .	98
<b>7 Wenige Elemente geben Auskunft über die Gesamtheit</b>	<b>105</b>
7.1 Was ist eine Stichprobe? . . . . .	105
7.2 Wie schätzen wir den Mittelwert? . . . . .	108
7.3 Wir schätzen Anteilswerte . . . . .	115
<b>8 Hypothesen werden geprüft</b>	<b>119</b>
8.1 Statistische Hypothesen und statistische Prüfverfahren . . . . .	119
8.2 Das Prüfen von Parameterhypothesen . . . . .	123
8.3 Wie prüfen wir Verteilungen? . . . . .	128
<b>9 Zeitreihen</b>	<b>135</b>
9.1 Die Grundrichtung der Entwicklung wird bestimmt . . . . .	136
9.2 Berechnung der Periodizität einer Erscheinung . . . . .	147
9.3 Die Veränderung von Zeitabschnitt zu Zeitabschnitt . . . . .	153
9.4 Korrelationsuntersuchungen bei Zeitreihen . . . . .	158

<b>10 Die Darstellung der Ergebnisse</b>	<b>162</b>
<b>11 Was sagt eine statistische Analyse aus?</b>	<b>167</b>
<b>12 Literaturhinweise</b>	<b>172</b>
<b>13 Kleines Lexikon der Statistik</b>	<b>174</b>

## Vorwort

Kennziffern, statistische Analysen, grafische und tabellarische Darstellungen statistischen Inhalts sind heute Bestandteile der beruflichen Tätigkeit vieler Menschen. Es gibt gegenwärtig kaum ein Tätigkeitsgebiet, das nicht direkt oder indirekt mit statistischen Daten in Berührung steht.

Sowohl zur Leitung und Planung der gesamten Volkswirtschaft als auch ihrer einzelnen Teile sind statistische Informationen unentbehrlich. Ebenso bedingt die Lösung naturwissenschaftlicher, technisch-konstruktiver und technologischer Aufgaben in vielen Fällen die Verwendung statistischer Werte. Ohne die Anwendung statistischer Instrumentarien ist ein tieferes Eindringen in den Gegenstand der einzelnen Wissenschaften nicht mehr vorstellbar.

Aufgabe des vorliegenden Bändchens soll es sein, den Leser in die Methoden der Statistik einzuführen und an einer Auswahl von Beispielen deren Bedeutung zu zeigen.

Allein die mathematische Statistik ist aber viel zu umfangreich, um sie in einer populärwissenschaftlichen Schrift erschöpfend darstellen zu können. Wir haben uns deshalb auf einige wesentliche Methoden beschränkt.

Mit der vorliegenden Arbeit möchten wir zur stärkeren Nutzung gerade der einfacheren statistischen Methoden auf den verschiedenen Arbeitsgebieten anregen. Für den Leser, der tiefer in die Statistik eindringen will, stehen ausgezeichnete Lehrbücher zur Verfügung, von denen einige im Literaturverzeichnis aufgeführt sind.

Die Verfasser danken allen, die auf direktem oder indirektem Wege das Entstehen dieses Buches förderten, und nehmen Anregungen und kritische Hinweise zur weiteren Verbesserung gern entgegen.

Dresden und Mittweida, im Februar 1974

Peter Fischer, Reinhard Göttner, Reinhard Krieg

# 1 Von der Staatenbeschreibung zur modernen Statistik

Wenn der einzelne heute das Wort Statistik hört, wird er an unterschiedliche Dinge denken. Vielleicht erinnern wir uns an einen Beitrag in einer Tageszeitung, in dem die Frage nach den Geldausgaben des "statistischen Durchschnittsbürgers" in unserer Republik beantwortet wurde.<sup>1</sup> Danach gab 1972 jeder von uns im Durchschnitt

1495,00 M für Nahrungsmittel,  
753,00 M für Genussmittel und  
1890,00 M für Industriewaren

aus. Konkreter Ausdruck dieser Ausgabe ist ein Pro-Kopf-Verbrauch u. a. von 133,6 kg Kartoffeln, 96,3 kg Mehl und Nahrungsmitteln, 91,2 kg Gemüse, 70,8 kg Fleisch und Fleischerzeugnisse, aber auch von 106,5 l Bier, 6,5 l Spirituosen und 1307 Zigaretten.

Diese und viele andere Ergebnisse sind im jährlich erscheinenden Statistischen Jahrbuch der Deutschen Demokratischen Republik enthalten. Die dort angeführten Zahlen geben unter anderem Auskunft über das Wachsen unserer Volkswirtschaft, der Außenhandelsbeziehungen sowie unseres Lebensstandards.

Verantwortlich für die Herausgabe dieses Werkes zeichnet die Staatliche Zentralverwaltung für Statistik beim Ministerrat der DDR, die mit ihrem Mitarbeiterstab in den Kreisen und Bezirken insbesondere die Ergebnisse der Volkswirtschaftspläne abrechnet und analysiert.

Die hier zusammengestellten und analysierten Daten sind wiederum wesentliche Anhaltspunkte für die Erarbeitung der künftigen Wirtschaftspläne, in denen, ausgehend von den erzielten Ergebnissen und den objektiven Erfordernissen unserer sozialistischen Gesellschaft, die neuen Aufgaben formuliert werden.

Im Rahmen der sich immer enger gestaltenden wirtschaftlichen Zusammenarbeit der sozialistischen Länder im RGW kommt neben der Abstimmung der Wirtschaftspläne der einzelnen Mitgliedsländer auch einer Abrechnung nach einheitlichen Kennziffern große Bedeutung zu.

Deshalb wurde auf der 16. Tagung des Rates für Gegenseitige Wirtschaftshilfe im Juni 1962 auch die Schaffung einer Ständigen Kommission für Statistik beschlossen, die durch ihre Arbeit im vergangenen Jahrzehnt wertvolle Beiträge zur weiteren ökonomischen Integration geliefert hat.

Alle statistischen Angaben haben ihren Ursprung in unserem täglichen Leben und in unserer Arbeit. Vielfach entstehen sie durch unsere eigenen Aufschreibungen im Produktionsprozess.

Die Arbeitsergebnisse, die wir beispielsweise auf unterschiedlichen Formblättern täglich abrechnen, werden im Betrieb auf der Basis des einheitlichen Systems von Rechnungsführung und Statistik zusammengefasst, verarbeitet und ausgewertet. Die Ergebnisse begegnen uns wieder in der Lohnabrechnung, aber ebenfalls in der Abrechnung der

---

<sup>1</sup>Knipping, F.: DDR-Bürger im Spiegel der Statistik. Neues Deutschland vom 21. Juli 1973, S. 9

Planerfüllung für den Arbeitsplatz, das Kollektiv und den Betrieb.

Hier treten uns auch Auswertungsergebnisse in Form von Kennziffern wie Bruttoproduktion, Arbeitsproduktivität und Grundfondsausstattung entgegen.

An das Wort Statistik werden wir auch unweigerlich denken, wenn wir die Aufnahme-prozedur in einem Krankenhaus über uns ergehen lassen müssen. Hier interessieren sowohl allgemeine Angaben zur Person als auch spezifische Daten, die unseren gegenwärtigen Gesundheitszustand genauer beschreiben. Ohne nähere Kenntnis dieser Fakten ist es für den Mediziner schwierig, wenn nicht sogar unmöglich, die zweckmäßigsten Behandlungsmethoden festzulegen.

Hingegen denken wir kaum an eine statistische Aufschreibung, wenn wir etwa eine Eintrittskarte für ein Theater oder Kino, ein Konzert oder ein Museum erwerben. Und doch dienen die auf den Eintrittskarten vermerkten laufenden Nummern zum einen der Finanzkontrolle und zum anderen der Erfassung der Zahl der Besucher.

Im Statistischen Jahrbuch ist dann in zusammengefasster Form die Zahl der Besucher bei kulturellen und sportlichen Veranstaltungen vergangener Jahre nachzulesen. Damit ist aber in der Regel erst eine Summation der Einzeldaten vorgenommen worden, an die sich eine differenzierte Auswertung noch anschließen muss.

Aus dem Gesagten erkennen wir, dass statistische Untersuchungen stets Massenerscheinungen zum Gegenstand haben.

Die Untersuchungen von Massenerscheinungen in Natur und Gesellschaft führen zu allgemeingültigen Erkenntnissen, die nicht gewonnen werden könnten, wenn wir nur die Einzelercheinung näher betrachteten. Ausschlaggebend ist dabei allerdings, dass diese statistisch erfassbaren Massenerscheinungen zufallsbedingt sind, d. h., dass sie stochastischen Charakter tragen.

So ist die Körpergröße erwachsener Menschen ebenso zufallsbedingt wie der Hektarertrag landwirtschaftlicher Erzeugnisse, die Leistungen der einzelnen Schüler in den jeweiligen Unterrichtsfächern und die Höhe des Einzelhandelsumsatzes an den einzelnen Wochentagen.

Diese Untersuchungen nun stützen sich auf ein methodisches Instrumentarium, das aus der Mathematik abgeleitet ist. Auf Grund ihrer Spezifik haben sich diese mathematischen Methoden zu einem eigenen Wissenschaftsgebiet, der auf der Wahrscheinlichkeitsrechnung aufbauenden mathematischen Statistik, entwickelt.

Die mathematische Statistik beschäftigt sich mit der Untersuchung statistischer Verteilungen, wobei zwischen empirischen und theoretischen Verteilungen unterschieden werden muss.

Eine Sonderstellung, da nicht zur mathematischen Statistik gerechnet, nehmen die statistischen Methoden zur Untersuchung von Entwicklungsvorgängen ein.

Dieser Komplex der statistischen Methoden ist nun gleichfalls mit dem Begriff "Statistik" belegt worden. So stehen wir vor der Tatsache, dass "Statistik" zum einen ein selbständiges Arbeitsgebiet charakterisiert, aus dem entsprechende Fachstatistiken, wie

die Bevölkerungsstatistik, die Medizinalstatistik oder die physikalische Statistik, hervorgehen, und zum anderen als Bezeichnung einer Wissenschaftsdisziplin (z. B. mathematische Statistik) verwendet wird.

Das mag zunächst verwundern, wird uns aber verständlich, wenn wir die historische Entwicklung der Statistik betrachten.

Wenn auch das Wort "Statistik" 1672 zum ersten Mal gedruckt wurde, so lassen sich doch Anfänge der Statistik im Sinne einer zahlenmäßigen Erfassung bestimmter Vorgänge bis in die frühe Sklavenhaltergesellschaft zurückverfolgen.

Aus dem dritten Jahrtausend v. u. Z. sind statistische Erhebungen aus Ägypten und China bekannt. Während sich die Ägypter durch das Sammeln von Zahlenmaterial u. a. einen Überblick über die für den Pyramidenbau verfügbaren Arbeitskräfte verschafften, überliefert Konfuzius in einer seiner Schriften über die alte Geschichte Chinas, dass unter dem Kaiser Yu eine Landesvermessung und Seelenzählung erfolgte, wie auch Anfänge einer Agrar-, Gewerbe- und Handelsstatistik festzustellen waren.

Herodot berichtet von dem Perserkönig Darius, dass dieser um 500 v. u. Z. die eroberten griechischen Ländereien ausmessen und nach ihrer Ertragsfähigkeit unterscheiden ließ, um danach die Kriegssteuern zu bestimmen.

Sowohl der Koran als auch die Bibel berichten über Volkszählungen. Um 1500 v. u. Z. ließ Moses nach dem Auszug der Juden aus Ägypten die männlichen Angehörigen des Volkes zählen, während König David 210 Jahre später erneut eine Volkszählung durchführen ließ, um die Siegesaussichten vor Beginn kriegesischer Auseinandersetzungen besser einschätzen zu können.

Der sogenannte Census wurde deshalb namentlich mit Angabe des Geschlechts, des Alters und der körperlichen Beschaffenheit erhoben.

Etwa 500 v. u. Z. wurde unter dem Ägypterkönig Amasis angeordnet, dass sich jeder Bewohner alljährlich dem Gouverneur des Bereiches vorzustellen und dabei seinen Namen, den Beruf und die Art und Menge seiner Erwerbs- und Unterhaltungsmittel anzugeben habe. Das führte dazu, dass Ägypten bereits Mitte des ersten Jahrhunderts v. u. Z. über ein Grundkataster sowie über Einwohnerlisten und Zivilstandsregister verfügte.

Aus all diesen historisch belegten Tatsachen ist ersichtlich, dass schon auf der frühen Stufe der Entwicklung der menschlichen Gesellschaft kein Staat ohne ein Minimum an Kenntnissen über seine Bevölkerung, seine Arbeitskräfte und die Produktion auskommen konnte. Das wird auch in den vielfältigen Aufschreibungen der griechischen Stadtstaaten und des römischen Reiches über Land und Bevölkerung deutlich, aus denen wir heute ein recht genaues Bild über die Klassenstruktur und die Besitzverhältnisse zur damaligen Zeit herleiten können.

Aus dem römischen Reich ist hier vor allem der Census zu erwähnen, der unter Tullius eingeführt und später entsprechend der Verfassung aller fünf Jahre abgehalten wurde. Jeder voll rechtsfähige Römer hatte danach vor dem zuständigen Censor zu erscheinen und öffentlich unter Eid seinen Namen, den Namen seines Vaters, sein Alter, seinen Wohnort sowie Namen, Geschlecht und Alter der Familienangehörigen anzugeben. Hinzu kamen noch Auskünfte über die Bestandteile des Vermögens und ihren Wert,

wonach dann die Abgaben an den Staat festgelegt wurden.

In all den erwähnten Aufschreibungen haben wir es mit Sammlungen wichtiger Daten über den jeweiligen Staat und seine Einwohner zu tun, die aus unserer heutigen Sicht als Verwaltungsstatistik bezeichnet werden können.

Werfen wir noch einen Blick nach Südamerika, wo sich Ende des 15. Jahrhunderts u. Z. von Südkolumbien im Norden über 4000 km nach Süden bis etwa in das Gebiet der heutigen Hafenstadt Valparaiso das Inkareich erstreckte.

Dieses Reich wurde durch eine strenge Unterteilung des Volkes sowie durch Vorschriften regiert, die das Leben der Staatsangehörigen bis ins kleinste regelten. Die im Inkareich betriebene Plan- und Vorratswirtschaft stützte sich auf das Sammeln und Auswerten umfangreicher statistischer Daten.

Da schriftliche Aufzeichnungen im Inkareich unbekannt waren, wurden die Angaben in der Hauptstadt Cuzco mit Hilfe eines wohldurchdachten Systems des Knüpfens in mehrfarbigen Knotenschnüren zusammengefasst und ausgewertet. Damit verfügten die Herrscher des Inkareiches stets über genaue Informationen zum Wirtschaftsleben des Landes und konnten auf der Grundlage dieser Kenntnisse auch weitgehend fundierte Entscheidungen über die nächsten Aufgaben in den einzelnen Teilen des Reiches treffen.

Sehen wir von diesen frühen Vorläufern der Statistik ab, so sind ihre eigentlichen Anfänge im 17. und 18. Jahrhundert in ganz verschiedenen Bereichen festzustellen.

Das Wort "Statistik" bedeutete ursprünglich Staatenwissenschaft und geht auf die deutsche Universitätsstatistik zurück. H. Conring an der Universität in Helmstedt und G. Achenwall an der Universität Göttingen hielten verbal und nicht mathematisch orientierte staatenkundliche Vorlesungen, die sich allmählich zu Beschreibungen der Staatseinrichtungen (Bevölkerung, Gewerbe, Armee usw.) entwickelten.

Ihrem Inhalt nach entsprachen diese Vorlesungen etwa dem, was wir heute als Bestandteil der ökonomischen Geographie verstehen. Dabei spielten in diesen Anfängen der Statistik Zahlen kaum eine Rolle.

Der tatsächliche Beginn unserer heutigen Statistik war jedoch die vornehmlich in England, aber auch in Holland betriebene "Politische Arithmetik", die 1661 mit J. Graunt einsetzt und ihren Höhepunkt mit R. Malthus um 1800 erreichte. Diese "politischen Arithmetiker" beschäftigten sich neben der Finanzmathematik vor allem mit Tauf-, Heirats- und Sterberegistern und dem daraus zusammengestellten statistischen Material, das sie auf darin enthaltene Regelmäßigkeiten untersuchten.

Als bedeutende Vertreter dieser neu entstandenen Schule sind der Astronom E. Halley, der 1694 die erste Sterbetafel aufstellte, und der Feldprediger J. P. Süßmilch zu nennen. In dem 1741 erstmalig erschienenen Buch von Süßmilch "Betrachtungen über die göttliche Ordnung in der Veränderung des menschlichen Geschlechts, aus der Geburt, dem Tod und der Fortpflanzung derselben erwiesen" bemerkt dieser erstmalig den Tatbestand des statistischen Ausgleiches und des Gesetzes der großen Zahlen.

Er erkannte diesen Tatbestand allerdings nicht wie J. Bernoulli (1713) als mathema-



tisches Gesetz, sondern deutete diese Erkenntnis als einen "Beweis" für die von Gott optimal eingerichtete Welt.

Diese Entwicklung setzte sich fort in der Suche nach neuen Methoden, um Bevölkerung, Sterblichkeit, Fruchtbarkeit usw. abzuschätzen und damit Unterlagen für die Bestimmung von Versicherungsprämien oder für naturwissenschaftliche und ökonomische Theorien zu gewinnen.

Es fehlte aber nach wie vor eine Nutzung dieser ersten Erkenntnisse in der staatlichen Verwaltung.

Nach der Französischen Revolution schuf das aufstrebende Bürgertum 1800 in Paris das erste staatliche "Bureau de Statistique", dessen Direktor der bekannte Mathematiker J. B. J. Fourier wurde. Im Jahre 1801 erfolgten gleichzeitig in Frankreich und England durch den Staat organisierte Volkszählungen. Allerdings konnte von einer weitgehenden Verarbeitung des Materials mit entsprechenden Aussagen noch keine Rede sein.

Erst der Belgier A. Quetelet lieferte in seinem 1835 erschienenen Buch "Sur l'homme et le développement de ses facultés ou Essai de physique sociale" ein erstes Beispiel für eine sinnvolle und zielgerichtete Verarbeitung statistischen Materials.

Die voneinander abweichenden Messergebnisse astronomischer, geodätischer und später physikalischer Größen führten, maßgeblich inspiriert durch C. F. Gauß und R. Helmert, zur Fehler- und Ausgleichsrechnung, in der bereits wichtige Methoden der Statistik - wenn auch zunächst nur für die Lösung von Spezialaufgaben genutzt - entstanden.

Mit der Beschreibung der Brownschen Bewegung durch ein statistisches Modell setzte Ende des vergangenen Jahrhunderts L. Boltzmann in bedeutendem Maße die Statistik für physikalische Untersuchungen ein. Er begründete damit die physikalische Statistik und bewies, dass statistische Methoden auch zur theoretischen Erklärung physikalischer Prozesse herangezogen werden können.

Einen besonderen Aufschwung erfuhr die Statistik durch die umfangreichen Arbeiten der Engländer K. Pearson (1857-1936) und R. A. Fisher (1890-1962).

Beide formulierten allgemeine Prinzipien zur Konstruktion statistischer Methoden und eine Reihe neuer Methoden der Statistik. Mit der Vervollkommenheit der Methodik ging in immer stärkerem Maße die Anwendung der Statistik in den Natur- und Gesellschaftswissenschaften einher. Genannt werden sollen hier vor allem die Meteorologie, die Biologie und die Agrar- und Wirtschaftswissenschaften. Im Zuge der weiteren Spezialisierung bestimmter Wissenschaften trug die Statistik mit zur Herausbildung neuer Wissenschaftsdisziplinen, wie etwa der Biometrie und der Psychometrie, bei.

Die natur- und gesellschaftswissenschaftlichen Disziplinen haben andererseits häufig zur Präzisierung und Erweiterung des methodischen Instrumentariums der Statistik beigetragen. So ist seit den zwanziger Jahren unseres Jahrhunderts verstärkt versucht worden, einzelne statistische Verfahren zur Beantwortung spezieller Fragen dahingehend zu verallgemeinern, dass sie darüber hinaus universell eingesetzt werden können.

Zu nennen sind hier

- die auf der Methode der kleinsten Abweichungssumme beruhende Regressi-

onsanalyse von K. Pearson

- das Maximum-Likelihood-Prinzip und die Varianzanalyse von R. A. Fisher
- der Konfidenzschluss und die allgemeine Testtheorie von E. S. Pearson und J. Neyman
- die grundlegenden Arbeiten zur Wahrscheinlichkeitsrechnung von A. A. Markow
- der nichtparametrische Test von G. N. Kolmogorow und N. W. Smirnow
- die Sequentialanalyse und
- die Theorie der Entscheidungsfunktionen von A. Wald.

Diese und noch einige andere Methoden der mathematischen Statistik lassen sich zu einer allgemeinen Theorie der Verteilungsfunktionen zusammenfassen, die wir heute als die Theorie der mathematischen Statistik bezeichnen.

Von der Theorie der mathematischen Statistik nicht zu trennen ist die Theorie der Wahrscheinlichkeitsrechnung, auf der die erstere mit aufbaut. Ihr Anfang liegt im 17. Jahrhundert, wobei ihre Grundlagen vor allem durch B. Pascal, Ch. Huygens, P. de Fermat sowie im besonderen Maße durch J. Bernoulli geschaffen wurden.

In seinem Buch "Versuch einer Philosophie der Wahrscheinlichkeit" fasste 1812 der Astronom, Physiker und Mathematiker P. S. Laplace den damaligen Erkenntnisstand der Wahrscheinlichkeitsrechnung zusammen und ergänzte ihn durch eigene fundamentale Forschungsergebnisse auf diesem Gebiet.

Laplace formulierte aber auch eine Theorie von Th. Bayes neu, die 1763/64 nach dessen Tode veröffentlicht worden war und heute als die Theorie der Wahrscheinlichkeit a posteriori bezeichnet wird.

Die großartigen theoretischen Ansätze wurden außer in den Naturwissenschaften aber auch auf Gebieten eingesetzt, wo eine objektive Einschätzung der Wahrscheinlichkeit unmöglich war.

Die Folge war, dass Mitte des 19. Jahrhunderts die Wahrscheinlichkeitsrechnung arg in Misskredit geriet und von kaum einem mehr ernst genommen wurde. Dem russischen Akademiemitglied P. L. Tschebyscheff (1821-1894) gebührt das Verdienst, die Wahrscheinlichkeitsrechnung von nicht zur Sache gehörenden Dingen bereinigt und sie zu einer exakten mathematischen Wissenschaft mit einer eigenen Theorie und einem spezifischen mathematischen Instrumentarium entwickelt zu haben.

In den letzten Jahrzehnten waren es insbesondere sowjetische Wissenschaftler, wie A. N. Kolmogorow, A. J. Chintschin, A. A. Markow und B. W. Gnedenko, die wesentlich zur Weiterentwicklung der Wahrscheinlichkeitsrechnung und der mathematischen Statistik in Theorie und Praxis beigetragen haben.

Die historische Entwicklung der Statistik macht auch verständlich, warum Karl Marx Mitte des 19. Jahrhunderts noch formulieren musste, dass eine Wissenschaft erst dann als voll entwickelt zu bezeichnen ist, wenn es ihr gelingt, sich der Mathematik zu bedienen.

Zu diesem Zeitpunkt war das methodische Instrumentarium der Statistik noch nicht so weit ausgebildet, um eine umfassende Analyse quantitativ fassbarer Erscheinungen und Prozesse zu ermöglichen. So mussten sich Marx und Engels bei ihren umfangreichen

Untersuchungen sozialökonomischer Erscheinungen und Prozesse auf die Anwendung der zu ihrer Zeit als zutreffend erkannten Methoden beschränken.

Dass aber die Mathematik zur Entwicklung der Produktivkräfte einen wesentlichen Beitrag zu leisten vermag, war schon in den Antworten auf ein Preisausschreiben der Berliner Akademie im Jahre 1769 zum Ausdruck gekommen. Die eingereichten Vorschläge zur Preisaufgabe "Über die Verbindung der Naturlehre mit der Landwirtschaft" wiesen auf die Notwendigkeit hin, die "Ökonomie mit der Mathematik" und die "theoretischen mit den praktischen Wissenschaften" zu verbinden, um wesentliche Fortschritte in der Wirtschaft zu erreichen.<sup>2</sup>

W. I. Lenin ist in zahlreichen seiner Arbeiten auf die Bedeutung der Statistik sowohl für theoretische Untersuchungen als auch für die Planung und Kontrolle durch den sozialistischen Staat eingegangen.

Er gelangte in seinen vielfältigen Analysen zu der Erkenntnis, dass es eine vordringliche Aufgabe der Statistik bei der Untersuchung gesellschaftlicher Erscheinungen und Prozesse sein muss, sozialökonomisch unterschiedliche Merkmalsvariationen auf den Grundlagen der Erkenntnisse der marxistischen Philosophie und der politischen Ökonomie exakt zu bestimmen und sie deutlich voneinander abzugrenzen.

Sehr anschaulich hat Lenin z. B. in seinem Werk "Neue Daten über die Entwicklungsgesetze des Kapitalismus" vom marxistischen Standpunkt die Gruppenbildung wissenschaftlich angewendet und darauf aufbauend die um die Jahrhundertwende tatsächlich bestehenden sozialökonomischen Unterschiede in der amerikanischen Landwirtschaft dargelegt.<sup>3</sup>

Er wies gleichzeitig an diesen und anderen Beispielen nach, in welchem erheblichem Maße die bürgerliche Statistik die in dieser Gesellschaftsordnung bestehenden Widersprüche verschleiert und falsch interpretiert.

Lenin stellte fest: "Will sich der wissenschaftliche Forscher frei halten von dem Vorwurf, dass er, bewusst oder unbewusst, der Bourgeoisie zu Gefallen die Lage der kleinen Landwirte beschönigt, die ruiniert und erdrückt werden, so muss er seine Aufgabe vor allem und am meisten darin sehen, die Merkmale der Ruinierung, die gar nicht so einfach und gleichartig sind, genau zu bestimmen, sodann diese Merkmale aufzudecken, zu verfolgen und den Grad ihrer Verbreitung sowie ihrer zeitlichen Veränderung möglichst zu berücksichtigen.

Diese besonders wichtige Seite der Sache pflegen die Ökonomen und Statistiker unserer Zeit am allerwenigsten zu beachten."<sup>4</sup> Diese Mahnung gilt gleichermaßen heute, wenn wir aus kapitalistischen Staaten stammendes Zahlenmaterial bearbeiten und verarbeiten wollen.

Lenin umriss aber auch die Rolle der Statistik als Kontrollinstrument für die Leitung und

---

<sup>2</sup>Müller, H.-H.: Streiflichter aus der Entwicklung der Produktivkräfte. Neues Deutschland vom 23. Juni 1973, S. 15

<sup>3</sup>Lenin, W.I.: Neue Daten über die Entwicklungsgesetze des Kapitalismus. Werke Band 22, Dietz Verlag, Berlin 1960, S.5-100

<sup>4</sup>Ebenda, S. 65

Planung des sozialistischen Staates. In seinem Werk "Staat und Revolution" schrieb er: "Rechnungsführung und Kontrolle - das ist das Wichtigste, was zum 'Ingangsetzen', zum richtigen Funktionieren der kommunistischen Gesellschaft in ihrer 1. Phase erforderlich ist."<sup>5</sup>

Das in unserer Republik angewendete einheitliche System von Rechnungsführung und Statistik trägt dieser Forderung Rechnung und sichert eine einheitliche Abrechnung der wert- und mengenmäßigen Seite aller Bestandteile des Produktions- und Reproduktionsprozesses.

Mit der fortschreitenden Vervollkommnung der mathematischen und mathematisch-statistischen Methoden und ihrer Nutzung in den einzelnen Wissenschaftsdisziplinen ist ein zunehmend tieferes Eindringen in den jeweiligen Untersuchungsgegenstand möglich geworden.

Das setzt allerdings voraus, dass der Fachwissenschaftler diese Methoden beherrscht und die erzielten Ergebnisse richtig zu interpretieren vermag.

Gerade durch falsche und unzutreffende Aussagen ist die Statistik häufig in das Kreuzfeuer der Kritik geraten, wobei es sich eben hier um eine Kritik nicht vornehmlich am methodischen Instrumentarium der Statistik, sondern an der Nutzung dieser Methoden in den einzelnen Fachdisziplinen handelt.

Die bürgerliche Statistik hat eine Vielzahl von Beispielen für die unzulässige Anwendung mathematisch-statistischer Methoden geliefert, die den Engländer Sir Francis Galton zu der sarkastischen Bemerkung veranlasste, dass es drei Arten von Lügen gäbe: erstens die Notlüge, die entschuldbar sei, zweitens die gemeine Lüge, für die keine Entschuldigung gelten könne, und drittens die Statistik.<sup>6</sup>

Die Anwendung der Statistik unter sozialistischen Produktionsverhältnissen geht, wie bereits hervorgehoben wurde, von den Erkenntnissen des dialektischen und historischen Materialismus und der marxistischen Politischen Ökonomie aus. Nur wenn sie sich dieser Grundlagen bedient, kann sie zu objektiv richtigen Erkenntnissen gelangen. Dies gilt gleichermaßen für die Anwendung der Statistik in den Natur- wie in den Gesellschaftswissenschaften.

Wenn wir beispielsweise von der Möglichkeit und Notwendigkeit der bewussten Ausnutzung der objektiv existierenden ökonomischen Gesetze des Sozialismus sprechen, so muss der Wirtschaftswissenschaftler sich auch statistischer Methoden bedienen, um die Wirkungsweise, -intensität und -richtung sowohl der einzelnen Gesetze als auch aller wirkenden ökonomischen Gesetze in ihrer Komplexität erkennen und ausnutzen zu können.

---

<sup>5</sup>Lenin, W. I.: Staat und Revolution. Werke Band 25, Dietz Verlag, Berlin 1960, S. 487

<sup>6</sup>Mises, R. v.: Wahrscheinlichkeit, Statistik und Wahrheit. Wien 1928, S.1

## 2 Einige wichtige Grundbegriffe der Statistik

Bei statistischen Untersuchungen treten spezifische Begriffe auf, deren Kenntnis für das Verstehen der durch die Statistik zu analysierenden Zusammenhänge erforderlich ist. Es erscheint daher sinnvoll, vor dem Behandeln bestimmter Instrumente der Statistik einige Grundbegriffe zu erläutern, die im weiteren häufig benutzt werden.

Wir gehen zunächst von dem Begriff der Menge aus. Unter einer Menge versteht man ganz allgemein eine Gesamtheit von Objekten oder Ereignissen, d. h. die Zusammenfassung von verschiedenen Einzelobjekten, an denen ein oder mehrere Merkmale beobachtet werden können, zu einem einheitlichen Ganzen. Die einzelnen Objekte oder Ereignisse werden dabei Elemente der Menge genannt.

In diesen wenigen Sätzen sind gleich mehrere Begriffe enthalten, die wir für die weiteren Betrachtungen benötigen. Es sind die Begriffe

- Menge
- Gesamtheit
- Element
- Merkmal.

Wenden wir uns dem Begriff Menge etwas näher zu. Wir können z. B. von der Menge des Wassers im Ozean, von der Menge aller Lebewesen auf der Erde, von der Menge der Schüler einer Klasse, von der Menge der erzeugten Produkte eines Betriebes oder von der Menge aller Punkte auf einer gegebenen Linie sprechen. Wir erkennen hierbei schon, dass die Menge ein sehr allgemeiner Begriff ist.

Die exakte Definition eines Begriffes erfordert, dass man vor allem sagen kann, von welchem umfassenderen Begriff er durch Spezialisierung abzuleiten ist. Das ist aber bei dem Begriff der Menge wegen dessen Allgemeinheit nicht möglich, denn es gibt in der Mathematik keinen allgemeineren Begriff als den der Menge.<sup>7</sup>

Es ist daher auch nicht möglich, eine strenge mathematische Definition des Begriffes Menge zu formulieren. Wir können deshalb die Menge, wie bereits erwähnt, lediglich als eine Gesamtheit von Objekten oder Ereignissen bezeichnen.

Zur Erläuterung der anderen Begriffe betrachten wir zunächst zwei Beispiele.<sup>8</sup>

In einer Schulklasse wird das Körpergewicht der Schüler festgestellt. Die Schulklasse, bestehend aus  $n$  Schülern, ist dabei die Menge. Die einzelnen Schüler sind die Elemente dieser Menge, und das Körpergewicht ist das untersuchte Merkmal. Das untersuchte Merkmal wird dabei durch die Variable  $X$  charakterisiert.

Diese Zusammenhänge können wie folgt dargestellt werden, wobei zum Vergleich auch die allgemeine Schreibweise angegeben ist:

---

<sup>7</sup>Vgl. Wilenkin, W. J.: Unterhaltsame Mengenlehre. B. G. Teubner Verlagsgesellschaft, Leipzig 1972, S. 9

<sup>8</sup>Vgl. Clauß, G., Ebner, H.: Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Volk und Wissen - Volkseigener Verlag, Berlin 1967, S. 13ff.

$$\text{Menge } M \left\{ \begin{array}{l} \text{Element 1} \\ \text{Element 2} \\ \text{Element 3} \\ \dots \\ \text{Element } n \end{array} \right\} \text{ Untersuchtes Merkmal } \begin{array}{l} \text{charakterisierbar durch eine} \\ \text{Variable } X \text{ mit den Werten} \\ x_i (i = 1, 2, \dots, n) \end{array} \quad (1)$$

Beispiel:

$$\begin{array}{l} \text{Menge } M \\ \text{Schulklasse} \\ \text{mit } n \text{ Schülern} \end{array} \left\{ \begin{array}{l} \text{Schüler 1} \\ \text{Schüler 2} \\ \text{Schüler 3} \\ \dots \\ \text{Schüler } n \end{array} \right\} \begin{array}{l} \text{Untersuchtes Merkmal} \\ \text{Körpergewicht} \end{array} \begin{array}{l} x_1 = 55,2\text{kg} \\ x_2 = 59,7\text{kg} \\ x_3 = 48,4\text{kg} \\ \dots \\ x_n = 53,1\text{kg} \end{array}$$

Die geschweiften Klammern sind bei der Darstellung von Mengen üblich. Sie weisen darauf hin, dass die einzelnen Elemente zu einer Menge  $M$  gehören.

Als zweites Beispiel soll die erreichte Augenzahl beim Würfeln mit einem Würfel betrachtet werden. Die Anzahl der Würfe bildet dabei die Menge, jeder einzelne Wurf ein Element, und die jeweils erreichte Augenzahl ist das untersuchte Merkmal. Wir können also in der Form des vorhergehenden Beispiels schreiben:

$$\begin{array}{l} \text{Menge } M \\ n \text{ Würfe mit} \\ \text{einem Würfel} \end{array} \left\{ \begin{array}{l} \text{Wurf 1} \\ \text{Wurf 2} \\ \text{Wurf 3} \\ \dots \\ \text{Wurf } n \end{array} \right\} \begin{array}{l} \text{Untersuchtes Merkmal} \\ \text{Augenzahl} \end{array} \begin{array}{l} x_1 = 2 \\ x_2 = 3 \\ x_3 = 5 \\ \dots \\ x_n = 3 \end{array}$$

Besteht eine Menge aus einer endlichen Zahl von Elementen, so ist das eine endliche Menge (z. B. Zahl der Schüler einer Klasse, Zahl der Würfe in einer Viertelstunde). Sind unendlich viele Elemente möglich, bezeichnet man die Menge als unendliche Menge (z. B. mögliche Würfe beim Würfeln überhaupt).

Die bisherigen Erläuterungen haben auch schon den Inhalt des Begriffes Element deutlich werden lassen: Ein Element ist die gewählte oder festgelegte kleinste Einheit einer Menge.

An dieser Stelle sei aber bereits darauf hingewiesen, dass in der Statistik nicht für alle Begriffe einheitliche Bezeichnungen gebraucht werden. So finden wir statt des Begriffes Element z. B. auch die Bezeichnung Einheit.

Die Elemente einer Menge bilden die Gesamtheit. In der Statistik ist der Begriff der Grundgesamtheit  $G$  wichtig. Darunter ist zunächst die Menge aller gleichartigen Objekte oder Ereignisse zu verstehen, die zu einer Menge überhaupt gehören können. Mit dieser Definition wollen wir uns vorerst begnügen. Sie wird später noch exakter formuliert.

An unseren beiden Beispielen haben wir vermerkt, dass das untersuchte Merkmal durch eine Variable  $X$  charakterisiert wird, die eine Zufallsgröße darstellt. Derartige Zufallsvariablen sind in der mathematischen Statistik bedeutungsvoll. Eine Zufallsvariable ist

eine Variable, die Werte in bestimmten vorgegebenen Intervallen der Zahlengeraden mit bestimmten Wahrscheinlichkeiten annimmt. Sie wird auch wie folgt definiert:

"Unter einer Zufallsvariablen  $X$  versteht man die Funktion  $X = X(i)$ ,  $i = 1, 2, 3, \dots, n$ , die für alle Elemente  $i$  der Menge  $M$  definiert ist und jedem Element  $i$  einen Zahlenwert  $x_i$  eindeutig zuordnet."<sup>9</sup>

Die quantitative Ausprägung des Merkmals an jedem Element der Menge wird durch die Werte  $x_i$  ( $i = 1, 2, 3, \dots, n$ ) charakterisiert, wobei die  $x_i$  als Realisationen von  $X$  bezeichnet werden.

In unseren Beispielen waren die Merkmale das Gewicht der einzelnen Schüler bzw. die erreichte Augenzahl bei jedem Wurf mit einem Würfel. Die Realisationen ändern sich von Element zu Element durch zufällige Einflüsse. Das heißt, dass man z. B. bei einem Wurf mit einem Würfel vorher nicht weiß, welche Augenzahl eintreffen wird; das Ergebnis eines Wurfs lässt auch keine Schlüsse auf das Ergebnis des folgenden Wurfs zu usw.

Oder beim anderen Beispiel: Wenn das Körpergewicht von Schülern bestimmt wird, hängt der Wert in jedem einzelnen Fall von dem jeweiligen Schüler ab.

Jedem Element der betreffenden Menge ist aber eindeutig ein Zahlenwert  $x_i$  (der also zufälligen Einflüssen unterliegt) zugeordnet. Daraus ergibt sich, dass Zufallsvariablen Funktionen sind.

Aussagen über Zufallsvariablen lassen sich nur treffen, wenn wir die Verteilung dieser Variablen kennen. Darunter ist zu verstehen, dass wir wissen müssen, welche möglichen Werte eine Zufallsvariable annehmen kann und mit welchen Wahrscheinlichkeiten sie diese Werte annimmt. Dieser Sachverhalt wird in den folgenden Kapiteln noch deutlich werden.

Wir bezeichnen die Zufallsvariablen mit großen lateinischen Buchstaben ( $X, Y, \dots$ ) und die Werte, die sie annehmen können, mit kleinen lateinischen Buchstaben ( $x, y, \dots$ ).

Wir kennen zwei Arten von Zufallsvariablen:

- die stetige Zufallsvariable und
- die diskrete Zufallsvariable.

Eine Zufallsvariable heißt stetig, wenn sie in einem bestimmten Intervall jeden beliebigen Wert annehmen kann. So kann das Körpergewicht eines Schülers jeden beliebigen Wert in einem Intervall von beispielsweise 35 kg bis 90 kg annehmen.

Weitere Beispiele für stetige Zufallsvariablen sind: die Geschwindigkeit von Kraftfahrzeugen, der Ernteertrag pro Hektar, die Abweichungen vom Sollmaß bei hergestellten Werkstücken usw.

Als diskret wird eine Zufallsvariable  $X$  bezeichnet, wenn sie nur endlich viele oder abzählbar unendlich viele Werte annehmen kann. So können beim Wurf mit einem Würfel nur die abzählbaren Werte 1, 2, 3, 4, 5 oder 6 auftreten. Auch bei der Anzahl der Schüler je Klasse, bei der Anzahl der in einem Betrieb pro Schicht aufgetretenen Störungen

---

<sup>9</sup>Ebenda, S. 14

oder bei der Anzahl der pro Stunde in einem Kaufhaus zu bedienenden Kunden handelt es sich um diskrete Größen.

Diskrete Zufallsvariablen erhalten wir ebenfalls, wenn bei der Qualitätskontrolle von Werkstücken lediglich nach dem Prinzip gut oder schlecht unterschieden wird. Dabei werden nicht die jeweiligen Messergebnisse notiert, sondern es wird vermerkt, ob das Werkstück eine vorgegebene Toleranzgrenze einhält oder nicht.

Wir können nun wieder auf den Begriff der Grundgesamtheit zurückkommen und nach diesen gewonnenen Erkenntnissen feststellen, dass die Grundgesamtheit die Menge aller möglichen Realisationen einer Zufallsvariablen bildet.

Die Menge aller möglichen Realisationen ist der Umfang der Grundgesamtheit  $N$  und stellt den Untersuchungsbereich dar, wobei  $N$  endlich (100 Würfe mit einem Würfel) oder unendlich groß (mögliche Würfe mit einem Würfel) sein kann.

Wenn wir der Grundgesamtheit durch zufällige Auswahl  $n$  Elemente entnehmen, dann erhalten wir eine Stichprobe. Die Anzahl  $n$  der in der Stichprobe auftretenden Zahlenwerte der Zufallsvariablen  $X$  wird dabei Umfang der Stichprobe genannt.

Aus den bisherigen Darlegungen ist bereits deutlich geworden, dass nur Untersuchungsergebnisse statistisch bearbeitet werden können, die sich auf Mengen beziehen. Einzelne Elemente (ein einzelnes Element für sich) eignen sich dagegen nicht für statistische Untersuchungen.

Wir haben weiter erkannt, dass die Untersuchungsergebnisse das Unterscheiden quantifizierbarer Eigenschaften zulassen müssen. Damit wären wir bei der Betrachtung des statistischen Merkmals angelangt, das auch schon in unseren zwei Beispielen eine Rolle spielte.

Unter den Merkmalen verstehen wir in der Statistik die betrachteten Charakteristika (Eigenschaften) der Untersuchungsobjekte.

In unseren Beispielen waren in diesem Sinne das Körpergewicht der Schüler bzw. die erreichte Augenzahl beim Würfeln Merkmale. Die aus der Untersuchung resultierenden Einzelergebnisse sind die Beobachtungswerte bzw. empirischen Werte.

Die an den einzelnen Elementen festgestellten Merkmale ergeben unterschiedliche Beträge. Sie werden Variationen des Merkmals bzw. Merkmalsvariationen genannt.

Statistische Merkmale können nach verschiedenen Kriterien eingeteilt werden. Wir wollen hier die wichtigsten dieser Einteilungen aufzeigen und dabei mit der Gliederung nach nominalen, ordinalen und metrischen Gesichtspunkten beginnen. Bei der Quantifizierung von Sachverhalten kann zwischen

- Nominalwerten
  - Ordinalwerten und
  - metrischen Werten
- unterschieden werden.

Definieren wir für eine Variable qualitativ verschiedene Klassen bzw. Gruppen und ermitteln anschließend die Häufigkeit, mit der jede Klasse bzw. Gruppe besetzt ist, erhalten



wir Nominalwerte. Die betreffenden Merkmale bezeichnen wir als nominale Merkmale. Typisch für diese Merkmale ist, dass die Merkmalsvariationen nicht durch Zahlen ausgedrückt werden können, sondern lediglich verbal, wie z. B. männlich oder weiblich bzw. rot, grün oder blau. In dieser Tatsache liegt auch die Bezeichnung nominal begründet, denn nomen ist das lateinische Wort für Namen.

Wir ermitteln dabei die Häufigkeit, mit der verschiedene Variationen eines bestimmten Merkmals auftreten. So kann gezählt werden, wie viele Einwohner einer Stadt männlichen Geschlechts und wie viele weiblichen Geschlechts sind, wie viele Beschäftigte eines Betriebes bestimmten Berufsgruppen angehören, welche Staatsangehörigkeit die in die DDR einreisenden Personen besitzen usw. Dabei wird aber keine Wertigkeit der Objekte vorgenommen, denn es gibt z. B. keine derartige Wertigkeit zwischen Männern und Frauen usw.

Das Feststellen der Häufigkeiten von Sachverhalten in Form der Nominalwerte ist die einfachste Form der Quantifizierung. Sie wird durch Zählen realisiert. Unter Zählen ist das Ermitteln einer Anzahl diskreter Größen zu verstehen.

Treten nur zwei Merkmalsvariationen auf, z. B. männlich oder weiblich, gibt es somit nur die Alternative Merkmal A vorhanden oder Merkmal A nicht vorhanden, dann liegen als spezielle nominale Merkmale alternative Merkmale vor.

Als zweite Form der Quantifizierung von Sachverhalten hatten wir die Ordinalwerte genannt. Ordinalwerte liegen dann vor, wenn eine bestimmte Rangfolge (ordo: lateinisch für Rang) durch Zahlenangaben ausgedrückt wird. Werden in einem Betrieb die erzeugten Produkte nach bestimmten Kriterien in Güteklassen eingeteilt, wie "sehr gut", "gut", "mittelmäßig", "nur bedingt brauchbar", "schlecht (unbrauchbar)", dann besteht hierin eine Rangfolge.

Man könnte statt dessen auch sagen 1. Qualität, 2. Qualität, 3. Qualität usw., was eine Rangordnung charakterisieren würde. Es erfolgt also im gewissen Sinne eine numerische Wertung. Sinnvolle Verhältniszahlen lassen sich damit jedoch nicht bilden, denn die 2. Qualität ist in der Regel nicht halb so gut oder halb so schlecht wie die 4. Qualität. Es bleibt undefiniert, wie stark die Differenzen zwischen den betrachteten Objekten sind. Das ist der Hauptmangel der Ordinalwerte. Der Informationsgewinn ist jedoch größer als bei den Nominalwerten.

Ordinalwerte eignen sich z. B. zur Quantifizierung von Merkmalen, für die keine Maßeinheiten vorliegen. Sie nehmen eine Zwischenstellung zwischen den Nominalwerten und den metrischen Werten ein. Den größten Informationsgewinn liefern metrische Werte (metrisch: durch Zahlen- und Maßangaben ausdrückbar).

Wir sprechen von metrischen Maßzahlen und von metrischen Merkmalen, wenn wir bei der Untersuchung eines Sachverhaltes die Merkmalsvariationen zahlenmäßig ausdrücken können, ohne dass die Zahlenangaben lediglich eine Ordnungszahl oder eine Rangfolge darstellen. Beispiele für metrische Maßzahlen sind die Werte beim Messen von Längen, beim Messen der Zeit, aber z. B. auch die zahlenmäßige Verteilung der LPG innerhalb der Bezirke der DDR im Hinblick auf die Nutzfläche der LPG in ha oder die Zahl der

Mädchen, aufgeteilt nach dem Alter, in dem sie heiraten. Für metrische Angaben lassen sich sinnvolle Verhältniszahlen bilden.

Nominal- und Ordinalwerte sind immer diskrete Größen.

Metrische Werte können diskrete (Anzahl der Gespräche in einer Telefonzentrale in bestimmten Zeitabschnitten) oder stetige Größen (Körpergewicht von Schülern) sein. In der Statistik wird oft auch eine Gliederung der Merkmale in quantitative und qualitative Merkmale vorgenommen. Unter quantitativen Merkmalen versteht man Merkmale, deren Merkmalsvariationen zu einer Kategorie gehören und sich nur hinsichtlich ihrer Quantität unterscheiden (z. B. Größe, Gewicht, Temperatur, Länge u. a.). Sie entsprechen somit den metrischen Merkmalen.

Ein qualitatives Merkmal liegt dann vor, wenn die Merkmalsvariationen in bestimmte, qualitativ unterschiedliche Kategorien eingeteilt werden. Das ist bei allen Nominalwerten eindeutig der Fall. Unter einer qualitativen Gliederung versteht man z.B. auch die Einteilung des Ernteertrages je Hektar nach der Rangordnung gut, mittel oder schlecht, obwohl hier natürlich auch quantitative Erwägungen von Bedeutung sind. Die Einteilung in quantitative und qualitative Merkmale ist häufig von fachlichen Überlegungen abhängig.

So kann z. B. der Ernteertrag pro Hektar sowohl quantitativ nach Dezitonnen als auch qualitativ nach der Rangordnung gut, mittel oder schlecht erfasst werden.

Eine letzte, für unsere Betrachtungen wichtige Untergliederung ist die Einteilung in sachliche, örtliche und zeitliche Merkmale. Sachliche Merkmale können z.B. das Gewicht, das Geschlecht oder die Länge sein. Örtliche Merkmale ergeben sich durch das Beachten territorialer Standorte (Kreis, Bezirk o.ä.). Zeitliche Merkmale werden u. a. aus dem Beachten des Zeitraumes der Erfassung hergeleitet.

### 3 Von der Vorbereitung zur Analyse

Nachdem wir einige wesentliche Begriffe und Grundgedanken erörtert haben, die erforderlich sind, um die statistischen Methoden anwenden zu können, soll uns im folgenden der konkrete Ablauf der statistischen Arbeit interessieren. Unabhängig vom Untersuchungsgegenstand müssen immer vier Etappen, die in ihrer Reihenfolge und Verbundenheit in Abbildung 1 dargestellt sind, nacheinander durchlaufen werden.

Ausgehend vom Gegenstand der Untersuchung, werden in der Vorbereitungsetappe sowohl die inhaltlichen als auch die organisatorischen Aufgaben der sich anschließenden Etappen festgelegt.

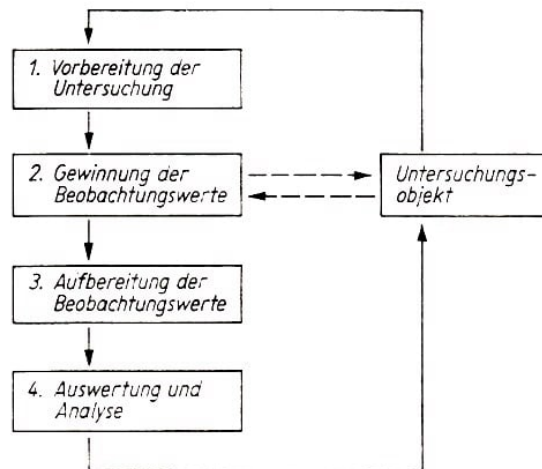


Abb.1 Die Etappen der statistischen Arbeit

An die Vorbereitung schließt sich dann die Datenerfassung und -aufbereitung an, wobei unter "Aufbereitung" das Zusammenstellen und Zusammenfassen des Urmaterials vor allem in Tabellen zu verstehen ist. Die am Ende stehende Auswertung und Analyse der Daten hat das in der Vorbereitungsetappe formulierte Ziel der Untersuchung zu erfüllen.

Während bei einer erstmaligen Untersuchung alle vier Etappen durchlaufen werden müssen, können bei wiederholten Untersuchungen einzelne Etappen in ihrem Umfang reduziert werden.

So wird bei regelmäßig wiederholten Untersuchungen volkswirtschaftlicher, naturwissenschaftlicher und technischtechnologischer Prozesse der Umfang der Vorbereitung, bedingt durch die bereits gewonnenen und berücksichtigten Erkenntnisse über die Datenerfassung, -aufbereitung und -auswertung, auf ein Minimum begrenzt werden können.

Von der sachgerechten Vorbereitung hängt maßgeblich die Zuverlässigkeit der Ergebnisse ab. Sind Versäumnisse in der Vorbereitung aufgetreten, so werden diese meist nicht sofort sichtbar, sondern führen erst in einer der folgenden Etappen zu Schwierigkeiten. Das ursprünglich fixierte Ziel der Untersuchung wird dann nicht erfüllt, es macht sich erforderlich, den gesamten Zyklus nach entsprechenden Korrekturen vollständig zu wiederholen.

Die vorbereitenden Arbeiten finden ihren Niederschlag in einem entsprechenden Untersuchungsplan. In ihm sind die inhaltlichen und organisatorischen Aspekte der Untersuchung präzisiert. Haben wir im Untersuchungsplan dann alle wesentlichen Gesichtspunkte berücksichtigt und die Notwendigkeit der Untersuchung nachweisen können, sind im nächsten Schritt die Beobachtungswerte zu gewinnen.

Je nach der Art des zu untersuchenden Prozesses oder der zu untersuchenden Erscheinung müssen die Beobachtungswerte durch Zählen oder Messen gewonnen werden (vgl. Abb. 2).

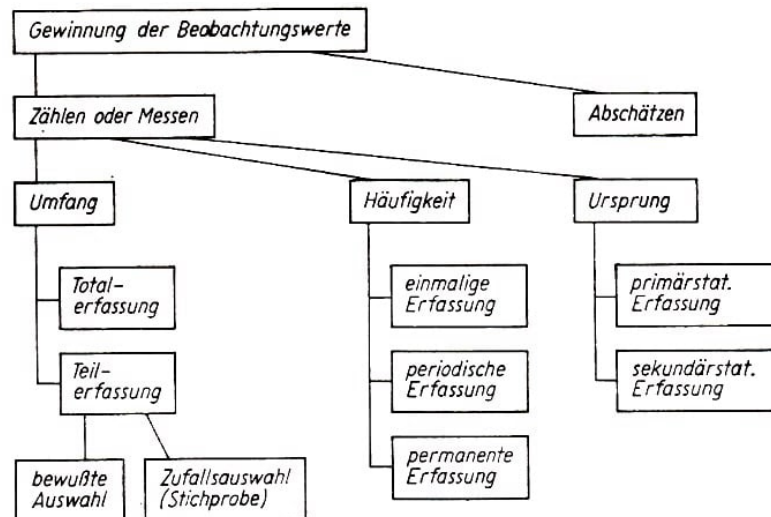


Abb.2 Arten der Gewinnung des Beobachtungsmaterials

Die Beobachtungswerte liegen zunächst in der Reihenfolge ihrer Gewinnung in Form einer Urliste vor, die weiter bearbeitet werden muss, damit die Daten später mit Hilfe mathematisch- statistischer Methoden ausgewertet werden können. Wir bezeichnen diesen Schritt als Aufbereitung der Beobachtungswerte, wobei es vordringlich um eine Verdichtung der in der Urliste enthaltenen Daten geht.

Denn nur, wenn die Zähl- oder Messwerte zu für die Erscheinung charakteristischen Klassen zusammengefasst werden, wird es möglich, das Typische der zu untersuchenden Erscheinung sichtbar zu machen.

Wir müssen jedoch beachten, dass einmal gebildete Klassen nicht ohne Rückgriff auf die Daten der Urliste wieder in kleinere Einheiten aufgegliedert werden können. Die erste Zusammenfassung soll sich deshalb immer auf die für die Auswertung erforderlichen kleinsten Klassen erstrecken, wobei sicherzustellen ist, dass durch Summation dieser kleinsten Klassen die für die Untersuchung notwendigen nächstgrößeren Klassen gebildet werden können.

Die einzelnen Beobachtungswerte werden mit dem Ziel der Aufbereitung je nach dem Umfang des statistischen Materials

- auf manuellem Wege
- durch den Einsatz der konventionellen Lochkartentechnik oder
- mit Hilfe von elektronischen Datenverarbeitungsanlagen

zusammengefasst. Entscheidend für die Wahl des jeweiligen Verfahrens muss der Vergleich zwischen Aufwand und Nutzen sein, doch können auch andere Kriterien, z. B. die für die Aufbereitung zur Verfügung stehende Zeit oder die vorhandenen Kapazitäten für die maschinelle Aufbereitung, die Wahl des Verfahrens beeinflussen.

In der Auswertung sind schließlich die gefundenen empirischen Verteilungen oder zeitlichen statistischen Reihen - entsprechend dem in der Vorbereitungsstufe formulierten Ziel - mit den geeigneten mathematisch-statistischen Methoden zu bearbeiten.

In der Analyse selbst werden die wesentlichsten Ergebnisse in Tabellen und Grafiken zusammengestellt und entsprechend der Zielstellung der Untersuchung verbal beurteilt. Wir sollen uns aber davor hüten, wie zuweilen noch anzutreffen, die Analyse allein als eine Beschreibung der Tabellen und Ergebnisse aufzufassen.

Vielmehr wird von den Bearbeitern erwartet, dass sie bereits bekanntes Wissen bestätigen oder neue Erkenntnisse gewinnen und entsprechende Schlussfolgerungen aus den Ergebnissen der Untersuchung zu ziehen in der Lage sind.

## 4 Empirische Verteilungen und ihre Auswertung

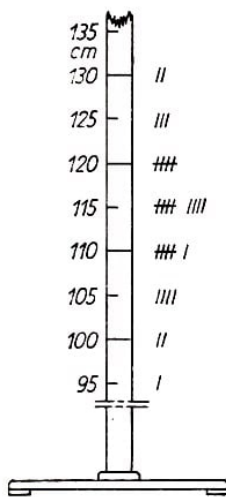
Wir haben aus den vorangegangenen Darlegungen schon erkannt, dass mit Hilfe der statistischen Methoden Massenerscheinungen analysiert werden. Über die Struktur eines statistischen Untersuchungsbereichs kann z. B. bei einem diskreten Merkmal schon etwas ausgesagt werden, wenn angegeben wird, wie oft eine Variation auftritt bzw. welcher absolute oder prozentuale Anteil auf jede Variation des betreffenden Merkmals entfällt.

Nehmen wir als Beispiel eine Gruppe von 32 Schülern, die sich im Hochsprung üben. Wir legen die Latte zunächst bei 90 cm auf, so dass alle Beteiligten sie überspringen können. Die Höhe von 95 cm wird ebenfalls noch von allen geschafft. In beiden Fällen sind somit 0 Schüler ausgeschieden.

Bei 100 cm kommt ein Schüler nicht mehr über die Latte. Bei 105 cm haben zwei weitere Schüler, bei 140 cm vier, bei 115 cm sechs, bei 120 cm neun, bei 125 cm fünf, bei 130 cm drei und bei 135 cm die letzten beiden Schüler "gerissen".

Wird die Anzahl der Schüler, die diese Höhe schafften (der jeweils niedrigere Wert), an einer gezeichneten Hochsprunglatte vermerkt, so erhalten wir die Abbildung 3.<sup>10</sup> Eine derartige Aufschreibung über die Häufigkeit des Erreichens einer bestimmten Höhe beim Hochsprung ist aus statistischer Sicht eine Häufigkeitsverteilung, die auch tabellarisch oder als Säulendiagramm dargestellt werden könnte.

Abb. 3 Häufigkeitsverteilung beim Hochsprung von Schülern



Durch die Häufigkeitsverteilung werden die Leistungen bestimmter Gruppen von Schülern dieser Klasse im Hochsprung charakterisiert. Über die Leistung des gesamten Klassenkollektivs kann etwas gesagt werden, wenn alle einzelnen Leistungen der 32 Schüler aufgeschrieben, die Werte zusammengezählt und die Summe durch die Zahl der Schüler (32) geteilt wird. Dieser Mittelwert kennzeichnet eben die mittlere Leistung dieses Klassenkollektivs. Im Abschnitt 4.3. kommen wir nochmals auf diesen Mittelwert zurück. Um aber ein exaktes Bild zu erhalten, ist es zweckmäßig, außer Mittelwerten weitere statistische Maßzahlen zu verwenden, z.B. die Abweichung der einzelnen Werte vom errechneten Mittelwert. Der Abschnitt 4.4. ist dieser Streuung der Einzelwerte gewidmet.

Insgesamt wird sich dieses 4. Kapitel unseres Bändchens mit solchen empirischen, aus Zählungen oder Messungen gewonnenen Häufigkeitsverteilungen beschäftigen - also mit empirischen Verteilungen im Unterschied zu den theoretischen Verteilungen, die im Kapitel 6 behandelt werden.

<sup>10</sup>Vgl. Stemmler, Becher, Reichstein und Steglich: Statistische Methoden im Sport. Sportverlag, Berlin 1973, S. 14

## 4.1 Flache, steile, schiefe und mehrgipflige Verteilungen

Zunächst wollen wir uns mit einigen Hilfsmitteln zur Untersuchung empirischer Verteilungen vertraut machen, vor allem mit der Urliste, der primären und der sekundären Häufigkeitstabelle. In diesem Zusammenhang werden wir auch die Bildung von Klassen und darauf aufbauend die Ermittlung der Summenhäufigkeit kennenlernen. Anschließend betrachten wir noch einige Formen empirischer Verteilungen, wie sie aus der grafischen Darstellung ersichtlich werden.

5	20	7	9	17	10	16	8	12	13	11	16	18
18	15	16	15	11	13	11	16	15	8	9	17	7
11	8	18	10	13	8	17	13	19	12	10	11	10
15	9	13	9	14	18	13	11	13	14	16	20	17
18	13	14	8	13	7	15	9	11	17	12	10	11
10	15	11	17	16	14	12	6	14	16	7	15	14
11	18	14	12	18	11	13	12	9	11	9	12	8
7	9	17	14	15	12	16	17	7	10	8	14	12
11	12	19	13	13	14	15	12	16	12	16	6	14
8	14	10	7	18	12	17	9					

Tabelle 1: Urliste

Das für diesen Abschnitt ausgewählte und durchgerechnete Beispiel bezieht sich auf die Auswertung von Messergebnissen, die im Verlaufe des Herstellungsprozesses eines Loses von Einzelteilen in einem feinmechanischen Betrieb gewonnen wurden.

Unter einem Los - auch als Produktionslos oder Partie bezeichnet - wird hierbei eine Anzahl konstruktiv gleichartiger Einzelteile verstanden, die in einem Arbeitsgang in zusammenhängender Folge unter Gewährung einer einmaligen Vorbereitungs- und Abschlusszeit gefertigt werden.

Die Messwerte stellen Abweichungen vom Nennmaß dar. Insgesamt sind während eines bestimmten Zeitabschnittes 125 verschiedene, voneinander unabhängige Messungen an Werkstücken gleicher Art vorgenommen und jeweils die Abweichungen vom Nennmaß ermittelt worden.

Die Ergebnisse dieser statistischen Untersuchungen befinden sich in der Tabelle 1, die als Urliste bezeichnet wird.

Die Angaben beziehen sich auf Mikrometer ( $\mu\text{m}$ ). In unserem Beispiel ergeben sich die Abweichungen in den Abmessungen der Werkstücke zufällig. Dabei hat jedes Werkstück eines Loses die gleiche Chance, in die Auswahl zu gelangen. Die Stichprobe mit den 125 zufällig ausgewählten Werkstücken umfasst also nur den Bruchteil eines solchen Loses. Die einzelnen Elemente, die gemessenen Abweichungen oder Messwerte können wir mit  $x_1, x_2, \dots, x_n$  bezeichnen.

Zunächst werden folgende Symbole eingeführt:

$x_i$ :  $i$ -ter Einzelwert der Zufallsvariablen  $X$ , z. B. die Messwerte  $x_1, x_2, \dots, x_n$

$n$ : Anzahl der Messwerte

$e_m$ : Klassennummer der  $m$ -ten Klasse einer empirischen Häufigkeitsverteilung ( $m = 1, 2, \dots, k$ )

$e_{mu}$ : untere Klassengrenze der  $m$ -ten Klasse einer empirischen Verteilung

$e_{mo}$ : obere Klassengrenze der  $m$ -ten Klasse einer empirischen Verteilung

$h_m$ : absolute Häufigkeit der  $m$ -ten Klasse einer empirischen Verteilung

$f_m$ : relative Häufigkeit der  $m$ -ten Klasse einer empirischen Verteilung

$SP$ : relative Summenhäufigkeit für eine empirische Verteilung in Prozent

$u_m$ : Klassenmitte der  $m$ -ten Klasse einer empirischen Verteilung

$k$ : Klassenanzahl einer empirischen Häufigkeitsverteilung

$d$ : Klassenbreite einer empirischen Verteilung

$x_{\max}$ : größter Einzelwert einer Messreihe

$x_{\min}$ : kleinster Einzelwert einer Messreihe

Die Urliste (Tab. 1) ist auf Grund der großen Anzahl nicht sortierter Werte unübersichtlich. Es ist deshalb zweckmäßig, dieses Zahlenmaterial zunächst der Größe nach zu ordnen. In unserem Beispiel verläuft die Messwertreihe von 5 bis 20  $\mu\text{m}$ .

Diese Werte schreiben wir in der Tabelle 2 in die Spalte 2. In der folgenden Spalte wird jeder Wert entsprechend seiner Größe durch einen Strich notiert und damit eine sogenannte Strichliste angelegt. Aus dieser kann für die Spalte 4 die Anzahl der Striche für jeden Messwert, d. h. seine absolute Häufigkeit, schnell ermittelt werden.

Auf diesem Wege entsteht aus der Tabelle 1 die Tabelle 2, die auch als primäre Verteilungstafel bezeichnet wird. Bereits jetzt ist zu erkennen, dass die Mehrzahl der gemessenen Werte zwischen 8 und 17  $\mu\text{m}$  liegt. Die absolute Häufigkeit der Messwerte steigt von einem Messwert mit 5  $\mu\text{m}$  bis zu je 13 Werten mit 11 und 12  $\mu\text{m}$  an. Sie fällt danach wieder ab, um schließlich bei 2 Messwerten von je 19  $\mu\text{m}$  anzugelangen. Die primäre Verteilungstafel gibt eine Häufigkeitsverteilung wieder, die im DDR-Standard "Statistische Qualitätskontrolle - Begriffe" als "Zuordnung der Häufigkeiten der Merkmalswerte zu den innerhalb des Wertebereiches eingeordneten Merkmalswerten<sup>11</sup>" erläutert wird.

Werden die Häufigkeiten der Messwerte grafisch dargestellt, so entstehen die Abbildungen 4 und 5. Die Abbildung 4 zeigt ein Histogramm oder Staffelpild für die Messwerte  $x_i$ . Es entsteht, indem über jedem Wert oder über jeder Klasse auf der Abszissenachse ein Rechteck errichtet wird.

Die Fläche oder die Höhe dieser Rechtecke sind entweder der absoluten bzw. der relativen Häufigkeit der Messwerte oder der Klassenhäufigkeit proportional.

Dem Staffelpild für die absoluten Häufigkeiten entspricht das Häufigkeitspolygon in der Abbildung 5. Es entstand, indem die Häufigkeiten der einzelnen auf der Abszisse aufgetragenen Messwerte durch Geraden miteinander verbunden wurden. Dieses Häufigkeitspolygon hat mehrere Gipfel.

---

<sup>11</sup>DDR-Standard, "Statistische Qualitätskontrolle - Begriffe". TGL 14449, Blatt 2, S. 6



Lfd. Nr. $i$	Messwert $x_i$ ( $\mu\text{m}$ )	Strichliste	Absolute Häufigkeit $h_i$
1	2	3	4
1	5	I	1
2	6	II	2
3	7	III II	7
4	8	III III	8
5	9	III IIII	9
6	10	III III	8
7	14	III III III	13
8	12	III III III	13
9	13	III III II	12
10	14	III III II	12
4	15	III IIII	9
12	16	III III	10
13	17	III IIII	9
14	18	III III	8
15	19	II	2
16	20	II	2
Umfang der Stichprobe			$n = 125$

Tabelle 2: Primäre Verteilungstafel

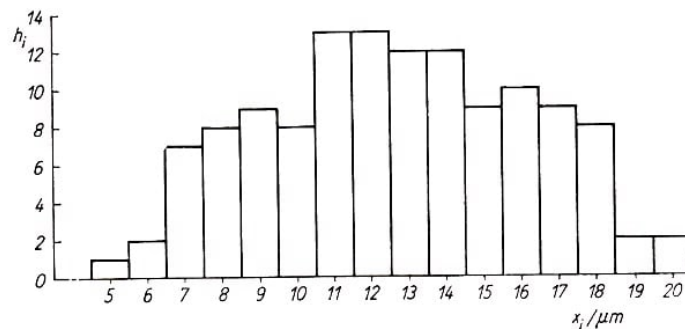


Abb. 4 Histogramm oder Staffelfeld - absolute Häufigkeiten

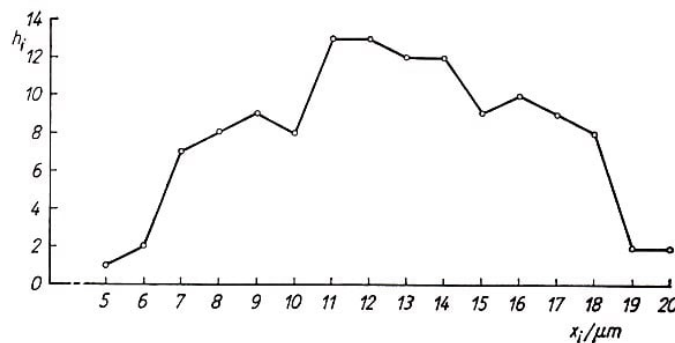


Abb. 5 Häufigkeitspolygon - absolute Häufigkeiten

Obwohl die in der Tabelle 2 vorgenommene Gruppierung schon eine bessere Übersicht als die Tabelle 4 gewährleistet, ist es insbesondere für die weiteren Berechnungen etwas umständlich, wenn mit 16 verschiedenen Werten gearbeitet werden soll. Deshalb wird - von der primären Verteilungstafel ausgehend - die empirische Verteilung weiter gestrafft, indem jeweils zwei unmittelbar nebeneinander liegende Messwerte zu

einer Klasse zusammengefasst werden. Bei einer Klasse handelt es sich - wie aus der Tabelle 3 deutlich zu erkennen ist - um mehrere aufeinanderfolgende Messwerte einer Messreihe.

Klassen-Nr.	Klassengrenzen untere/obere	Klassen- mitte	absolute Häufigkeit	Produkt aus Spalte 3 und 4	relative Häufigkeit	relative Summen- häufigkeit
$e_m$	$e_{mu}, e_{mo}$	$u_m$	$h_m$	$u_m \cdot h_m$	$f_m$ in %	SP $\sum$ in %
1	2	3	4	5	6	7
1	4,5...6,5	5,5	3	16,5	2,4	2,4
2	6,5...8,5	7,5	15	112,5	12,0	14,4
3	8,5...10,5	9,5	17	161,5	13,6	28,0
4	10,5...12,5	11,5	26	299,0	20,8	48,8
5	12,5...14,5	13,5	24	324,0	19,2	68,0
6	14,5...16,5	15,5	19	294,5	15,2	83,2
7	16,5...18,5	17,5	17	297,5	13,6	96,8
8	18,5...20,5	19,5	4	78,0	3,2	100,0
		$n = \sum_{m=1}^k h_m$ = 125	1583,5	100,0 %		

Tabelle 3 Sekundäre Verteilungstafel (Häufigkeitstabelle)

Unter der Klasseneinteilung wird in der statistischen Qualitätskontrolle die sinnvolle Unterteilung des Wertebereiches in Intervalle verstanden, in die die Werte eindeutig eingeordnet werden können.

Oft schafft man gleich große Klassen, d. h. mit gleicher Breite, z. B. mit Abständen von 2 oder 3 oder 4 oder ... 10, 12, ..., 20, 50, 100, 500, 1000 Einheiten.

Bei biologischen Messreihen wird die Größengleichheit der Klassen in der Regel sogar gefordert, um die quantitative Vergleichbarkeit auf jeden Fall zu gewährleisten. Diese Forderung wird meist auch für naturwissenschaftliche oder technische Messreihen erhoben.

In anderen Fällen, wie in der Ökonomie, müssen häufig unterschiedlich große Klassen gebildet werden, z. B. um verschiedene qualitative Aspekte zum Ausdruck zu bringen. Bei der Klassenbildung wird für jede Klasse eine untere und eine obere Grenze angegeben.

Sie sind so festzulegen, dass eine eindeutige Zuordnung der Messwerte zu einer Klasse ermöglicht wird. Deshalb legen wir im Beispiel die Klassengrenzen um 0,5 unter- bzw. oberhalb der gemessenen Werte fest. Damit wird gewährleistet, dass kein Messwert genau mit einer Klassengrenze identisch ist.

In anderen der Praxis entlehnten Beispielen können Häufigkeiten auch auf Klassengrenzen fallen. Diese Häufigkeiten werden zur Hälfte der einen und zur Hälfte der anderen Klasse zugerechnet. Allerdings erhalten wir dadurch unter Umständen gebrochene Häufigkeitszahlen, die sich bei der Ermittlung weiterer statistischer Maßzahlen nachteilig im Hinblick auf den Rechenaufwand auswirken können.

Deshalb wird in der Regel versucht, ohne gebrochene Häufigkeitszahlen auszukommen, indem - wie in unserem Beispiel - jeweils zwischen den Messwerten liegende Werte als Klassengrenzen festgelegt werden.

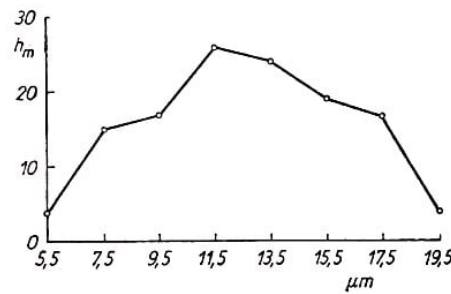


Abb. 6 Häufigkeitspolygon - absolute Häufigkeiten bei Verwendung von Klassen

Jetzt ist aus der primären Verteilungstafel (Tab. 2) eine sekundäre Verteilungstafel (Tab. 3), auch oft als Häufigkeitstabelle bezeichnet, entstanden.

Sie vermittelt einen Überblick darüber, wie die 125 Messwerte  $x_1, x_2, \dots, x_{125}$  (zusammengefasst in Klassen) über den gesamten Wertebereich verteilt sind, also über die Häufigkeitsverteilung. Dieser Tabelle entspricht die Abbildung 6. Sie zeigt das Häufigkeitspolygon (Häufigkeitsdiagramm) für die hier behandelte empirische Verteilung auf der Grundlage der gebildeten Klassen.

Wie ersichtlich ist, befindet sich ein großer Teil der Messergebnisse in den mittleren Klassen. Nach den Rändern zu verringern sich die Häufigkeiten. Prinzipiell sollte bei der Wahl der Klassenzahl beachtet werden, dass sich diese in der Größenordnung zwischen 6 und 20 bewegt.

Bei weniger als 6 Klassen werden die Berechnungen zu ungenau, und bei mehr als 20 Klassen entsteht ein relativ großer Arbeitsaufwand. Die Anzahl der Klassen ( $k$ ) und ihre Breite ( $d$ ) kann für die empirischen Verteilungen nach folgenden Formeln ermittelt werden.

Zur Abschätzung der Anzahl an Klassen können verwendet werden:

$$k \leq 5 \cdot \lg n \quad (2)$$

Hierbei ergibt sich:

$n$	50	100	500	1000	10000
$k$	8	10	13	15	20

Man kann ebenso

$$k \approx \sqrt{n} \quad (\text{bei } k \text{ zwischen } 5 \dots 25) \quad (3)$$

benutzen.

Im Beispiel ergibt sich für die Anzahl der Klassen:

$$k \leq 5 \cdot \lg 125, \quad k \leq 5 \cdot 2,095, \quad k \leq 10$$

Da die Beziehung kleiner oder gleich 10 lautet und sich auf Grund der Verteilungsbreite von 16 eine Einteilung in Zweierklassen, also in 8 Klassen, anbietet, wird diese gewählt. Eine sinnvolle und für die Berechnung zweckmäßige Einteilung der Klassen kann unter Berücksichtigung der Klassenbreite im vorliegenden Fall mit 10 Klassen nicht erreicht werden.

Die Klassenbreite wird angenähert ermittelt als

$$d \approx \frac{x_{\max} - x_{\min}}{k} \quad (4)$$

Für unser Beispiel bedeutet das:

$$d = \frac{20 - 5}{8} = \frac{15}{8} = 1,88$$

Die dem errechneten Wert von 1,88 unter Berücksichtigung sinnvoller Abstände zwischen den Klassen entsprechende günstigste Klassenbreite beträgt 2. Bei der Wahl der Klassenbreite, d. h. der Differenz zwischen der unteren und der folgenden oberen Klassengrenze, ist neben der Spannweite (vgl. Abschnitt 4.4.), der Anzahl an Messungen  $n$ , vor allem auch der Zweck, dem die statistische Untersuchung dient, mit zu beachten.

Die Klassenmitten sind jeweils der Durchschnitt aus dem Betrag der unteren und oberen Klassengrenzen, d.h.

$$u_m = \frac{e_{mu} + e_{mo}}{2} \quad (6)$$

Das Zusammenfassen der Messwerte in Klassen hat zweifellos Auswirkungen auf die grafische Darstellung der empirischen Verteilung, z. B. kann eine zunächst vorhandene Symmetrie sich beim Auftragen der Messwerte unter Verwendung der Klassen in eine Asymmetrie verwandeln.

Die Abbildungen 5 und 6 zeigen für das gewählte Beispiel den Unterschied in der grafischen Darstellung bei der Verwendung der Häufigkeiten für die einzelnen Messwerte bzw. der Häufigkeiten der Klassen. Auf die spätere Berechnung des Mittelwertes oder der Streuung wirkt sich die Zusammenfassung in Klassen jedoch nur geringfügig aus.

Aus der sekundären Verteilungstafel (Tab. 3) sind die erforderlichen Angaben für die Klassen (Klassennummern, -grenzen und -mitten) für das Beispiel ersichtlich. Die absoluten Häufigkeiten für die Klassen wurden aus der Tabelle 2 ermittelt.

Weiterhin sind in der Tabelle 3 die relative Häufigkeit für die einzelnen Klassen ( $f_m$ ) in Prozent und darauf aufbauend die relative Summenhäufigkeit ( $SP$ ) berechnet worden. Die Summe der relativen Häufigkeit beträgt 1. Demzufolge kann  $f_m$  als Stellenwert von 1 oder unter der Voraussetzung, dass 1 gleich 100 gesetzt wird, auch in Prozent angegeben werden.

$$f_m = \frac{h_m}{n} \quad (6)$$

Wird das Ergebnis in Prozent gewünscht, so ist die Gleichung mit 100 zu multiplizieren. Die relative Summenhäufigkeit ergibt sich daraus durch Aufsummierung der Häufigkeiten der einzelnen Klassen mit

$$SP = \frac{h_1 + h_2 + \dots + h_k}{n} \cdot 100\% = \frac{\sum_{m=1}^k h_m}{n} \cdot 100\% \quad (7)$$

(Zum Rechnen mit dem Summenzeichen vgl. a. Abschnitt 4.3.!)

In diesen Formeln wurde im Nenner  $n$  eingesetzt, d.h. die Gesamtzahl der gemessenen Werte, obwohl die Bezugsbasis eng genommen die  $\sum_{m=1}^k h_m$  ist. Das ist jedoch zulässig, da hier  $\sum_{m=1}^k h_m = n$  ist.

Die Angaben in der Spalte 5 der Tabelle 3 dienen bereits bestimmten Vorbereitungen für weitere Untersuchungen an diesem Beispiel in einem folgenden Abschnitt.

Die grafische Darstellung der Summenhäufigkeiten ergibt eine Treppenkurve gemäß der Abbildung 7. Sie stellt über den Klassen die relativen Summenhäufigkeiten dar, wobei in diesem Falle auf der Abszissenachse die Klassengrenzen angegeben wurden. Sie bringt auch entsprechend der zugrunde liegenden empirischen Verteilung eine Annäherung an die Verteilungsfunktion der Zufallsgröße  $X$  zum Ausdruck.

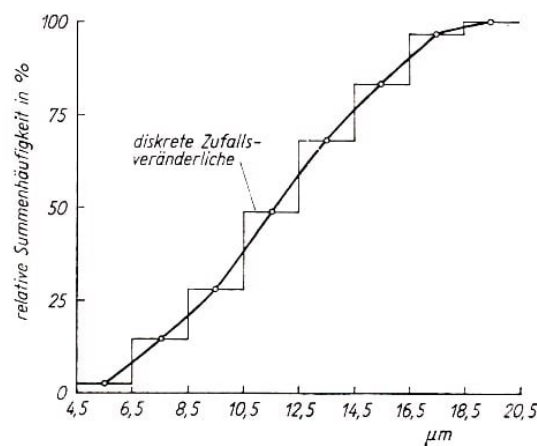
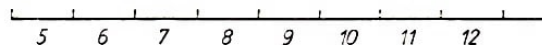


Abb.7 Treppenkurve der Summenhäufigkeiten

Besonders deutlich wird das, wenn man die Klassenmitten mit einem Polygonzug verbindet.

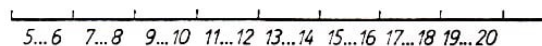
Aus den bisherigen Ausführungen ergab sich bereits, dass die Abszissenachse bei der grafischen Darstellung der Häufigkeitsverteilungen mit unterschiedlichen Bezeichnungen versehen werden kann.

Einzelwerte:

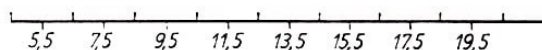


In Klassen zusammengefallte Werte:

Klasse:



Klassenmitten:



Klassengrenzen:

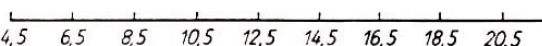


Abb.8 Abszissenbezeichnung für Häufigkeitsverteilungen

Die Abbildung 8 gibt hierüber unter Bezugnahme auf das erörterte Beispiel zusammenfassend Auskunft.

Unser Beispiel hat erkennen lassen, dass Untersuchungsergebnisse aufbereitet und ausgewertet werden können, wenn die einzelnen Mess- oder Beobachtungsergebnisse über verschiedene Werte oder Klassen verteilt sind. Die Messergebnisse  $x_1, x_2, \dots, x_n$  sind reelle Zahlen. Diese untersuchten Werte verkörpern eine Zufallsvariable, für die wir symbolisch  $X$  setzen (vgl. Kap. 2). Die  $x_1, x_2, \dots, x_n$  sind dann die Realisationen von  $X$ .

Die empirische Verteilung ist die Verteilung der in der Praxis tatsächlich gemessenen oder gezählten Werte, der Realisationen der Zufallsvariablen  $X$ . In den Abbildungen 4 bis 7 wurden die unserem Beispiel entsprechenden Verteilungen grafisch dargestellt. Der Statistiker unterscheidet zwischen derartigen empirischen Verteilungen und den theoretischen Verteilungen, auf die im 6. Kapitel näher eingegangen werden soll.

Empirische Verteilungen können in verschiedenen Formen auftreten. Sie sind anhand der grafischen Darstellungen deutlich zu unterscheiden. Im Prinzip sind diese Formen aber ebenso aus der Aufschreibung der sortierten Beobachtungswerte zu ersehen. Je nachdem, welches Bild sich ergibt, spricht man von flachen, steilen, schiefen, ein-, zwei- oder mehrgipfligen empirischen Verteilungen. Zahlreiche empirische Verteilungen deuten eine Glockenform an, wie das z.B. aus den Häufigkeitspolygonen in der Abbildung 9 zu erkennen ist.

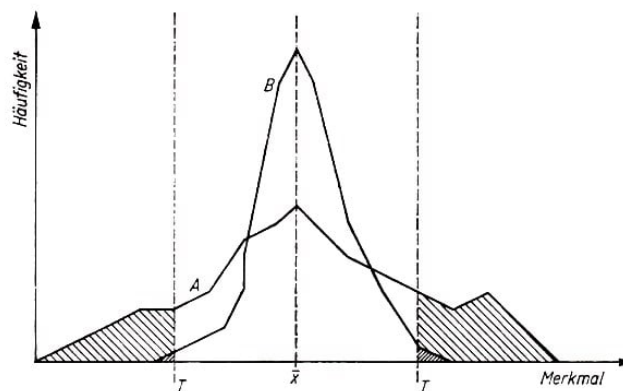


Abb.9 Häufigkeitspolygone, die der Glockenform ähnlich sind - Ergebnisse von Qualitätskontrollen

Sie spiegelt Häufigkeitsverteilungen aus zwei vergleichbaren statistischen Qualitätskontrollen bei der Herstellung elektronischer Bauelemente wider. In diesen Fällen verteilen sich die Mess- bzw. Beobachtungswerte symmetrisch um den Mittelwert. Die Verteilung B lässt erkennen, dass die extremen Messwerte verhältnismäßig wenig, die mittleren jedoch häufig vorkommen.

Die geringere Ballung ähnlicher Werte bei der Verteilung A zeigt sich an dem flachen Kurvenverlauf.

Je nach der Streuung der Messwerte um den Mittelwert ist der Anstieg in einer Verteilung steiler oder flacher. Bei der Verteilung A in der Abbildung 9 ist eine breite Streuung und demzufolge ein flacherer Anstieg als bei der Verteilung B mit ihrem relativ steilen Anstieg zu erkennen. Eine Übertreibung gegenüber der Glockenform wird als Exzess bezeichnet.

Als Vergleichsmaßstab dient dabei die Normalverteilung, auf die im Abschnitt 6.4. eingegangen wird. Ein positiver Exzess liegt vor, wenn eine empirische Verteilung steiler ist als die Normalverteilung (s. Abb. 9, Verteilung B). Von einem negativen Exzess spricht man, wenn die Verteilung flacher ist (s. Abb. 9, Verteilung A).



Abb. 10 Linksschiefe Verteilung— Messergebnisse von Bohrungen

In anderen Fällen liegt der Gipfel der Verteilung rechts oder links von der Mitte, wie in den Abbildungen 10 und 11. Im ersten Fall handelt es sich um die Häufigkeiten der Abweichungen vom Nennmaß bei der Messung von Bohrungen in einem Maschinenbaubetrieb.

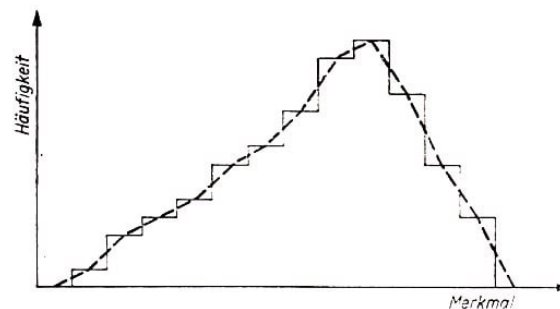


Abb. 11 Rechtsschiefe Verteilung — Fehlerhäufigkeiten in elektronischen Geräten

Beim zweiten Beispiel geht es um eine Verteilung von Fehlerhäufigkeiten in elektronischen Geräten. In beiden Beispielen verteilen sich die Werte ungleichmäßig, also asymmetrisch um den Mittelwert. Derartige Verteilungen werden als schief charakterisiert.

Bei einer positiven Schiefe ist die Mehrzahl der Werte auf der linken Seite über der Abszisse konzentriert (s. Abb. 10). Das gilt unter der Voraussetzung, dass auf der Abszisse links die kleineren Werte aufgetragen wurden. Eine negative Schiefe liegt vor, wenn sich rechts, wo die höheren Werte stehen, die Mehrzahl der Mess- bzw. Beobachtungsdaten befindet (s. Abb. 11).

In der Abbildung 11 wird, da das Histogramm und das Häufigkeitspolygon einer empirischen Verteilung übereinander gelegt wurden, noch ein weiterer Sachverhalt deutlich. Einerseits werden vom Histogramm Dreiecke abgetrennt, andererseits wird eine bestimmte Fläche außerhalb des Histogramms hinzugefügt. In jedem Falle stimmen der abgetrennte und der hinzugefügte Teil überein.

Eine ganz extreme Schiefe ist festzustellen, wenn die Verteilung einseitig begrenzt wird. Das trifft z. B. zu, wenn der Wert 0 nicht überschritten werden darf.

Der Anstieg der Verteilung, der sich gegebenenfalls im negativen Zahlenbereich bewegen würde, fällt in einem solchen Falle weg. Man spricht deshalb auch von einer einseitig abfallenden Verteilung.

Derartige Kurven bringen zum Ausdruck, dass die Werte eines Extrems innerhalb der Verteilung am häufigsten auftreten, während die Häufigkeit des anderen Extrems ein Minimum darstellt. Ein Beispiel hierfür ist die in Abbildung 12 enthaltene Verteilung, die sich bei einer technisch-ökonomischen Untersuchung ergeben hat. Sie zeigt die absolute Häufigkeit des Ausfalls von Maschinen in einem Betrieb.

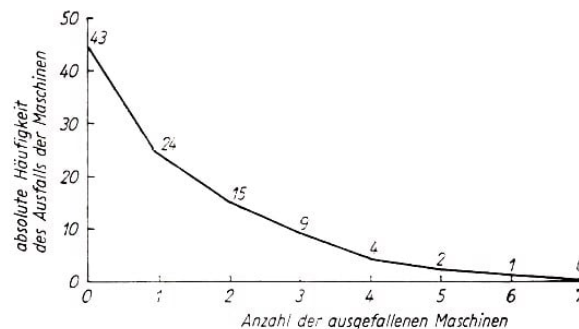


Abb.12 Von links nach rechts abfallende Verteilung - Ausfall von Maschinen

Als Beispiel für eine nach rechts ansteigende Verteilung dient die relative Häufigkeit der Wahl von Farbtönen für den Anstrich von Wohnzimmern nach dem Grad ihrer Beliebtheit, und zwar vom unbeliebten bis zum beliebtesten Farbton (vgl. Abb. 13).

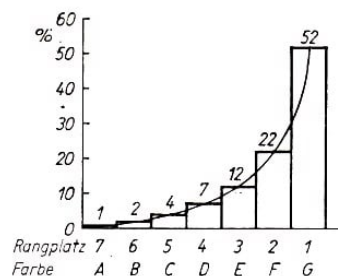


Abb. 13 Von links nach rechts ansteigende Verteilung - Rangplätze von Farben entsprechend ihrer Beliebtheit

Für dieses Beispiel ist typisch, dass keine Intervalleinteilung gewählt werden konnte, sondern eine Reihenfolge bzw. Rangfolge, die dem Grad der Beliebtheit entspricht. Es ist offensichtlich, dass bei einer entgegengesetzten Anordnung (Austausch der Farben, so dass A die Stelle von G in der Abbildung einnimmt usw.) eine von links nach rechts abfallende Verteilung entstehen würde.

Ein Beispiel für eine U-förmige Verteilung ist die Statistik der Grippesterblichkeit bezogen auf die Altersklassen der Bevölkerung. Da vorwiegend nur einige wenige Säuglinge und ältere Bürger an dieser Krankheit sterben, ergibt diese empirische Verteilung etwa das in der Abbildung 14 angedeutete Bild. U-förmige Verteilungen sind dadurch gekennzeichnet, dass - im Gegensatz zur glockenförmigen Verteilung - die extremen Messwerte häufig und die mittleren Werte verhältnismäßig selten auftreten. Die U-förmige Verteilung ist oft zur Verteilungsmitte symmetrisch. Sie ist zweigipflig.



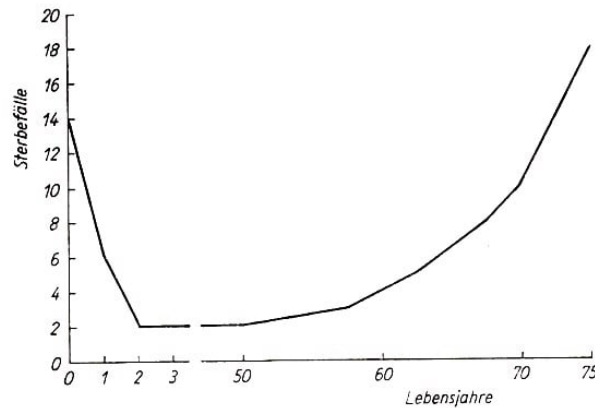


Abb. 14 U-förmige Verteilung - Sterbefälle an Grippe

Die grafische Darstellung einer empirischen Verteilung zeigt meist das Bild eines Berges oder eines Gebirges. Hiervon ausgehend hat es sich in der Statistik eingebürgert, von eingipfligen, zweigipfligen und mehrgipfligen Verteilungen zu sprechen.

Oft handelt es sich hierbei um die Überlagerung mehrerer Verteilungen. Bei der Abbildung 15 könnte z. B. angenommen werden, dass es sich um einen Fall handelt, bei dem sich zwei Normalverteilungen überlagern. Wir wissen es aber nicht genau und könnten das nur mit Hilfe eines Tests (vgl. Kap. 8) überprüfen.

Im Beispiel handelt es sich um die Verteilung für einen Qualitätsparameter von Thyristoren. Diese Produkte wurden nach zwei etwas voneinander abweichenden technologischen Varianten hergestellt, jedoch gemeinsam geprüft.

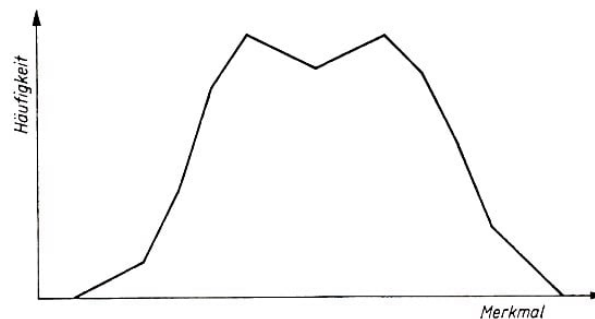


Abb. 15 Zweigipflige Verteilung - Verteilung für einen Qualitätsparameter

Eine empirische Verteilung mit mehreren Gipfeln wird mit der Abbildung 16 gezeigt. Es handelt sich hierbei um eine Häufigkeitsverteilung für Fehler in der Endfertigung elektronischer Geräte.

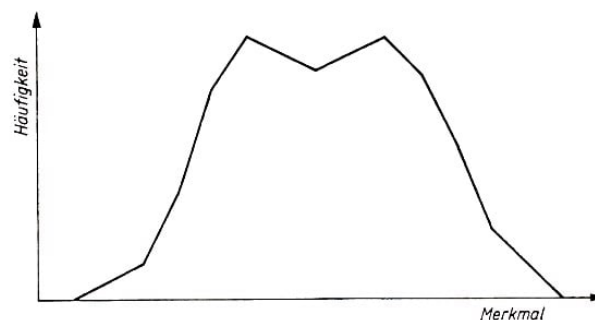


Abb. 16 Mehrgipflige Verteilung - Fehlerverteilung in der Endfertigung

Für bestimmte, insbesondere mehrgipflige Verteilungen ist es unter Umständen zweckmäßig, die Ausgangsdaten nach zusätzlichen Merkmalen zu überprüfen. Das gilt vor allem dann, wenn der Verdacht auf Mischverteilungen vorliegt, z. B. wenn vermutet wird, dass sich hinter einer zweigipfligen Verteilung in Wirklichkeit zwei homogene Verteilungen verbergen.

Eine homogene Verteilung ist in diesem Zusammenhang eine Verteilung, die durch einen einheitlichen Ursachenkomplex hervorgerufen worden ist.

In der Praxis treten Mischverteilungen verhältnismäßig häufig auf, u. a. bei der statistischen Qualitätskontrolle. Die Ursachenkomplexe können z. B. unterschiedliche Maschinen oder Maschinengruppen, verschiedenartige technologische Verfahren oder Varianten sein. Ein Beispiel zeigte die Abbildung 15.

Hier empfiehlt es sich, für die Erzeugnisse getrennt nach jeder Maschine oder Maschinengruppe oder den unterschiedlichen technologischen Verfahren bzw. Varianten die Häufigkeitsverteilung zu untersuchen. Ein formales Zerlegungsverfahren wird nur dann angewandt, wenn die sachliche Zusammengehörigkeit die Aufgliederung der Mischverteilung verbietet und die Annahme begründet ist, dass eine Mischverteilung vorhanden ist.

Ein solches Zerlegungsverfahren wurde z.B. von Daeves und Beckell<sup>12</sup> konzipiert. Hierbei wird die Zerlegung auf grafischem Wege angestrebt.

Liegen empirische Verteilungen vor, so ist es in der Regel notwendig, sie mit Hilfe von Maßzahlen auszuwerten, die jeweils bestimmte Erfahrungen der Verteilungen charakterisieren. Zu diesen Maßzahlen gehören Mittelwerte und Streuungsmaße.

Wir wollen uns deshalb hier noch einen Überblick über einige statistische Maßzahlen verschaffen.

In der Tabelle 4 wird zunächst eine Übersicht über einige statistische Maßzahlen gegeben, die zur Charakterisierung von Verteilungen verwendet werden können.<sup>13</sup> In diese Tabelle sind nur wenige Maßzahlen als Beispiele aufgenommen.

Auf einige sei lediglich hingewiesen und nicht näher eingegangen, da sie in diesem Buch nicht weiter verwendet werden. Die Maßzahlen werden nach zwei hauptsächlichen Kriterien unterteilt: einerseits nach der zu kennzeichnenden Eigenschaft, wobei Mittelwerte, Streuungsmaße und jene Maße, die einen Zusammenhang, eine Abhängigkeit charakterisieren, zu erwähnen sind; andererseits nach der Art der Merkmalsvariation, wobei zwischen nominalen, ordinalen und metrischen Merkmalsvariationen sowie entsprechenden Maßzahlen unterschieden wird (vgl. Kap. 2).

Einteilung statistischer Maßzahlen

---

<sup>12</sup>Daeves, K., u. Beckel, A.: Großzahl-Methodik und Häufigkeits-Analyse. 2. Aufl., Weinheim 1958, S. 143

<sup>13</sup>Vgl. Marinell, G.: Statistische Rezeptsammlung. R. Oldenbourg Verlag, München und Wien 1970, S. 15ff.

Nach der zu kennzeichnenden Eigenschaft	Nach der Art der Merkmalsvariation		
	Nominal	Ordinal	Metrisch
Mittelwerte	Häufigste Werte (z.B. Dichtemittel Anteilswerte)	Zentralwert Perzentile Quartile	Arithmetisches Mittel
Streuungsmaße	Anteilstreuung	Spannweite Perzentilabstand Quartilabstand	Standardabweichung Varianz
Zusammenhangs- oder Abhängigkeitsmaße	Kontingenzkoeffizient	Rangkorrelationskoeffizient	Maßhangskorrelationskoeffizient

Tabelle 4 Übersicht über einige statistische Maßzahlen

Die Art der Merkmalsvariation ist entscheidend dafür, ob wir metrische, ordinale oder nominale Mittelwerte, Streuungsmaße oder Maße für den Zusammenhang bzw. die Abhängigkeit ermitteln können. Es ist möglich, aus metrischen Verteilungen metrische oder ordinale oder nominale Maßzahlen zu berechnen. Liegen ordinale Verteilungen vor, können nur ordinale oder nominale Maßzahlen fixiert werden. Bei einer nominalen Verteilung ist es nur möglich, nominale Maßzahlen zu bestimmen.

Die Mittelwerte sollen das Niveau einer Verteilung durch einen einzigen zahlenmäßigen Ausdruck beschreiben. Jeder Mittelwert kennzeichnet auf Grund seines Bildungsprinzips einen anderen Tatbestand, ein anderes Charakteristikum einer Verteilung. In der Praxis werden dazu bei unterschiedlichen Fragestellungen unterschiedliche Mittelwerte verwendet.

Streuungsmaßzahlen erfassen im allgemeinen die Abweichungen der Einzelwerte von fixierten Werten. Man unterscheidet u.a. Maße des Abstands zwischen markanten Werten (z. B. die Spannweite) und Maße für die Abweichungen der Elemente einer Verteilung von einem die Verteilung charakterisierenden Wert (z. B. die Standardabweichung).

Die Maße zur Kennzeichnung eines Zusammenhangs bzw. einer Abhängigkeit zwischen zwei oder mehreren Verteilungen bzw. Zufallsveränderlichen werden ebenfalls unter Berücksichtigung der Art der Merkmalsvariation und demzufolge der Maßzahlen berechnet. Während z.B. der Zusammenhang zwischen zwei oder mehreren metrischen Verteilungen mit Hilfe des Maßkorrelationskoeffizienten bestimmt wird, dient der Rangkorrelationskoeffizient zur Ermittlung der Abhängigkeit zwischen ordinalen Größen.

Der Kontingenzkoeffizient gibt als Maßzahl Aufschluss über den Zusammenhang zwischen nominalen Verteilungen. Anhand von Beispielen wird im Kapitel 5 dieses Buches näher auf die Korrelation und die damit verbundene Regression eingegangen.

## 4.2 Zur Bildung von Gruppen oder Klassen

Bei empirischen Untersuchungen kommt es zweifellos auf die richtige Planung dieser statistischen Zählungen bzw. Messungen und der Auswertung und Analyse der Ergeb-

nisse an. Die sachlich einwandfreie Fragestellung und eine eindeutige Definition der benutzten Begriffe sind dafür notwendige Voraussetzungen. Ebenso wichtig ist in der Statistik die Auswahl der einzusetzenden Zähl- und Messverfahren sowie der anderen methodischen Hilfsmittel, die bei der Aufbereitung und Auswertung von Statistiken verwendet werden sollen.

Es war bereits darauf hingewiesen worden, dass dieses Instrumentarium dem Inhalt und Zweck der Untersuchung und dem jeweiligen Wissenschaftsgebiet, in dem es angewandt wird, untergeordnet werden muss. Das gilt auch für die Bildung von Gruppen oder Klassen.

Betrachten wir die einschlägige gesellschaftswissenschaftliche, naturwissenschaftliche, technische und mathematisch-statistische Literatur, so können wir verschiedene Auffassungen zur Gruppierung bzw. Klassifizierung von statistischen Elementen feststellen. Sie unterscheiden sich vor allem dadurch, dass die Begriffe "Gruppe", "Gruppierung", "Klasse" und "Klassifizierung" jeweils mit einem anderen Inhalt, mit einer anderen Definition bedacht werden. Dabei gibt es folgende wesentliche Varianten:

- Unter Gruppierung oder Gruppenbildung wird die Untergliederung eines Untersuchungsobjektes in Gruppen verstanden, die sich vor allem qualitativ unterscheiden. Die Gliederung von Betrieben nach Eigentumsformen (Volkseigentum, genossenschaftliches Eigentum, Privateigentum) ist ein Beispiel hierfür.

Der quantitative, zahlenmäßige Unterschied zwischen diesen Gruppen steht dabei zunächst im Hintergrund. Eine derartige, vorrangig qualitative Gruppierung ist z. B. bei ökonomischen Untersuchungen erforderlich, bevor andere methodische Hilfsmittel angewandt werden. Überhaupt ist die Gruppenbildung bei statistischen Untersuchungen gesellschaftswissenschaftlicher Erscheinungen und Prozesse von entscheidender Bedeutung. Sie stellt die wesentliche Voraussetzung für alle anderen statistischen Auswertungsarbeiten dar.

Demgegenüber wird unter Klassifizierung oder Klassenbildung häufig die nach der bereits erfolgten qualitativen Gruppierung vorzunehmende Einteilung vor allem nach quantitativen Aspekten, d.h. insbesondere die Einteilung in Größenklassen, verstanden.

Es ist z. B. möglich, die Gruppe der volkseigenen Betriebe in folgende Klassen aufzugliedern: Betriebe mit bis zu 100 Beschäftigten, von 101 bis 500, von 501 bis 1000, von 1001 bis 2000 und mit 2001 und mehr Beschäftigten.

Die Genossenschaften des Friseurhandwerks können z.B. eingeteilt werden in Klassen mit einer Breite von jeweils 10, d. h. also Genossenschaften mit 21 bis 30, mit 31 bis 40, mit 41 bis 50 Mitgliedern usw.

Bei der hier geschilderten Art des Herangehens bestehen also innerhalb der qualitativ sich unterscheidenden Gruppen mehrere Größenklassen, wobei im ersten Fall ungleich große und im zweiten gleich große Klassen gebildet wurden.

- Die Begriffe "Gruppe" und "Klasse" werden als synonym betrachtet und ebenfalls die Begriffe Gruppierung oder Gruppenbildung und Klassifizierung oder Klassenbildung als identisch verwendet. Aber auch bei einer Gleichsetzung dieser Begriffe bleibt die For-

derung bestehen, dass die Einteilung sowohl qualitativ als auch quantitativ sachgerecht sein und dem Untersuchungsgegenstand voll entsprechen muss.

In Veröffentlichungen auf dem Gebiet der mathematischen Statistik wird z. B. häufig nur von Klassen gesprochen und dabei sowohl quantitativen als auch qualitativen Gesichtspunkten Rechnung getragen.

Zu diesen beiden grundsätzlichen Möglichkeiten können weitere, mehr oder wenig logisch variierte Kombinationen hinzukommen. Es lohnt sich hier aber nicht, näher darauf einzugehen.

Entscheidend ist, dass eindeutig definiert wird, was im jeweiligen Anwendungsfall, z. B. unter "Gruppe", zu verstehen ist und wie die Einheit von qualitativen und quantitativen Aspekten gewährleistet wird.

Betrachten wir die Gruppenbildung noch etwas näher, und zwar vorwiegend unter dem Blickwinkel der Anwendung auf gesellschaftswissenschaftliche Gebiete. Beispielsweise trägt sie in der ökonomischen Statistik erheblich dazu bei, dass die Statistik ihre Aufgabe, wesentliche Seiten und Züge der untersuchten Massenerscheinungen widerzuspiegeln und die objektive Realität erkennen zu helfen, erfüllen kann.

Auf der Grundlage der gebildeten Gruppen werden dann jeweils bestimmte statistische Methoden angewendet. Das lässt zugleich einen relevanten Zusammenhang zwischen der Gruppenbildung und den statistischen Methoden sichtbar werden. Es ist deshalb nicht zu viel behauptet, wenn festgestellt wird, dass die Gruppenbildung ein zentrales Problem der Statistik ist.

Ausschlaggebend für die Bildung der Gruppen sind das Untersuchungsobjekt, die Anforderungen der verschiedenen Leitungsebenen oder Forschungsstellen sowie das unter Umständen schon vorhandene Zahlenmaterial. Die Klassiker des Marxismus-Leninismus bewiesen in ihren Arbeiten, dass die Gruppenbildung entscheidend ist, um u. a. das Ausbeutungsverhältnis im Kapitalismus auf der Grundlage des dialektischen und historischen Materialismus und der politischen Ökonomie analysieren zu können.

Die Gruppenbildung ist - wie Lenin hervorhob - notwendig, denn "... hinter Bergen von Zahlen verschwinden die ökonomischen Typen der Erscheinungen, Typen, die nur bei vielseitig und rationell zusammengestellten Gruppen- und Kombinationstabellen zum Vorschein kommen können".<sup>14</sup>

Die Klassiker des Marxismus-Leninismus mussten für ihre Analysen das Zahlenmaterial der bürgerlichen Statistiken neu aufbereiten, eben neue Gruppen bilden. Die bürgerlichen Statistiker sind - ausgehend von den Interessen der Klasse, der sie dienen - ebenso wie die herrschende kapitalistische Klasse selbst, nicht daran interessiert, die wahren sozialen Strukturen und Verhältnisse mit Hilfe statistischer Untersuchungen voll aufzudecken. Sie müssten sonst das Ausbeutungsverhältnis bloßlegen.

Da die Gruppierung das Wesentliche gesellschaftlicher Erscheinungen sichtbar macht, drückt sie mehr als alle anderen statistischen Methoden den Klassencharakter der Statistik, ihre Parteilichkeit bzw. Wahrheitstreue (für kapitalistische Verhältnisse besser -untreue) aus.

---

<sup>14</sup>Lenin, W. I.: Werke, Band 20, Dietz Verlag, Berlin 1961, S. 73

Im Kapitalismus wird die Statistik im Klasseninteresse zur Verschleierung bestehender gesellschaftlicher Verhältnisse verwandt. Im Sozialismus ist sie eine Grundlage für die Arbeit der Organe des sozialistischen Staates und der Wirtschaft. Sie gibt Impulse zur Gestaltung der gesellschaftlichen Beziehungen der Menschen in der Produktion und zur Beherrschung der Natur durch den Menschen.

Wird die Gruppenbildung nur nach einem Merkmal vorgenommen, kann man lediglich die Struktur des Untersuchungsobjektes ansehen. Bei Untergliederung nach mehreren Merkmalen führt die Gruppenbildung dazu,

- dass die Struktur der untersuchten Massenerscheinungen zu erkennen ist, weil die Untergliederung nach den Variationen hervorgehoben wird
- dass die Bedeutung der Gruppen sichtbar wird, weil die Gruppenbildung die Merkmalsgrößen anderer Merkmale für die einzelnen Gruppen hervortreten lässt
- dass wesentliche Beziehungen zwischen verschiedenen Seiten der untersuchten Massenerscheinungen deutlich werden, weil unterschiedliche Gruppierungsmerkmale miteinander verbunden werden können.

Bei der Gruppenbildung für statistische Analysen müssen

- die Merkmale für die Gruppierung genau definiert und
- ihre Variationen unter den für das Untersuchungsobjekt gegebenen konkreten Bedingungen ermittelt werden.

Die verschiedenen Merkmale sind nach sachlichen, örtlichen und zeitlichen Aspekten (vgl. Kap. 2) gegliedert, so dass in der Regel ebenfalls eine diesen Gesichtspunkten entsprechende Gruppenbildung vorgenommen werden kann.

### 4.3 Der Durchschnitt - ein charakteristischer Wert?

Jeder Erwachsene und fast jeder Schüler hat in seinem Leben schon irgendwann einmal mit Durchschnittswerten zu tun gehabt.

Das können z. B. die durchschnittliche Körpergröße, die Durchschnittsnoten der Schulklassen in bestimmten Fächern, die Durchschnittsausgaben eines Bürgers der DDR für Nahrungsmittel oder für Industriewaren, der durchschnittliche Verbrauch pro Einwohner eines Landes an Kartoffeln, Mehl, Fleisch und ähnlichen Erzeugnissen, der Durchschnittslohn, eventuell auch die unter dem Durchschnitt liegende Leistung oder andere durchschnittlichen Angaben gewesen sein.

Wird die Frage aufgeworfen, wie Durchschnittsgrößen berechnet werden, so verweist bestimmt der eine oder andere Leser auf das arithmetische Mittel. Diese Antwort kann richtig sein.

Sie muss es aber nicht, denn in der Statistik wird zwischen recht verschiedenartigen Mittelwerten unterschieden, z. B. dem arithmetischen, geometrischen, harmonischen, quadratischen Mittel, Integralmittel, Zentralwert, Dichtemittel.

Allerdings gehen gerade hinsichtlich der Verwendung des Begriffs "Durchschnitt" bei den Autoren statistischer Fachbücher die Meinungen auseinander.

Weiter können die Mittelwerte in Mittelwerte der Lage und in eigentliche Durchschnittswerte unterteilt werden. Die Mittelwerte der Lage sind dadurch charakterisiert, dass ein Element (oder eine Gruppe von Elementen) durch seine (ihre) Lage in einer geordneten Folge von Elementen den Mittelwert dieser Folge repräsentiert (dazu gehören u. a. der Zentralwert und das Dichtemittel).

Bei den eigentlichen Durchschnittswerten hingegen tragen alle Elemente der Folge zur Durchschnittsbildung bei (z. B. arithmetisches oder harmonisches Mittel).

Bei dem bereits eingangs dieses Kapitels erwähnten Beispiel über das Hochspringen kann das arithmetische Mittel sowohl Maßzahl der Lage für die empirische Verteilung als auch kennzeichnend für das mittlere Niveau, die Durchschnittsleistung dieser Klasse im Hochsprung sein.

Wir hatten schon festgestellt, dass die mittlere Leistung des gesamten Klassenkollektivs im Hochsprung dadurch berechnet werden kann, dass wir die von jedem Schüler erreichte Leistung aufschreiben, diese Einzelwerte zusammenzählen, also die Gesamtsumme bilden, und durch 32, d. h. durch die Gesamtzahl der Schüler in dieser Klasse, teilen.

Bezeichnen wir die Einzelleistung mit  $x$  und unterscheiden die einzelnen Schüler durch den Index  $i$  ( $i = 1, 2, \dots, n$ ), so können wir aus dieser Überlegung für das einfache arithmetische Mittel, das mit  $\bar{x}$  symbolisiert wird, folgende Formeln aufstellen:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (8)$$

Wir können aber diese Berechnung auch vereinfachen, indem wir die Anzahl der Schüler, die die gleiche Leistung zeigten, eben mit diesem Wert der erreichten Leistung multiplizieren und die sich ergebenden Produkte danach zu einer Summe zusammenfassen.

Auf diese Weise erhält jeder erreichte Wert ein bestimmtes Gewicht, das hier durch die Anzahl der Schüler, die die jeweilige Höhe sprangen, ausgedrückt wird. Dieses Gewicht soll mit  $g$  symbolisiert werden, so dass wir ausgehend von der Gleichung (8) jetzt schreiben können:

$$\bar{x} = \frac{a_1 g_1 + a_2 g_2 + \dots + a_m g_m}{g_1 + g_2 + \dots + g_m} = \frac{\sum_{j=1}^m a_j g_j}{\sum_{j=1}^m g_j} \quad (9)$$

Mit  $\bar{x}$  symbolisieren wir dabei zugleich das gewogene arithmetische Mittel. Für den jeweils im Hochsprung erzielten Wert einer Leistungsgruppe von Schülern setzen wir  $a$ .

825		1720		2650			
Schüler 1	95	Schüler 9	110	Schüler 17	115	Schüler 25	120
Schüler 2	100	Schüler 10	110	Schüler 18	115	Schüler 26	120
Schüler 3	100	Schüler 11	110	Schüler 19	115	Schüler 27	120
Schüler 4	105	Schüler 12	110	Schüler 20	115	Schüler 28	125
Schüler 5	105	Schüler 13	110	Schüler 21	115	Schüler 29	125
Schüler 6	105	Schüler 14	115	Schüler 22	115	Schüler 30	125
Schüler 7	105	Schüler 15	115	Schüler 23	120	Schüler 31	130
Schüler 8	110	Schüler 16	115	Schüler 24	120	Schüler 32	130
$\Sigma$	825		1720		2650		3645

Tabelle 5 Geordnete Aufschreibung der Ergebnisse des Hochspringens

Der Index  $j$  steht hier für die Leistungsgruppen von Schülern der Schulklasse, wobei  $j$  von  $1, 2, \dots, m$  läuft und  $m$  die Gesamtzahl dieser Leistungsgruppen innerhalb der Schulklasse darstellt. Das sind in unserem Beispiel 8 Leistungsgruppen.

Berechnen wir das arithmetische Mittel für unser Beispiel mit dem Hochspringen zunächst nach der Formel (8), so ergibt sich laut Tabelle 5 für die  $\sum_{i=1}^n x_i$  ein Gesamtwert von 3645, der durch die Gesamtzahl der Schüler, also durch 32, geteilt wird, d.h.,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{3645}{32} \text{ cm} = 113,9 \text{ cm}$$

Aufgerundet erhalten wir  $\bar{x} = 114 \text{ cm}$ . Dasselbe Ergebnis entsteht, wenn unter Verwendung der Tabelle 6 die Berechnung nach der Formel für das gewogene arithmetische Mittel (9) vorgenommen wird:

$$\bar{x} = \frac{\sum_{j=1}^m a_j g_j}{\sum_{j=1}^m g_j} = \frac{3645}{32} \text{ cm} = 113,9 \text{ cm}$$

Wir erkennen dabei, dass in diesem Falle im Nenner des Bruches anstelle von  $\sum_{j=1}^m g_i$  auch  $n$  geschrieben werden kann.

$a_j$	$g_j$	$a_j \cdot g_j$	
95	1	95	
100	2	200	
105	4	420	
110	6	660	
115	9	1035	
120	5	600	
125	3	375	
130	2	260	
$\sum_{j=1}^m g_j = 32$		$\sum_{j=1}^m a_j g_j = 3645$	

Tabelle 6 Aufbereitung der Daten zur Ermittlung des gewogenen arithmetischen Mittels - Beispiel "Hochspringen"



Hätte man bei diesem Hochspringen die Latte jeweils nicht um 5 cm, sondern nur um 1 oder 2 cm höher gelegt, so wären sicher auch Leistungen zustande gekommen, die sich zwischen den erreichten Resultaten mit jeweils 5 cm Abstand befinden, z. B. 111, 112, 113, 114 cm u.a. Betrachten wir die vorliegenden Ergebnisse mit den Abständen von jeweils 5 cm als Klassengrenzen, die absoluten Häufigkeiten der Klassen als Gewichte, bilden für jede Klasse die Klassenmitte und multiplizieren diese mit der absoluten Häufigkeit in der jeweiligen Klasse, so können wir unter Verwendung der im Abschnitt 4.1. hierfür eingeführten Symbole anstelle von (9) schreiben:

$$\bar{x} = \frac{1}{n} \sum_{m=1}^k h_m u_m \quad (10)$$

In diesem Falle steht  $n$  anstelle von  $\sum_{m=1}^k h_m$ , wobei  $\bar{x}$  der Quotient aus der Summe der Produkte der Klassenmitten und der ihnen zugeordneten absoluten Häufigkeiten sowie der Summe der absoluten Häufigkeiten  $\sum_{m=1}^k h_m$  oder der Gesamtzahl der Messwerte  $n$  ist.

Werden relative Häufigkeiten verwendet, ergibt sich das arithmetische Mittel als

$$\bar{x} = \sum_{m=1}^k f_m u_m \quad (11)$$

Die Berechnung mit Hilfe der Klassen ist bei großem  $n$  mit weniger Arbeitsaufwand verbunden, als wenn alle einzelnen Messwerte für die Berechnung des Mittels herangezogen werden.

Ergeben sich in einigen Fällen Differenzen zwischen den Berechnungen nach den verschiedenen Formeln, so sind diese - eine richtige numerische Ermittlung vorausgesetzt - vor allem durch die Klassenbreite  $d$  und das Nivellieren aller Klassenhäufigkeiten auf die Klassenmitte verursacht.

Berechnen wir das gewogene arithmetische Mittel unter Berücksichtigung einer Klasseneinteilung für das Beispiel des Hochspringens, so erhalten wir ein Resultat, das die Leistungen des gesamten Klassenkollektivs sowie auch der einzelnen Gruppen von Schülern besser widerspiegelt als die vorangegangenen Berechnungsergebnisse. Es trägt der Realität, den tatsächlichen Leistungen der Schüler, besser Rechnung.

Unter Verwendung der Werte aus der Tabelle 7 und der Formel (10) zeigt sich:

$$\bar{x} = \frac{1}{n} \sum_{m=1}^k h_m u_m = \frac{3725}{32} \text{ cm} = 116,4 \text{ cm}$$

Würden wir unsere Häufigkeitsverteilung, die beim Hochspringen der Schüler unseres Klassenkollektivs entstand, grafisch in Form eines Polygonzuges darstellen, so könnten wir darin  $\bar{x}$  als senkrechte Linie einzeichnen.

$h_m$	$u_m$	$h_m \cdot u_m$
1	97,5	97,5
2	102,5	205,0
4	107,5	430,0
6	112,5	675,0
9	117,5	1057,5
5	122,5	612,5
3	127,5	382,5
2	133,5	265,0
$\sum_{m=1}^k h_m = 32$		$\sum_{m=1}^k h_m \cdot u_m = 3725,0$

Tabelle 7 Aufbereitung der Daten zur Ermittlung des gewogenen arithmetischen Mittels auf der Grundlage von Klassen - Beispiel "Hochspringen"

In diesem Falle würde  $\bar{x}$  als Mittelwert der Lage unserer Verteilung fungieren und diese zugleich charakterisieren. Die zeichnerische Darstellung soll jedoch erst beim nächsten Beispiel erfolgen, bei dem wir mehrere Mittelwerte gegenüberstellen wollen. Interessiert uns nur die mittlere Leistung, das mittlere Niveau, eben der Durchschnitt des Klassenkollektivs, so geben die für  $\bar{x}$  berechneten Werte hinreichend Auskunft.

Das arithmetische Mittel ist vor allem durch zwei Eigenschaften gekennzeichnet:

- Die Summe der Abweichungen der Einzelwerte vom arithmetischen Mittel ist gleich Null.
- Die Summe der Quadrate der Abweichungen der Einzelwerte vom arithmetischen Mittel ist ein Minimum im Vergleich zu anderen Mittelwerten.

Wird das arithmetische Mittel aus wenigen  $x_i$  gebildet und weichen diese sehr stark voneinander ab, so kann in bestimmten Fällen ein falsches oder unzureichendes Bild entstehen.

Nehmen wir als Beispiel hierfür die Messwerte für die Tagestemperatur über 24 Stunden, die an verschiedenen Tagen zu gleichen Zeitpunkten gemessen wurden. Für jeden Tag liegen vier Messwerte (in °C) vor, und zwar u.a.

$$\begin{aligned}
 (+8) + (+10) + (+9) + (+7) &= \frac{34}{4} = 8,5 \\
 (+7) + (+13) + (+8) + (+6) &= \frac{34}{4} = 8,5 \\
 (+4) + (+20) + (+8) + (+2) &= \frac{34}{4} = 8,5 \\
 (+2) + (+24) + (+6) + (+2) &= \frac{34}{4} = 8,5
 \end{aligned}$$

In jedem dieser Fälle beträgt die mittlere Tagestemperatur 8,5 °C. Die einzelnen Werte schwanken jedoch zwischen +2 °C und +24 °C.

Dieses Beispiel deutet an, dass es sinnvoll ist, zur Beurteilung bestimmter Situationen nicht allein das arithmetische Mittel, sondern auch andere Angaben wie weitere Mittelwerte, Minimal- und Maximalwerte, Streuung oder Variationsbreite heranzuziehen.

Vor allem sollte das arithmetische Mittel in den nachstehenden Situationen nicht oder höchstens im Zusammenhang mit anderen Maßen verwendet werden:

- bei einer mehrgipfligen Verteilung
- bei sehr kleiner Stichprobe
- bei Verteilungen mit offenen Klassen
- bei stark asymmetrischer Verteilung.

Als typische Maßzahlen der Lage werden von den Statistikern deshalb für bestimmte Aufgaben empfohlen:

- der Zentralwert  $Z$  (auch der Median genannt)
- das Dichtemittel  $D$  (auch als der Modalwert oder der Mode oder der Gipfelwert bezeichnet).

Der Zentralwert  $Z$  oder  $\tilde{x}$  (sprich:  $x$  Schlange) einer Verteilung kann bei geringer Zahl an Beobachtungswerten von der Urliste oder der primären Verteilungstafel ausgehend bestimmt werden. Dazu sind die Messwerte zunächst der Größe nach zu ordnen.

Bei geradem  $n$  ist  $Z$  das arithmetische Mittel aus den beiden in der Mitte liegenden Werten der geordneten Messwerte und bei ungeradem  $n$  der mittelste Wert. Es gibt also ebenso viele größere wie kleinere Messwerte. Bei vielen Messwerten und deren Einteilung in Klassen wird folgende Beziehung verwendet:

$$Z = \tilde{x} = x_u + (x_o - x_u) \frac{\frac{1}{2} \sum_{k=1}^m h_m - \sum_{i=1}^{z-1} h_m}{h_z} \quad (12)$$

Dabei bedeuten

$x_u$ : untere Klassengrenze der Zentralwertklasse

$x_o$ : obere Grenze der vor der Zentralwertklasse liegenden Klasse

$h_m$ : absolute Klassenhäufigkeit

$h_z$ : Klassenhäufigkeit der Zentralwertklasse

$z$ : Zeiger für die Zentralwertklasse

Die Zentralwertklasse bzw. der Zeiger, der sie uns anzeigt, wird bestimmt durch die Beziehung

$$z = \frac{\sum_{k=1}^m h_m + 1}{2} \quad (13)$$

Der Zentralwert verfügt über die sogenannte lineare Minimumseigenschaft, d. h., dass die sich aus den absoluten Beträgen der Abweichungen der Einzelwerte vom Zentralwert ergebende Summe ein Minimum ist, und zwar niedriger als die Summe der Abweichungen der Einzelwerte von irgendeinem anderen Wert.

Untersuchen wir eine Verteilung, so können wir mit dem Zentralwert dasjenige Element der Verteilung bestimmen, unter dem und über dem je 50% aller Elemente liegen. Diese Eigenschaft ist die Ursache dafür, dass in bestimmten Fällen das arithmetische Mittel durch den Zentralwert ersetzt wird.

So können vor allem bei kleinen Messwertreihen stark streuende, nicht typische Werte das arithmetische Mittel erheblich beeinflussen, während der Zentralwert davon unberührt bleibt.

Der Zentralwert bringt nicht wie das arithmetische Mittel die zahlenmäßige Größe aller Einzelwerte zum Ausdruck, er spiegelt nur die Größenordnung wider. Seine Höhe hängt von der Größe des mittelsten Wertes oder der beiden mittleren Werte unserer geordneten Messwertreihe ab.

Werden diese erheblich vergrößert oder verkleinert, ohne dass sich die anderen Werte verändern, so kann sich das auf den Zentralwert maßgeblich auswirken. Für Verteilungen mit an den Enden offenen Klassen eignet sich der Zentralwert besonders, weil seine Höhe nur von dem in der Mitte der Reihe liegenden Wert oder den beiden mittleren Werten abhängt.

Bei medizinischen Beobachtungsergebnissen wird z. B. vom Zentralwert der Verteilung ausgegangen, wenn für das Körpergewicht Grenzwerte festgelegt werden sollen, um den Bereich normaler Werte abzustecken. Hat sich eine Verteilung als Folge von Intensitätsgraden, z. B. der mittleren körperlichen Leistung, ergeben, so ist der Zentralwert gut geeignet, diese Leistung zu charakterisieren.

Bei Versuchen in der Pharmakologie kommt es manchmal darauf an, die für 50% der Versuchstiere tödliche Dosis festzustellen. Um die Versuchsreihen mit möglichst wenig Tieren durchführen zu können, wird hier der Zentralwert statt des arithmetischen Mittels verwendet, da bei kleinem  $n$  das Gesamtergebnis bereits durch einen extremen Wert stark beeinflusst werden könnte.

Das Dichtemittel  $D$  einer Verteilung ist der Wert  $x_i$ , der mit der größten Häufigkeit auftritt. Im Französischen wird diese Maßzahl der Lage in Übereinstimmung mit dem üblichen Sprachgebrauch bezeichnet als "valeur normale" (Normalwert) oder einfach als "normale". Hieraus werden richtig das "normale Heiratsalter" oder das "normale Einkommen" der Umgangssprache abgeleitet.

Das Dichtemittel hängt von den Größenverhältnissen an einer bestimmten Stelle der Verteilung ab und nicht wie das arithmetische Mittel von der Größe aller einzelnen Werte. Gehen in der Verteilung Veränderungen vor, die die Häufungsstelle nicht berühren, so wirken sich diese auf das Dichtemittel nicht aus. Da Klassen mit großer Spannweite die Stelle mit der größten Dichte in der Verteilung gut hervortreten lassen, ist die Bildung großer Klassen für die Ermittlung des Dichtemittels vorteilhaft.

Bei eingipfligen, symmetrischen Verteilungen stimmt das Dichtemittel mit dem Zentralwert überein. Bei linksseitig schiefen Verteilungen ist das Dichtemittel größer als der Zentralwert, und bei rechtsseitig schiefen Verteilungen ist es umgekehrt.

Bei mehrgipfligen Verteilungen liegen mehrere Dichtemittel vor. In dem Fall, in dem sogar alle  $x_i$  verschieden voneinander sind, ist jedes  $x_i$  Dichtemittel.

Wir berechnen das Dichtemittel wie folgt:

$$D = x_u + \frac{d_D(h_D - h_{D-1})}{(h_D - h_{D-1}) + (h_D - h_{D+1})} \quad (14)$$

$x_u$ : untere Klassengrenze der Dichtemittel-Klasse

$h_D$ : absolute Häufigkeit der Dichtemittel-Klasse

$h_{D-1}$ : absolute Häufigkeit der der Dichtemittel-Klasse vorhergehenden Klasse

$h_{D+1}$ : absolute Häufigkeit der der Dichtemittel-Klasse nachfolgenden Klasse

$d_D$ : Breite der Dichtemittel-Klasse

Die Dichtemittel-Klasse ist jene, die die größte Häufigkeit aufweist.

Die Bildung des Dichtemittels setzt voraus, dass gleich breite Klassen gebildet werden. Demgegenüber ist das Dichtemittel bei Gruppen mit unterschiedlicher Gruppenbreite auf Grund der nicht vergleichbaren Häufigkeiten in den einzelnen Gruppen nicht zu berechnen.

Das Dichtemittel hat für die Praxis besonders deshalb große Bedeutung, weil es in einer Häufungsstelle der Verteilung liegt und somit keine rein rechnerische Abstraktion darstellt wie etwa das arithmetische Mittel.

Für den Arzt, der sich z. B. mit Forschungsarbeiten auf dem Gebiet der Säuglingserkrankungen beschäftigt, ist vor allem die Veränderung der Bedingungen im Zusammenhang mit den am häufigsten auftretenden Krankheiten wichtig, um diese erfolgreich bekämpfen zu können. Hier kann die Verwendung des Dichtemittels günstig sein.

Das Dichtemittel bietet auf Grund seiner geschilderten Eigenschaften weiterhin die Möglichkeit, Schlussfolgerungen abzuleiten, z. B. bei der Untersuchung des örtlichen und des zeitlichen Verbrauchs bestimmter Waren, der Analyse von Fehlerursachen in Geräten, bei Kundenstromanalysen im Handel oder in Postämtern usw.

Wir wollen jetzt einmal zu der Aufgabe zurückkehren, für die wir im Abschnitt 4.1. die primäre und die sekundäre Verteilungstafel aufgestellt hatten. Es handelt sich darum, dass Einzelteile gemessen wurden, um die jeweilige Abweichung vom Nennmaß zu ermitteln. Aus den Daten dieser Aufgabe wollen wir nun noch das arithmetische Mittel  $\bar{x}$ , den Zentralwert  $Z$  und das Dichtemittel  $D$  errechnen.

In dem behandelten Beispiel beträgt das arithmetische Mittel  $\bar{x}$  berechnet auf der Grundlage von Formel (10)

$$\bar{x} = \frac{1}{n} \sum_{m=1}^8 h_m u_m = \frac{1583,5}{125} \mu\text{m} = 12,67 \mu\text{m}$$

Wird die Berechnung gemäß Formel (8) vorgenommen, so ergibt sich für dieses Beispiel dasselbe Resultat.

Um den Zentralwert  $Z$  zu berechnen, beginnen wir mit  $z$  nach (13) und kommen mit Hilfe des dabei gewonnenen Ergebnisses nach (12) zu  $Z$ .

$$z = \frac{125 + 1}{2} = 63$$

$$Z = 12,5 \mu\text{m} + (14,5 - 12,5) \frac{0,5 \cdot 125 - 61}{24} \mu\text{m} = 12,5 \mu\text{m} + \frac{3}{24} \mu\text{m} = 12,6 \mu\text{m}$$

Um den Zentralwert  $Z$  für unser Beispiel näherungsweise zu ermitteln, kann in der nach der Größe geordneten Reihe der Messwerte der 63. Wert aufgesucht werden. Er beträgt

13  $\mu\text{m}$  (vgl. auch Tab. 2).

Bei der Berechnung des Dichtemittels  $D$  erhalten wir nach Formel (13), d. h. unter Berücksichtigung der Klassen:

$$D = 10,5\mu\text{m} + \frac{2(26 - 17)}{(26 - 17) + (26 - 24)}\mu\text{m} = 10,5\mu\text{m} + \frac{18}{11}\mu\text{m} = 12,1\mu\text{m}$$

Das Dichtemittel  $D$  kann eventuell auch aus der sekundären Verteilungstafel abgelesen werden, wenn die Messwerte Realisationen einer diskreten Zufallsvariablen  $X$  sind.  $D$  ist gleich dem häufigsten Wert.

Im Beispiel erhalten wir - ausgehend von der Klassenmitte - 11,5  $\mu\text{m}$ . Handelt es sich bei den Messwerten um Realisationen einer stetigen Zufallsvariablen  $X$ , so wird auf die primäre Verteilungstafel zurückgegangen. Der Wert mit der größten absoluten Häufigkeit entspricht  $D$ .

Ist die primäre Verteilungstafel nicht vorhanden, kann man  $D$  auch anhand der Häufigkeitstabelle (vgl. Tab. 3) näherungsweise abschätzen. Nehmen wir an, im Beispiel handle es sich um eine solche stetige Zufallsvariable. Auch dann würde bei einem derartigen Herangehen der Wert 11,5  $\mu\text{m}$  ermittelt.

Werden die verschiedenen Aussagen zu den unterschiedlichen Mitteln dieser Aufgabe zusammengestellt, so erhalten wir folgende Werte:

$$D = 12,1\mu\text{m}; Z = 12,6\mu\text{m}; \bar{x} = 12,7\mu\text{m}$$

Ihre grafische Wiedergabe zeigt Abbildung 17.

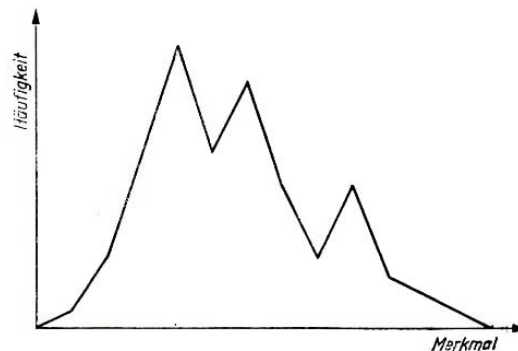


Abb. 17 Häufigkeitspolygon mit  $D$ ,  $Z$  und  $\bar{x}$

Wenn eine Verteilung symmetrisch und eingipflig ist, fallen das arithmetische Mittel, der Zentralwert und das Dichtemittel zusammen ( $D = Z = \bar{x}$ ).

Unterscheiden sich bei einer Verteilung das arithmetische Mittel, der Zentralwert und das Dichtemittel, so sprechen wir von einer schiefen Verteilung. Eine große Differenz zwischen arithmetischem Mittel und Zentralwert oder Dichtemittel deutet auf eine erhebliche Schiefe hin.

Es kann in diesen Fällen erforderlich sein, ein Maß der Verlagerung, eben die Schiefe einer Verteilung, zu berechnen. Die Schiefe ist das "Maß für die Abweichung des häufigsten Wertes einer Verteilung vom Mittelwert" dieser Verteilung.

Sie wird gemessen mit Hilfe der "Summe der dritten Potenzen der Abweichungen der Einzelwerte vom Mittelwert, dividiert durch die mit der Anzahl der Einzelwerte multiplizierte dritte Potenz der Standardabweichung".<sup>15</sup>

Für die linksschiefe Verteilung gilt unter der Voraussetzung, dass nur ein Dichtemittel vorhanden ist,

$$D < Z < \bar{x} \quad (15)$$

Man bezeichnet diese Verteilung auch als "linkssteil", weil auf der linken Seite ein steiler und auf der rechten ein flacher Anstieg vorhanden ist. Im Gegensatz dazu gilt bei einer rechtssteilen Verteilung, auch wieder unter der Voraussetzung, dass nur ein Dichtemittel vorliegt,

$$\bar{x} < Z < D \quad (16)$$

Im Unterschied zur Schiefe ist der Exzess das "Maß für die Abweichung einer Verteilung von der Normalverteilung mit gleichem Mittelwert und gleicher Standardabweichung". Er wird gemessen mit Hilfe der "Summe der vierten Potenzen der Abweichungen der Einzelwerte vom Mittelwert, dividiert durch die mit der Anzahl der Einzelwerte multiplizierte vierte Potenz der Standardabweichung".<sup>16</sup> Im übrigen sei hierzu auch auf die Abbildung 9 verwiesen.

Bei der praktischen Anwendung der drei hier erörterten Mittelwerte  $\bar{x}$ ,  $D$  und  $Z$  können wir uns an die nachstehenden Faustregeln halten (wobei es allerdings auch Ausnahmen gibt):

- Für Vergleichszwecke ist beim Vorliegen mehrerer Reihen im allgemeinen das arithmetische Mittel zu bevorzugen.
- Hat der Mittelwert die Aufgabe, für eine einzelne Reihe das Wesentliche auszudrücken, so ist diese zunächst auf das Vorhandensein eines dichtesten Wertes hin zu analysieren. Erscheint das Dichtemittel nicht typisch oder kann es beim Vorliegen von Klassen nur mit Schwierigkeiten ermittelt werden, so wird überprüft, ob der Zentralwert im gegebenen Fall aussagekräftig ist. Sollte diese Entscheidung auch nicht positiv ausfallen, so wird das arithmetische Mittel gewählt.

Bei wirtschaftsstatistischen Untersuchungen wird in der Regel vom durchschnittlichen oder vom allgemeinen Niveau oder einfach vom Durchschnitt gesprochen. Dabei wird in der Wirtschaftsstatistik der Durchschnitt prinzipiell als ein Verhältnis zweier statistischer Gesamtheiten gekennzeichnet. Soll für eine Erscheinung  $x_i$  (wobei  $i$  der Index für die Unterscheidung der einzelnen Elemente ist) der Durchschnitt  $\bar{x}$  ermittelt werden, so geschieht das, indem eine Beziehung hergestellt wird zwischen der Gesamtheit der Merkmalserscheinung und der Gesamtheit der Bezugserscheinung.

Die Erscheinung  $x_i$  kann z. B. der Pro-Kopf-Verbrauch in Mengeneinheiten für ein bestimmtes Lebensmittel sein. Die Merkmalserscheinung  $m_i$  wäre der Verbrauch in Mengeneinheiten und die Bezugserscheinung  $n_i$  die Zahl der Personen ( $i = 1, 2, \dots, k$ ).

---

<sup>15</sup>DDR-Standard, „Statistische Qualitätskontrolle - Begriffe“. TGL 14449, Blatt 2, S.8

<sup>16</sup>Ebenda, S.8

Der Durchschnitt würde sich dann ergeben als

$$\bar{x} = \frac{\sum_{i=1}^k m_i}{\sum_{i=1}^k n_i} \quad (17)$$

Ist die Bezugserscheinung  $n_i$  für alle Einheiten gleich 1, so gilt die Formel für das einfache arithmetische Mittel - vgl. die Beziehung (8).

"Als Durchschnitt wird das allgemeine, alle einzelnen Größen  $x_i$  unter einer bestimmten Zielstellung charakterisierende Niveau bezeichnet, in dem das Wirken der die Erscheinungen bestimmenden wesentlichen Faktoren zum Ausdruck kommt.

Der Durchschnitt hat stets Beziehungen zweier Erscheinungen zum Inhalt. Seine zahlenmäßige Größe ergibt sich aus der Gegenüberstellung der Gesamtgrößen der beiden Erscheinungen."<sup>17</sup>

Allgemein sind Durchschnittswerte nur dann aussagefähig, wenn eine qualitative Gleichheit für die Gesamtheit vorliegt, aus der der Durchschnitt ermittelt wurde. In den folgenden Beispielen ist das nicht der Fall.

In der BRD werden z. B. die Einkommen verschiedener sozialer Gruppen der Bevölkerung, die von ungelernten Arbeitern gleich denen der Direktoren, "in einen Topf geworfen", wenn das Durchschnittseinkommen eines "Arbeitnehmers" ermittelt werden soll. Für diese unterschiedlichen Gruppen liegt zweifellos keine qualitative Gleichheit vor.

So beträgt das Einkommen eines Direktors das 10- bis 100fache des Lohnes eines ungelernten Arbeiters. Es liegt nahe, dass aus Gründen der Verschleierung im Interesse der herrschenden Klasse bewusst eine falsche Gruppenbildung erfolgte.

Es ist auch nach wie vor eine Tatsache, dass in der BRD der Frau ein erheblich niedrigerer Lohn gezahlt wird als dem Mann, auch wenn sie dieselbe Arbeit verrichtet. Je nach dem Aufbau der Durchschnittsberechnungen kann der bürgerliche Statistiker diesen krassen Unterschied ebenfalls "verschwinden lassen". Der apologetische Charakter der bürgerlichen Statistik offenbart sich hier erneut.

Durchschnittsangaben reichen auch meist nicht aus, wenn es gilt, detaillierte Schlussfolgerungen für Veränderungen in einem konkreten Bereich zu ziehen, da die gegebene Situation anhand von Durchschnittswerten kaum genügend eingeschätzt werden kann. Alle positiven und negativen Ergebnisse "gehen im Durchschnitt unter".

In bestimmten Fällen sind aber gerade die sogenannten Außenseiter von besonderem Interesse. Deshalb muss man - wie dieser gesamte Abschnitt über Mittelwerte andeuten sollte -, bevor die Entscheidung zugunsten eines bestimmten Mittels getroffen wird, für jeden Anwendungsfall gründlich prüfen, welches Mittel den in der Praxis gegebenen Sachverhalt mit hinreichender Aussagekraft widerspiegeln kann bzw. welche einschränkenden Bedingungen oder welche anderen Kennziffern jeweils zu beachten sind.

---

<sup>17</sup>Donda, Herrde, Kuhn, Struck: Statistik. Verlag Die Wirtschaft Berlin 1972, S. 139



Zum Schluss dieses Abschnittes sollen einige weitere statistische Mittel lediglich erwähnt werden.

In soziologischen, pädagogischen und psychologischen Untersuchungen muss das harmonische Mittel verhältnismäßig oft benutzt werden. Das quadratische Mittel, das nur für nicht negative Werte berechnet werden kann, hat für die Praxis weniger Bedeutung. Auf das geometrische Mittel wird im Zusammenhang mit den Zeitreihen (vgl. Kap. 9) eingegangen.

Das Integralmittel wird verwendet, wenn die Gesamtheit unendlich ist, was z. B. bei kontinuierlichen Temperaturmessungen in technologischen Prozessen oder in der Natur der Fall sein kann.

#### 4.4 Die Einzelwerte weichen vom Durchschnitt ab

Die in dieser Überschrift zum Ausdruck kommende Tatsache ist für uns nicht neu. Im folgenden wollen wir uns mit einigen Maßzahlen näher befassen, die dieses Streuen der Einzelwerte charakterisieren.

Der Streubereich, die Spannweite, die Streuung, die Standardabweichung und der Variationskoeffizient einer empirischen Verteilung sind hier zu nennen.

In bestimmten Fällen ist es notwendig, dass man - außer der Kenntnis eines mittleren Wertes für eine Stichprobe oder eine Gesamtheit - auch weiß, wie stark die Beobachtungswerte vom Mittelwert abweichen, wie weit sie unter Bezugnahme auf den Mittelwert streuen.

An den Polygonen der Häufigkeitsverteilungen A und B in der Abbildung 9 war dieser Sachverhalt schon zu erkennen. Für beide Häufigkeitsverteilungen gilt zwar derselbe Mittelwert, die Spannweite und der Streubereich sind bei der Verteilung A aber wesentlich breiter bzw. größer als bei B.

Beim Los B, dem die Messwerte, die zum Häufigkeitspolygon B führten, entnommen wurden, ist der Ausschuss wesentlich geringer, denn der Teil, der links und rechts außerhalb der durch die Qualitätsvorschriften vorgegebenen Grenze T liegt, ist wesentlich kleiner. Bei der Häufigkeitsverteilung B gruppieren sich die Messwerte enger um den Mittelwert als bei A.

Soll z. B. von einer Einheit der Nationalen Volksarmee ein Fluss überwunden werden, so ist es für den Kommandeur zwar ein Anhaltspunkt, wenn er weiß, dass dieser Fluss eine durchschnittliche Tiefe von 1,60 m hat.

Noch mehr ist ihm aber gedient, wenn er bei seiner Entscheidung davon ausgehen kann, dass die Flusstiefe in dem für die Überquerung vorgesehenen Abschnitt vom km  $x$  bis zum km  $y$  zwischen 0,50 m und 2,20 m schwankt. Die Größe 1,60 m gibt die durchschnittliche Tiefe an.

Für den erwähnten Zweck ist aber der Bereich wichtig, innerhalb dessen die verschiedenen gemessenen Tiefen liegen. Sind 0,50 m die geringste Tiefe und 2,20 m die größte Tiefe für diesen Abschnitt des Flusses, haben wir es mit dem Streubereich, der von  $x_{\min}$  bis  $x_{\max}$  reicht, zu tun.

Die Spannweite  $R$ , wobei

$$R = x_{\max} - x_{\min} \quad (18)$$

ist auch als Variationsbreite ( $V$ ) oder Variationsweite bekannt.

Da diese Maßzahl nur von den beiden Extremwerten einer Verteilung ausgeht, die zwischen diesen liegenden Werte jedoch unberücksichtigt lässt, ist sie allerdings für eine genauere Analyse weniger zu empfehlen.

Dem Begriff der Streuung entspricht wohl am besten die Standardabweichung  $s$ , die auch manchmal mittlere Abweichung oder mittlere quadratische Abweichung oder eben Streuung genannt wird. Das Quadrat von  $s$ , also  $s^2$ , wird entweder als Varianz oder Dispersion, aber auch als Streuung bezeichnet.

Wir wollen hier vor allem die Bezeichnungen "Standardabweichung" und "Varianz" verwenden.

Die Varianz ist die "Summe der quadratischen Abweichungen der Einzelwerte vom arithmetischen Mittel, dividiert durch die um Eins verminderte Anzahl der Einzelwerte"<sup>18</sup>.

Die Standardabweichung ist das "Maß für die Streuung der Einzelwerte um ihren Mittelwert", ausgedrückt als "positive Quadratwurzel aus der Varianz".<sup>19</sup>

Die Standardabweichung hat die Dimension des jeweiligen Merkmals, z. B. im vorangegangenen Fall die Dimension  $m$ . Für eine Messreihe  $x_1, x_2, \dots, x_n$  ergibt sich die Varianz nach der Beziehung

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (19)$$

und die Standardabweichung daraus als positive Quadratwurzel

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (20)$$

Es entsteht hier die Frage, warum nicht ebenso wie beim arithmetischen Mittel durch  $n$  dividiert wird, sondern durch  $n-1$ , also durch die um Eins verminderte Zahl der Beobachtungs- bzw. Messwerte.

Das hängt damit zusammen, das von der ermittelten Varianz  $s^2$  der empirischen Verteilung auf die Varianz  $\sigma^2$  in der Grundgesamtheit geschlossen werden soll. Die statistischen Maßzahlen aus Grundgesamtheiten bzw. aus theoretischen Verteilungen werden meist mit griechischen Buchstaben symbolisiert.

Man nennt diese Maßzahlen "Parameter" im Unterschied zu den "Schätzungen der Parameter", die aus den empirischen Verteilungen - mit denen wir es hier zu tun haben - errechnet und mit lateinischen Buchstaben bezeichnet werden.

Ist eindeutig erkennbar, dass "Schätzungen der Parameter", also Maßzahlen einer empirischen Verteilung, vorliegen, werden in der praktischen Arbeit oft die Worte "Schätzungen der" weggelassen.

<sup>18</sup>DDR-Standard, „Statistische Qualitätskontrolle - Begriffe“. TGL 14449, Blatt 2, S. 7

<sup>19</sup>Ebenda, S. 7

Würden wir nun in Formel (19) in den Nenner nur  $n$  aufnehmen, so würde  $\sigma^2$ , die Varianz der Grundgesamtheit, aus  $s^2$ , der Varianz der Messreihe, zu klein geschätzt. Für die statistische Maßzahl  $s^2$  ergibt sich aus wiederholten Messreihen als Mittelwert nicht die Varianz  $\sigma^2$  der Grundgesamtheit, sondern

$$\frac{n-1}{n}\sigma^2$$

Damit würde die Varianz  $\sigma^2$  in der Grundgesamtheit zu niedrig angenommen, weshalb bei der Stichprobenvarianz durch  $n-1$  dividiert werden muss.

Für die  $n-1$  Messwerte ist auch der Begriff Freiheitsgrade gebräuchlich, weil nur über  $n-1$  Mess- oder Beobachtungswerte "frei" verfügt werden kann. Die Freiheitsgrade drücken die voneinander unabhängigen Beobachtungen oder Messungen aus oder - anders formuliert - die "Anzahl der unabhängigen Summanden auf einer Seite einer Bestimmungsgleichung, wenn die andere Seite der Bestimmungsgleichung konstant gehalten wird"<sup>20</sup>.

Wir wollen uns diesen Zusammenhang mit Hilfe des Mittelwertes  $\bar{x}$  verdeutlichen. Bevor das arithmetische Mittel bestimmt wurde, konnten wir über alle  $n$  Glieder unserer Messreihe frei verfügen. Nach der Berechnung des Mittelwertes vermindern sich die Freiheitsgrade um einen. Wir wissen, dass  $n \cdot \bar{x}$  gleich der Summe aller Einzelwerte sein muss, d.h.

$$n \cdot \bar{x} = \sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

Einer dieser Einzelwerte  $x_i$  - wir nehmen den letzten  $x_n$  - ist jedoch dadurch festgelegt, dass

$$x_n = n\bar{x} - \sum_{i=1}^{n-1} x_i \quad (21)$$

ist.

Das lässt sich auch an unserem Beispiel mit dem Hochspringen demonstrieren. Wir hatten 32 am Hochspringen beteiligte Schüler,  $\bar{x}$  betrug 113,9 cm, und die Summe aller Hochsprungergebnisse - außer dem letzten Resultat des 32. Schülers, der 130 cm hoch sprang — war 3515 cm. Setzen wir diese Zahlen in die Formel (21) ein, so erhalten wir

$$x_{32} = 32 \cdot 113,9 \text{ cm} - 3515 \text{ cm} = 130 \text{ cm}$$

Es können also tatsächlich nur  $n-1$  Werte frei gewählt, bzw. es kann nur über  $n-1$  Werte frei verfügt werden.

Wie wir an anderer Stelle schon sahen, kann ein Einzelwert bei kleinen Mess- oder Beobachtungsreihen das Berechnungsergebnis unter Umständen erheblich beeinflussen. Bei  $n > 100$  ist das jedoch nicht mehr der Fall. Für entsprechende Berechnungen darf dann im Nenner der Varianz -1 entfallen und lediglich  $n$  verwendet werden. Die sich ergebende Differenz ist so gering, dass sie vernachlässigt werden kann.

---

<sup>20</sup>DDR-Standard, „Statistische Qualitätskontrolle - Begriffe“. TGL 14449, Blatt 2, S. 20

Bei den Streuungsmaßen werden für die praktische Arbeit im allgemeinen die Standardabweichung  $s$  und die Varianz  $s^2$  bevorzugt. Sie geben einerseits in geeigneter Weise Auskunft über die Streuung der Einzelwerte innerhalb der Mess- oder Beobachtungsreihe und dienen andererseits zur Schätzung der Streuung in der Grundgesamtheit.

Wir wollen für das Beispiel mit den 125 Werkstücken, bei denen jeweils durch Messung die Abweichung vom Nennmaß geprüft wurde, die Varianz  $s^2$  im Interesse eines möglichst geringen Rechenaufwandes aus der sekundären Verteilungstabelle ermitteln. Hierbei muss der Tatsache Rechnung getragen werden, dass Klassen gebildet und diesen entsprechende Häufigkeiten zugeordnet wurden. Demzufolge ist die Formel (19) etwas zu verändern, wobei sich jetzt ergibt:

$$s^2 = \frac{1}{n-1} \sum_{m=1}^k (u_m - \bar{x})^2 h_m \quad (22)$$

Die Bedeutung der einzelnen Symbole ist bereits in den Abschnitten 4.1. und 4.3. erläutert worden. Zur Berechnung sowie zur besseren Übersicht wird die schon vorhandene sekundäre Verteilungstafel (Tab. 3) um drei Spalten ergänzt, so dass die Auswertungstabelle 8 entsteht. (siehe nächste Seite)

Ergebnis:  $\sum_{m=1}^8 (u_m - \bar{x})^2 h_m = 1433,59$

Rechnerisch ergibt sich:

$$s^2 = \frac{1}{125-1} \cdot 1433,59 \mu\text{m}^2 = 11,56 \mu\text{m}^2$$

Nachdem  $s^2$  bekannt ist, wird daraus die Standardabweichung berechnet.

$$s = \sqrt{11,56 \mu\text{m}^2} = 3,4 \mu\text{m}$$

Der Wert  $3,4 \mu\text{m}$  besagt, wie stark die gemessenen Werte in der Verteilung um den Mittelwert streuen. Varianz und Standardabweichung lassen als Maßzahlen bereits Schlüsse über die vorliegende Mess- oder Beobachtungsreihe zu. Sie sind aber ebenfalls für weitere statistische Berechnungen notwendig.

Liegen mehrere empirische Verteilungen aus einem Untersuchungsbereich vor, die z. B. zu unterschiedlichen Zeiten ermittelt worden sind, kann es notwendig sein, die verschiedenen Werte von Maßen der Lage oder der Streuung miteinander zu vergleichen. Haben die Verteilungen erheblich voneinander abweichende Mittelwerte, genügt es nicht, die Streuungsmaße einfach einander gegenüberzustellen.

Fälle mit stark voneinander abweichenden Mittelwerten sind in der Wirtschaft, in der Technik und Technologie sowie im naturwissenschaftlichen Bereich relativ oft festzustellen. Beispiele hierfür sind:

- die Behandlung von an derselben Krankheit leidenden Patienten in einer Klinik mit unterschiedlichen Medikamenten und nicht übereinstimmender Streuung im Hinblick auf die Krankheitsdauer

Klassen-Nr.	Klassengrenzen untere/obere	Klassenmitte	absolute Häufigkeit	Produkt aus Spalte 3 und 4	relative Häufigkeit $f_m$ in %	relative Summenhäufigkeit SP $\Sigma$ in %	$u_m - \bar{x}$	$(u_m - \bar{x})^2$	$(u_m - \bar{x})^2 h_m$
$e_m$	$e_{mu}, e_{mo}$	$u_m$	$h_m$	$u_m \cdot h_m$	$f_m$ in %	SP $\Sigma$ in %	$u_m - \bar{x}$	$(u_m - \bar{x})^2$	$(u_m - \bar{x})^2 h_m$
1	2	3	4	5	6	7	8	9	10
1	4,5...6,5	5,5	3	16,5	2,4	2,4	-7,18	51,55	154,65
2	6,5...8,5	7,5	15	112,5	12,0	14,4	-5,18	26,83	402,45
3	8,5...10,5	9,5	17	161,5	13,6	28,0	-3,18	10,11	171,87
4	10,5...12,5	11,5	26	299,0	20,8	48,8	-1,18	1,39	36,04
5	12,5...14,5	13,5	24	324,0	19,2	68,0	0,82	0,67	16,08
6	14,5...16,5	15,5	19	294,5	15,2	83,2	2,82	7,95	71,55
7	16,5...18,5	17,5	17	297,5	13,6	96,8	4,82	23,23	394,91
8	18,5...20,5	19,5	4	78,0	3,2	100,0	6,82	46,51	186,04
		$n = \sum_{m=1}^k h_m = 125$	1583,5	100,0 %					

Tabelle 8 Auswertungstabelle

- der Vergleich der Streuung bei der Ermittlung der Reißfestigkeit des Garnes in der Textilindustrie
- eine Gegenüberstellung bei starker Streuung der Niederschlagsmenge in örtlicher und zeitlicher Hinsicht oder
- bei einem Vergleich der Streuung für die Höhe der Sparguthaben in einer Großstadt und einem Kreis mit vorwiegend ländlicher Bevölkerung.

In derartigen Fällen kann als Vergleichsmaß der Variationskoeffizient  $v$ , eine dimensionslose Größe, berechnet werden, der manchmal auch als Variabilitätskoeffizient bezeichnet wird.

Der Variationskoeffizient setzt die Standardabweichung zum arithmetischen Mittel ins Verhältnis und wird in Prozent angegeben.

$$v = \frac{s}{\bar{x}} \cdot 100\% \quad (23)$$

Wir wollen die unterschiedlichen Größen des Variationskoeffizienten an Beispielen betrachten. In einem Betrieb der elektronischen Industrie werden Bauelemente hergestellt. Ihre Qualität wird an einem Merkmal ermittelt.

Bei einer Messwertreihe aus einem Los vom November 1973 berechnete man  $s_1 = 3,4$  und  $\bar{x}_1 = 12,7$ . Bei einer anderen Prüfung aus dem im Monat zuvor produzierten Los betrugen die Werte  $s_2 = 7,3$  und  $\bar{x}_2 = 10,6$ . Der Vergleich der beiden Variationskoeffizienten ergibt:

$$v_1 = \frac{3,4}{12,7} \cdot 100\% = 26,8\% \quad , \quad v_2 = \frac{7,3}{10,6} \cdot 100\% = 69,0\%$$

Bei der empirischen Verteilung mit  $v_1 = 26,8\%$  war ein als normal zu bezeichnender Teil des Loses Ausschuss. Bei der zweiten empirischen Verteilung mit  $v_2 = 69\%$  wurde dagegen ein sehr hoher Anteil an Ausschuss festgestellt. Die Kontrolle des Ausgangsmaterials ergab, dass es erhebliche Mängel aufwies. An diesem Beispiel ist eine grundsätzliche Tendenz zu erkennen: Je mehr sich  $v$  dem Wert 0 nähert, um so günstiger ist das Ergebnis zu werten.

Stichprobe aus dem Monat	Arithmetisches Mittel $\bar{x}$	Standardabweichung $s$	Variationskoeffizient $v$
November 1973	$\bar{x}_1 = 12,67$	$s_1 = 3,4$	$v_1 = 26,8\%$
Oktober 1973	$\bar{x}_2 = 10,6$	$s_2 = 7,3$	$v_2 = 69,0\%$
September 1972	$\bar{x}_3 = 10,8$	$s_3 = 2,9$	$v_3 = 26,9\%$
Oktober 1972	$\bar{x}_4 = 6,2$	$s_4 = 1,7$	$v_4 = 27,4\%$

Tabelle 9 Vergleich von Variationskoeffizienten

Im Vorjahr waren im gleichen Betrieb ebenfalls zwei gleichgroße Lose von demselben Produkt aufgelegt worden, wobei auch die Qualität geprüft wurde. Die Variationskoeffizienten, die sich bei den verschiedenen Qualitätskontrollen ergaben, gehen aus der Tabelle 9 hervor.

Es ist zu erkennen, dass die Variationskoeffizienten für die erste, die dritte und die vierte empirische Verteilung sehr dicht beieinander liegen, während der Koeffizient für die zweite Prüfung mehr als doppelt so groß ist. Die Untersuchung, ob die geringen Unterschiede für  $v_1$ ,  $v_3$  und  $v_4$  zufällig aufgetreten sind oder nicht, muss anderen Analysen überlassen bleiben.

Die Anwendung des Variationskoeffizienten ist jedoch an zwei Voraussetzungen gebunden, und zwar muss die verwendete Skale einmal einen absoluten Nullpunkt und zum anderen konstante Klassengrößen haben. In der Medizin, der Psychologie und zum Teil auch bei soziologischen Untersuchungen sind diese beiden Bedingungen nicht immer erfüllt. Manchmal werden dann spezielle Verfahren angewendet, um einen absoluten Nullpunkt künstlich zu schaffen.

Die Ermittlung des Variationskoeffizienten erleichtert das in Abbildung 18 dargestellte Nomogramm. Aus der Abbildung geht hervor, dass in der linken Skale  $\bar{x}_3 = 10,8$  und in der mittleren  $s_3 = 2,9$  gekennzeichnet wurden. Die Verbindung beider Punkte mit einer Geraden, die weitergeführt wird bis in die dritte Skale, lässt uns am Schnittpunkt bei etwa 27% den gesuchten Wert  $v_3$  finden.

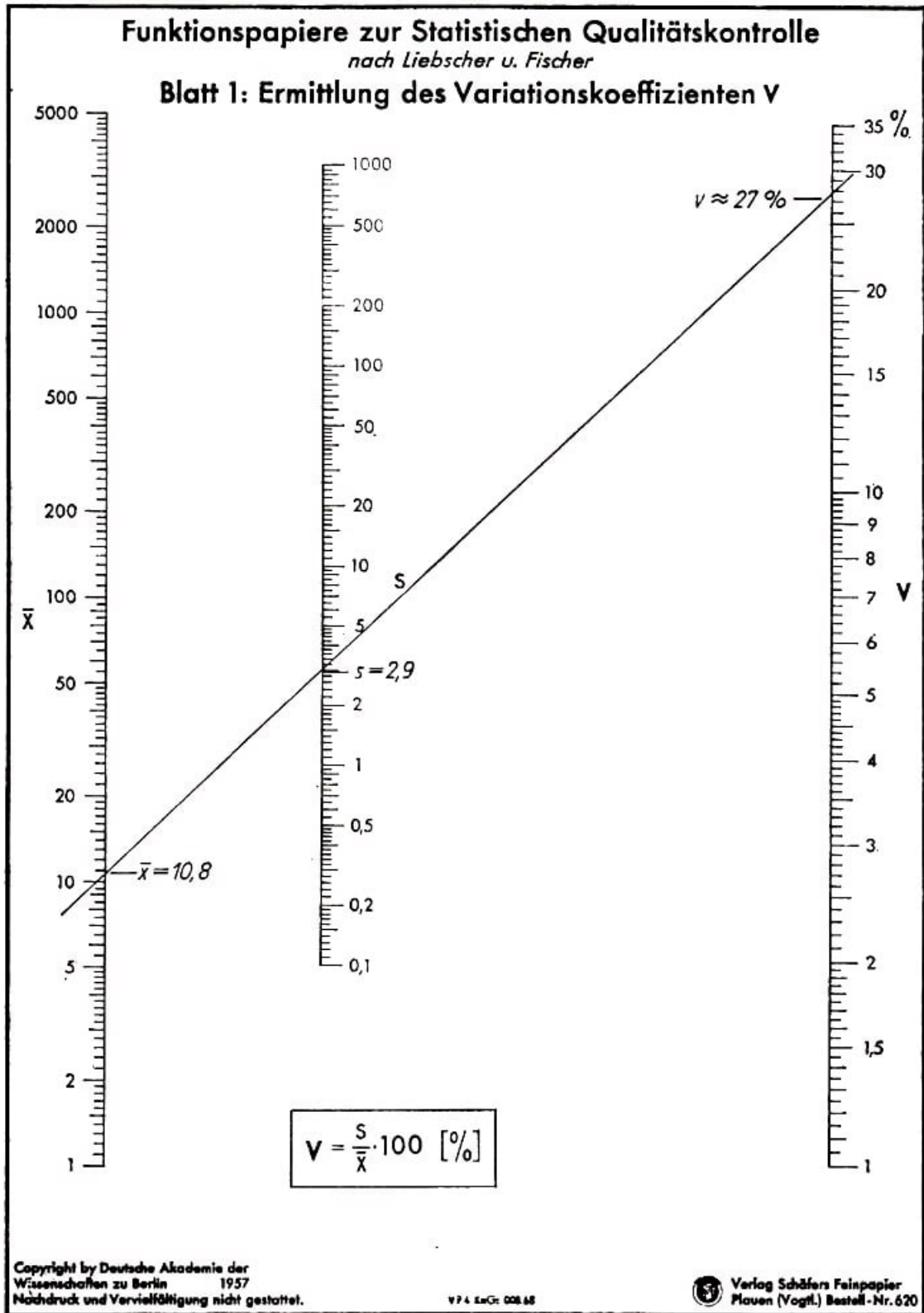


Abb. 18 Grafische Ermittlung des Variationskoeffizienten



## 5 Der Zusammenhang zwischen Ursache und Wirkung wird zahlenmäßig bestimmt

Im vorangegangenen Kapitel untersuchten wir in unseren Beispielen jeweils eine Zufallsvariable bzw. eine eindimensionale (monovariabel) empirische Verteilung. In diesem Abschnitt wollen wir zweidimensionale (bivariable) empirische Verteilungen betrachten. Sie sind dadurch charakterisiert, dass zwei Zufallsvariablen erfasst werden und der zwischen ihnen bestehende Zusammenhang untersucht wird.

In der Praxis interessieren nur solche Fälle, in denen ein ursächlicher, ein kausaler Zusammenhang besteht. Die Zufallsvariablen spiegeln dann objektiv vorhandene Abhängigkeiten wider.

Betrachten wir z. B. den Kohleverbrauch ( $y_i$ ) einer Lokomotive beim Transport von Wagenzügen unterschiedlicher Masse ( $x_i$ ) auf einer bestimmten Eisenbahnstrecke oder den getesteten Erinnerungswert ( $y_i$ ), d.h. jenen Anteil, der nach einer bestimmten Zeit von einem aufgenommenen kurzen Text in Erinnerung geblieben ist bezogen auf die Größe der Darstellung ( $x_i$ ) eines Textes in einer Zeitung, so führt neben anderen die Ursache  $x$  zur Wirkung  $y$ .

Zu solchen Zusammenhängen stellt die marxistisch-leninistische Philosophie fest: "Ursache und Wirkung sind korrelative Begriffe. Eine Erscheinung, die eine andere hervorruft, tritt in Bezug auf diese als Ursache auf. Das Resultat des Wirkens der Ursache ist die Wirkung."<sup>21</sup>

Auf der Grundlage dieser Erkenntnis könnte man aus der Sicht der Statistik konstatieren: In den geschilderten Beispielen besteht eine Regression von  $y$  bezüglich  $x$ . Daraus erhebt sich die Frage nach der Bedeutung von "Regression" und "Korrelation" in der Statistik.

### 5.1 Was bedeuten „Korrelation“ und „Regression“?

Die Worte Korrelation und Regression wurden von F. Galton im vorigen Jahrhundert eingeführt, wobei von ihm "Co-Relation" (Wechselbeziehung, Zusammenhang, Abhängigkeit) geschrieben wurde. Er wollte damit einen objektiv existierenden und mit stochastischen Mitteln zu erfassenden Zusammenhang zwischen Messreihen oder allgemein zwischen Zahlenreihen charakterisieren.

Ausgehend vom Studium des seinerzeit stark beachteten biologischen Werkes "Der Ursprung der Arten" (Charles Darwin, 1859), untersuchte Galton statistisch den Zusammenhang zwischen der Körpergröße von Vätern und Söhnen. Dabei stellte er fest, dass einerseits zwar große Väter große Söhne haben, aber andererseits diese Beziehung nicht in jedem Falle gegeben ist.

Würden große Väter nur große Söhne haben und kleine Väter nur kleine Söhne und würde sich diese Entwicklung immer wieder fortsetzen, so müsste es unter den Menschen nur "Riesen" und "Zwerge" geben. Wie jeder selbst sehen kann, ist das aber nicht

---

<sup>21</sup>Autorenkollektiv: Grundlagen der marxistischen Philosophie. Dietz Verlag, Berlin 1965, S. 169

der Fall.

Aus seinen Ergebnissen erkannte Galton als Massenerscheinung ein "Zurückdrücken" oder "Zurückführen" der Körpergröße der Kinder auf das mittlere Maß, die Durchschnittsgröße der Bevölkerung. Diesen Vorgang bezeichnete er als "Regression".

Die von Galton gewonnenen Erkenntnisse auf dem Gebiet der Regression wurden im Laufe der Zeit verallgemeinert. Das Wort Regression wird in der Gegenwart vor allem verwendet, um "einseitige" stochastische Abhängigkeiten zu charakterisieren. Einseitig bedeutet hier, dass bei jeder Regressionsanalyse, mit deren Hilfe die Abhängigkeit zwischen zwei Variablen untersucht werden soll, eine dieser Variablen unabhängig ist. Ihre Werte werden vor Beginn der Untersuchung festgelegt.

Die Regressionsanalyse hebt den Einfluss der Hauptursache, die Art und die Verlaufsform des Zusammenhangs hervor. Die Nebenursachen werden ausgeschaltet. Bei der Korrelationsanalyse wird vom Erscheinungsbild des Zusammenhangs, z. B. den Ergebnissen der Regressionsanalyse, die die Art und die Verlaufsform des Zusammenhangs verdeutlichen, ausgegangen.

Ein Ausdruck hierfür ist die Regressionsfunktion mit ihrem Regressionskoeffizienten, einem Parameter dieser Funktion. Im Unterschied zur Regression ist für die Korrelation typisch, dass beide Zufallsvariablen als abhängige Variablen betrachtet werden. Dabei liegt eine Korrelation allgemein dann vor, wenn zwei Zufallsvariablen mit ihren in Mess- oder Beobachtungsreihen zum Ausdruck kommenden Realisationen den ursächlichen Zusammenhang widerspiegeln, der zwischen Erscheinungen oder Prozessen der Natur, der Technologie, der Technik oder der Gesellschaft existiert.

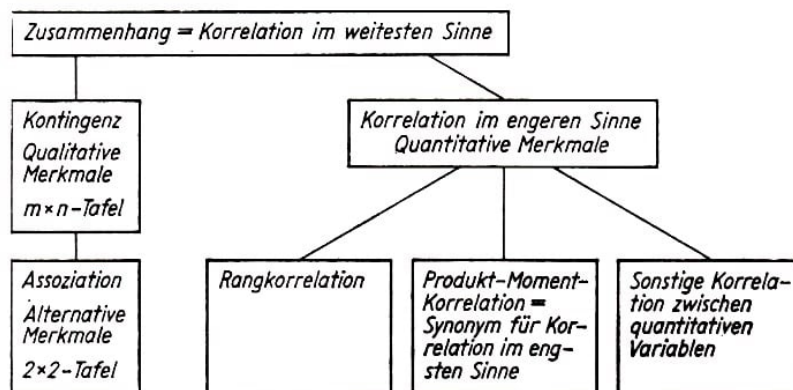


Abb. 19 Arten der Korrelation

Mit Hilfe des Korrelationskoeffizienten wird der Grad (die Intensität, Stärke, Straffheit, Enge, Dichte) dieses Zusammenhangs gemessen. Diese Charakterisierung der Stärke des Zusammenhangs zwischen den untersuchten Erscheinungen ist gleichbedeutend mit der Einschätzung des Einflusses der restlichen Ursachen, die nicht unmittelbar in der Berechnung berücksichtigt wurden.

Ebenfalls wird überprüft, wie die Regressionsfunktion den Zusammenhang widerspiegelt, insbesondere "ob die in die Regression aufgenommene Hauptursache nach dem

Grad ihres Einflusses auf die Wirkung ausreicht, um durch die Regression das Wesentliche des Zusammenhangs widerzuspiegeln, oder ob mehr als eine Hauptursache einbezogen werden muss".<sup>22</sup>

Die Abbildung 19 gibt schematisch einen Überblick über verschiedene Arten der Korrelation.<sup>23</sup> Es handelt sich in allen diesen Fällen um einen stochastischen und keinen funktionalen Zusammenhang im Sinne einer eindeutig determinierten mathematischen Funktion. Der jeweils vorliegende Zusammenhang kann nicht exakt determiniert, sondern nur auf der Grundlage von Wahrscheinlichkeiten wiedergegeben werden. Derartige vom Zufall bedingte Zusammenhänge nennt man stochastische oder korrelative Zusammenhänge oder kurz Korrelation.

Sie werden mit Hilfe der Korrelationsanalyse oder Korrelationsstatistik bzw. der Regressionsanalyse oder Regressionsstatistik untersucht.

Auch wenn - wie wir noch im einzelnen sehen werden - die Regression in einer Funktion sichtbar gemacht wird, so unterscheidet sich diese Regressionsfunktion doch von der strengen mathematischen Funktion. Für die letztere können wir z. B. sagen  $y = 3x$  und  $x = \frac{1}{3}y$ .

Der Zusammenhang zwischen diesen beiden Variablen ist umkehrbar eindeutig. Ein solcher Zusammenhang gilt oder gilt nicht. Eine derartige Eindeutigkeit ist bei der Regressionsfunktion nicht gegeben, da mindestens eine stochastische Zufallsvariable vorliegt.

Sind es zwei Veränderliche, dann ist für jedes  $x_i$  eine Anzahl von  $y$ -Werten (also nicht nur einer wie bei der strengen mathematischen Funktion) vorhanden, genauso wie bei der Korrelation zu jedem  $y$ -Wert eine Menge von  $x$ -Werten gehört oder umgekehrt.

Das Hauptanliegen der Regressions- bzw. Korrelationsrechnung kann mit wenigen Worten etwa wie folgt formuliert werden:

Die Regressionsanalyse ist darauf gerichtet,

- die Art, die Verlaufsform des Zusammenhangs zu fixieren
- festzustellen, wie eine oder mehrere Erscheinungen auf Folgeerscheinungen wirken, und zwar indem die Einzelwerte zu Regressionsfunktionen ausgeglichen, diese Funktionen selbst und damit der Einfluss der unabhängigen Variablen auf die abhängige Variable ermittelt werden
- bisher noch unbekannte Werte der abhängigen Variablen bzw. der Folgeerscheinung durch Interpolation oder Extrapolation abzuschätzen.

Die Zielstellung der Korrelationsanalyse besteht darin,

- den Grad, die Stärke des korrelativen Zusammenhangs zwischen zwei oder mehreren Erscheinungen zu bestimmen
- zu helfen, vermutete, jedoch noch nicht bewiesene oder noch nicht bekannte Zu-

---

<sup>22</sup>Donda, Herrde, Kuhn, Struck: Statistik. Verlag Die Wirtschaft, Berlin 1972, S. 187

<sup>23</sup>Vgl. Autorenkollektiv: Biometrisches Wörterbuch. Band 1. VEB Deutscher Landwirtschaftsverlag, Berlin 1968, S. 331

sammenhänge aufzudecken bzw. zusätzliche Informationen über existierende allgemein bekannte Zusammenhänge zu erhalten

- die Einflussfaktoren und den Grad ihrer Auswirkungen auf bestimmte Prozesse oder Erscheinungen zu erfahren, und danach wiederum durch Erhöhung oder Verminderung des Einflusses eines einzelnen Faktors (oder einiger Faktoren) eine andere Gesamtwirkung zu erreichen.

## **5.2 Kausale Zusammenhänge**

Kausale Zusammenhänge lassen sich theoretisch immer entweder funktional oder korrelativ darstellen. Für die praktische Arbeit sind dabei die Grenzen der jeweiligen statistischen Methode zu beachten. Jede Funktion oder Korrelation drückt aber nicht notwendigerweise einen objektiv bestehenden ursächlichen Zusammenhang aus.

Ein Beispiel für die zuletzt erwähnte These ist eine Korrelation, die einmal zum Spaß berechnet wurde und seitdem wohl in jeder Vorlesungsreihe über die Korrelationsrechnung erwähnt wird.

Für mehrere Jahre wurde die Anzahl der in Südschweden nistenden Störche mit der Geburtenhäufigkeit in diesen Jahren gegenübergestellt. Dabei ergab sich eine positive Korrelation. Es ist offensichtlich, dass es sich hier um keinen ursächlichen Zusammenhang im Sinne von Ursache und Wirkung handelt.

Ein ähnliches Beispiel, das ebenfalls nur ein Lächeln auslösen kann, ist eine Untersuchung mit Hilfe der Daten von 643 Studentinnen der "University of California" über eine Korrelation zwischen der Intelligenz und der Schönheit der Studentinnen.

Alle in dieser Untersuchung einbezogenen Studentinnen mussten mindestens zwei Jahre oder länger dieser Hochschule angehört haben. Sie wurden in drei Gruppen eingeteilt, und zwar: Jüngere, Ältere, Diplomierte.

Für jede dieser Gruppen wurde wiederum die "persönliche Erscheinung" klassifiziert, wobei vier Möglichkeiten bestanden, und zwar schön, gut aussehend, gewöhnlich und hausbacken. Danach wurde versucht, mit Hilfe der Korrelationsrechnung zu analysieren, ob ein korrelativer Zusammenhang zwischen den fachlichen Leistungen und der beschriebenen Art der "persönlichen Erscheinung" besteht. Ganz abgesehen davon, dass "Schönheit" ein sehr unbestimmter subjektiver Begriff ist, wurde hier kein kausaler Zusammenhang erforscht.

Wird in solchen Beispielen wie der "Störche-Geburten-Beziehung" oder der "Schönheit-Intelligenz-Relation" rechnerisch ein Zusammenhang nachgewiesen, kann es sich im allgemeinen nur um eine Scheinkorrelation - auch als mittelbare Korrelation bezeichnet - oder um eine unsinnige, also eine Nonsense- Korrelation handeln.

Für das Beispiel mit den Störchen findet man in der Literatur beide Bezeichnungen (Scheinkorrelation und Nonsense-Korrelation). Das zeigt, wie schwierig es manchmal ist, Nonsense- bzw. Scheinkorrelationen auseinanderzuhalten.

Für die Scheinkorrelation zwischen Geburten und Störchen gibt es in der Literatur z. B. mehrere Begründungen:

- Mit zunehmender Einwohnerzahl nimmt in einem Dorf auch die Zahl der Gebäude und damit die Zahl der Nistplätze für die Störche zu.
- Es werden vorwiegend städtische und ländliche Gegenden gegenübergestellt. In ländlichen Gebieten war bekanntlich die Geburtenhäufigkeit größer als in den Städten. Da Störche nur in ländlichen Gebieten nisten und kaum in Städten, kann aus der Zunahme der Storchennester ein scheinbarer Zusammenhang mit den zunehmenden Geburten in ländlichen Gebieten entstehen.
- Die rein zufällig zeitlich in gleicher Richtung sich bewegenden Variablen "Geburten" und "Storchennester" mit dem Ergebnis der positiven Korrelation.

Oft sind bei Scheinkorrelationen die eine (die Wirkung) oder beide korrelierenden Erscheinungen von einer dritten Erscheinung abhängig. Ein Beispiel hierfür ist der statistisch nachgewiesene Zusammenhang zwischen dem niedrigen Ernteertrag und einer hohen Zahl von Bränden in ländlichen Gebieten des zaristischen Russlands.

Auch hier besteht kein ursächlicher Zusammenhang. Die Ursache für beide Erscheinungen, die Brände und die niedrigen Ernteerträge, ist in einer dritten Erscheinung zu sehen, nämlich der sehr trockenen Witterung dieser Jahre.

Die Nonsense- und die Scheinkorrelation zeigen erneut, wie wichtig es ist, dass der Fachmann des betreffenden Gebietes, der über die unbedingt notwendige Sachkenntnis verfügt, statistische Angaben analysiert. Die Frage nach der kausalen Abhängigkeit ist dabei eine Grundfrage jeder statistischen Untersuchung.

Gerade auf die richtige, den objektiven Realitäten voll Rechnung tragende Beurteilung der stochastischen Zusammenhänge kommt es für die Praxis an. Deshalb "darf der Forscher beim Feststellen, dass zwischen seinem  $X$  und  $Y$  ein Zusammenhang besteht, nie stehenbleiben; er soll stets nach Möglichkeit danach streben, zu ergründen, was der beobachtete Zusammenhang eigentlich bedeutet, worauf er letzten Endes beruht.

In den Fällen, wo die vom Statistiker erzielten Ergebnisse für praktische Ratschläge und Entschlüsse verwertet werden, ist das halbe Wissen von Zusammenhängen, die ohne Deutung bleiben oder gar unrichtig gedeutet werden, oft schlimmer als Nichtswissen."<sup>24</sup>

Die Erfassung von Ursache und Wirkung ist Voraussetzung, damit der zu analysierende Zusammenhang wahrheitsgetreu und umfassend dargestellt werden kann. Da sich außer der hauptsächlichen Ursache meist Nebenursachen auswirken, ist - je nach dem Grad ihres Einflusses - unter Umständen auch diesen entsprechendes Augenmerk zu schenken. Gleichzeitig ist zu beachten, dass sich mit dem Zurücktreten der Neben- sowie der Zufallseinflüsse ein korrelativer Zusammenhang einem funktionalen Zusammenhang annähern kann.

Dasselbe Ergebnis könnte eintreten, wenn mehrere Zufallsvariablen in bestimmter Form verbunden sind und zugleich untersucht werden. Infolge von Messfehlern können auch funktionale Zusammenhänge als Korrelationen in Erscheinung treten. Das trifft u. a. dann zu, wenn sich ein physikalisches Gesetz unter unterschiedlichen realen Bedingun-

---

<sup>24</sup>Tschuprow, A. A.: Grundbegriffe und Grundprobleme der Korrelationstheorie. Leipzig-Berlin 1925, S. 17/18, (Moskau 1926)

gen durchsetzt.

Sowohl beim Vorliegen einer Funktion als auch bei einer Korrelation handelt es sich um die Verbindung von Erscheinungen oder Prozessen, wobei diese zahlenmäßig wiedergegeben wird.

Der Korrelationskoeffizient sagt uns, wie stark oder wie streng ein statistischer Zusammenhang ist. Es wird also weder eine Auskunft darüber gegeben, welche der beiden Veränderlichen als Ursache die Wirkung auslöste, noch ob ein solcher sachlichlogischer Zusammenhang existiert. Hier wird wieder deutlich sichtbar, dass die statistischen Methoden nur ein Hilfsmittel sein können und die Kenntnis des Untersuchenden vom Prozess der ausschlaggebende Faktor bei der Durchführung statistischer Analysen ist.

Innerhalb des methodischen Instrumentariums der Statistik ist aber die Korrelationsanalyse ein vorteilhaftes Hilfsmittel zur Unterstützung der schöpferischen Arbeit des Menschen, zur Gewinnung neuer Erkenntnisse eben durch das Aufdecken und Messen von Zusammenhängen.

Um die Korrelations- und Regressionsrechnung, überhaupt die statistischen Analysemethoden richtig anwenden zu können, benötigt zwangsläufig jeder Fachmann, der sich mit statistischen Angaben beschäftigen muss, zumindest ein Minimum an Kenntnissen auf dem Gebiet der Statistik. Bei umfangreichen Untersuchungen, wie sie z. B. auf verschiedenen wissenschaftlichen Gebieten notwendig sind, ist eine enge Zusammenarbeit zwischen Fachwissenschaftlern und Statistikern unbedingt erforderlich.

Manchmal ist der Statistiker mit der Entscheidung, ob eine Scheinkorrelation vorliegt oder nicht, einfach überfordert.

Das gilt insbesondere für jene Fälle, in denen die mathematisch- statistischen Methoden formal richtig angewendet wurden. Dem Statistiker fehlen zum Teil die sachlichen Voraussetzungen, die Fachkenntnisse, um bei solchen Fragen, ob eine Nonsense- oder eine Scheinkorrelation oder eine echte, eine unmittelbare Korrelation vorliegt, richtig entscheiden zu können.

Hier muss eben vor allem der Fachmann des betreffenden Gebietes seinen Standpunkt äußern, denn bei statistischen Untersuchungen in der Technik, Technologie oder in den Natur- bzw. Gesellschaftswissenschaften kommt es auf objektiv vorhandene kausale Zusammenhänge an. Eine enge, kameradschaftliche Zusammenarbeit zwischen den Statistikern und den Fachvertretern anderer Gebiete wird also am Beispiel der Korrelationsrechnung besonders deutlich.

Den Ingenieur oder den Ökonomen, den Wissenschaftler oder den vorwiegend politisch tätigen Funktionär interessieren nur jene Fälle, bei denen die unabhängige Variable unmittelbar auf die abhängige Variable einen bestimmten Einfluss ausübt. Deshalb sollen mit Hilfe der Statistik auch nur diese objektiv bestehenden unmittelbaren Korrelationen nachgewiesen werden. Man nennt sie auch echte Korrelationen.

Unmittelbare Korrelationen bestehen z. B. innerhalb bestimmter Grenzen zwischen der Bruchfestigkeit und der Härte einer bestimmten Materialsorte, zwischen der Zahl der Trainingsstunden und den sportlichen Leistungen von Sportlern, zwischen den Akku-

mulationsquoten und der Größe der Zuwachsrates des Nationaleinkommens (wobei es sich u. a. in diesem Falle um eine Zeitreihenkorrelation handelt - vgl. Abschnitt 9.4.), zwischen der Höhe der Ausfallzeiten in einer Betriebsabteilung und deren Kapazitätsauslastung, zwischen der Höhe des Ausschusses und dem Betriebsergebnis, zwischen der Arbeitsproduktivität und den Selbstkosten eines Erzeugnisses, zwischen der Höhe des Einkommens von Personengruppen und der Höhe ihres Verbrauchs an bestimmten Konsumgütern usw.

### **5.3 Zur Regressionsanalyse**

Wir wollen uns jetzt der Regressionsrechnung zuwenden und - darauf aufbauend - im folgenden Abschnitt die Korrelation erläutern. Schauen wir uns zunächst drei Beispiele für die Anwendung der Regressionsanalyse an:

- In einem volkseigenen Betrieb wird ein bestimmtes Produkt in Serie hergestellt. Die Größe der Serie wirkt sich auf den Materialverbrauch aus. Diese Tatsache ist im Betrieb bekannt.

Man weiß aber nichts Genaues über die quantitative Abhängigkeit des Materialverbrauchs von der Seriengröße oder - anders ausgedrückt - über den Einfluss der Seriengröße ( $x_i$ ) auf die Menge des eingesetzten Materials ( $y_i$ ).

Dieser Zusammenhang kann mit Hilfe der Regressionsanalyse zahlenmäßig erfasst und geklärt werden.

Außerdem möchte der verantwortliche Abteilungsleiter gern wissen, wie die Wirtschaftlichkeit von Serien einzuschätzen ist, für die keine empirischen Werte vorliegen. Auch auf diese Frage kann mit Hilfe der Regressionsrechnung eine Antwort gegeben werden. Wir werden dieses Beispiel später noch berechnen.

- Die Vorbereitung für Rationalisierungsmaßnahmen und die Durchführung von Investitionen erfordern manchmal zum Nachweis der Wirtschaftlichkeit dieser Maßnahmen, die Reparaturkosten für vorhandene technische Einrichtungen in die Betrachtung mit einzubeziehen. Das ist auch im Verkehrswesen notwendig.

Zweifellos sind die Reparaturkosten eines Fahrzeugs u. a. von seinem Alter abhängig. Dieser Zusammenhang ist messbar, indem die Regression der Reparaturkosten ( $y_i$ ) bezüglich des Fahrzeugalters ( $x_i$ ) untersucht wird.

- Der Ernteertrag ( $y_i$ ) einer bestimmten Pflanzenart ist u. a. vom Einsatz künstlichen Düngers ( $x_i$ ) und von der Niederschlagsmenge ( $z_i$ ) abhängig. Damit besteht eine Regression von  $y$  bezüglich  $x$  und  $z$ .

Greifen wir die  $x$  und die  $y$  heraus und betrachten sie allgemein, so können wir erst einmal feststellen, dass die Werte  $x_i = x_1, x_2, \dots, x_n$  sowie  $y_i = y_1, y_2, \dots, y_n$  Realisationen der Zufallsvariablen  $X$  und  $Y$  sind.

In allen drei Fällen beziehen sich diese Zufallsvariablen bzw. ihre Realisationen auf Paare von Beobachtungen für jeweils ein Objekt. Für jede Serie liegen - um bei unseren Beispielen zu bleiben - die Angaben über den Materialverbrauch und die Seriengröße vor, und für jedes Fahrzeug ist das Alter und die Höhe der Reparaturkosten bekannt.

Für unsere Pflanzenart konnte der Ernteertrag zum einen mit dem Aufwand an künstlichem Dünger oder zum anderen mit der Niederschlagsmenge "gepaart" werden. Die aus den Beobachtungen hervorgegangenen Paare  $(x_1, y_1; x_2, y_2; \dots; x_n, y_n)$  ergeben also eine zweidimensionale Beobachtungs- oder Messreihe.

Die Verteilung der einen der beiden Zufallsvariablen - wir hatten uns schon für  $Y$  entschieden - wird für gegebene Werte der anderen Variablen, also  $X$ , analysiert. Durch unsere Entscheidung werden vor einer Berechnung die Zufallsvariable  $X$  und ihre Werte  $(x_1, x_2, \dots, x_n)$  als unabhängig festgelegt.

Dieses Arbeiten mit einer von vornherein festgelegten, einer unabhängigen Zufallsvariablen unterscheidet - das sei hier nochmals hervorgehoben - die Regressionsrechnung von der Korrelationsrechnung.

Bei der Regressionsanalyse wird also von den Verteilungsfunktionen

$$F(Y/x_1), F(Y/x_2), \dots, F(Y/x_n)$$

ausgegangen, wobei die  $x_1, x_2, \dots, x_n$  eben die Realisationen von  $X$  sind. Unter  $F(Y/x_i)$  ist die Verteilungsfunktion von  $Y$  bei gegebenem  $x_i$  zu verstehen, wobei  $Y$  als abhängige Zufallsvariable betrachtet wird, deren Verteilung eben von  $X$  abhängt.

Um den Zusammenhang zwischen den beiden Zufallsveränderlichen mit Hilfe einer Regressionsrechnung analysieren zu können, sind folgende Aufgaben zu erfüllen:

- Der Typ der Funktion ist zu bestimmen.
- Der Regressionskoeffizient ist zu ermitteln und zu prüfen.
- Die Funktion und deren Einzelwerte sind zu berechnen.
- Auf der Grundlage der Abweichungen der Funktionswerte von den empirischen Werten ist die Streuung zu untersuchen.

Ein Streuungsdiagramm mit den Beobachtungs- oder Messwerten bringt auf jeden Fall schon Hinweise, um welchen Funktionstyp es sich handelt. Ein solches Streuungsdiagramm entsteht, wenn wir jedes Wertpaar  $(x_1, y_1; x_2, y_2; \dots; x_n, y_n)$  jeweils als Punkt in ein Koordinatensystem mit den  $x_i$  auf der Abszisse und den  $y_i$  auf der Ordinate einzeichnen.

Dadurch erhalten wir einen Punkteschwarm oder eine Punktwolke, in die hinein eine Kurve, z. B. die Regressionsgerade, gelegt werden kann, wodurch der Funktionsverlauf und der Funktionstyp bereits angedeutet werden. Das Streuungsdiagramm liefert auch Anhaltspunkte über die Gegen- oder Gleichläufigkeit, über progressiven oder degressiven, positiven oder negativen Verlauf der Funktion. Schauen wir uns das im einzelnen anhand von Abbildung 20 (nächste Seite) an.

Die Streuungsdiagramme a), b) und c) in der Abbildung zeigen Gleichläufigkeit. In den Fällen d) und e) liegt Gegenläufigkeit vor. Es ist aber auch denkbar, dass der Zusammenhang zwischen Gleich- und Gegenläufigkeit wechselt.

Ebenso ist zu beachten, dass der Verlauf nicht nur gleichmäßig, sondern auch beschleunigt, also progressiv, bzw. verzögert, also degressiv, sein kann. Abbildung 20 zeigt in den



Streuungsdiagrammen g) und i) einen progressiven und in h) sowie j) einen degressiven Verlauf der Funktion.

Die Streuungsdiagramme k) und l) sind durch eine jeweils unterschiedliche Kombination dieser Arten gekennzeichnet, und zwar bringt k) einen positiv progressiven Verlauf zum Ausdruck, der an einem bestimmten Punkt in einen degressiv ansteigenden Auslauf überwechselt, während er bei l) negativ progressiv ist und dann ebenfalls degressiv weiterverläuft.

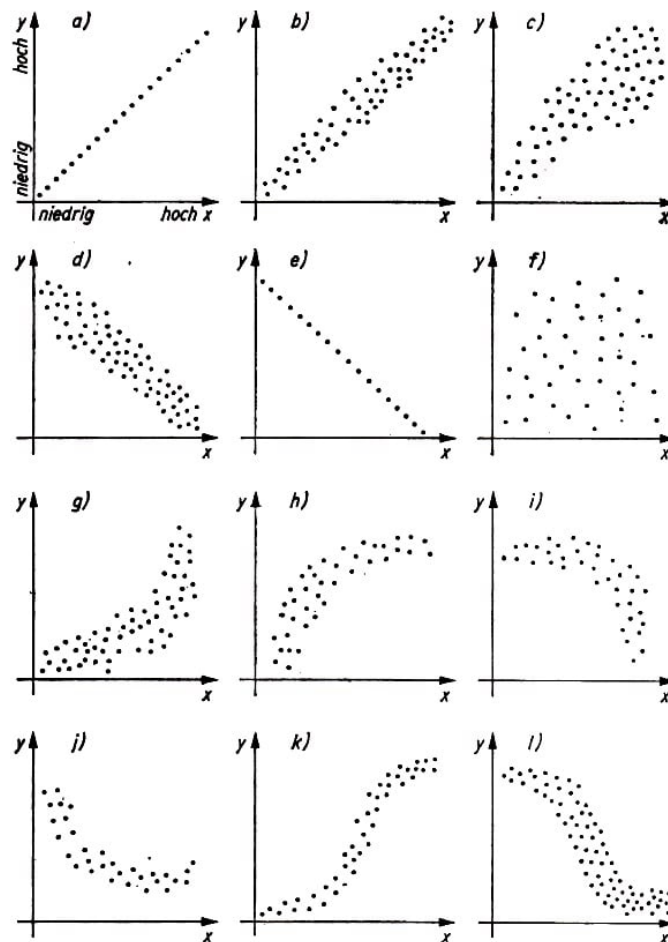


Abb. 20 Unterschiedliche Streuungsdiagramme

Die Streuungsdiagramme a) bis c) spiegeln graduell unterschiedliche positive lineare Zusammenhänge wider. Eine Zunahme von  $x$  bewirkt in diesen Fällen auch eine Zunahme von  $y$ .

Aus den Diagrammen d) und e) sind negative lineare Regressionen zu erkennen. Nimmt in diesen Fällen  $x$  zu, so bewirkt das eine Abnahme von  $y$ .

Das Bild f) zeigt keine oder eine sich nahe 0 bewegende Regression. Die Streuungsdiagramme g) bis j) deuten Formen eines nichtlinearen Verlaufs an, wobei g) und h) positive, i) und j) negative Regressionen ausdrücken.

Die beiden letzten Streuungsdiagramme k) und l) zeigen schließlich einen Verlauf, der sich aus unterschiedlichen Formen zusammensetzt.

Bei der Untersuchung derartiger Abhängigkeiten kommt es also darauf an, die Art des

jeweiligen Zusammenhanges zu ergründen. Insbesondere ist festzustellen,

- ob eine positive oder negative Regression vorliegt
- ob es sich um einen linearen oder nichtlinearen Zusammenhang handelt
- ob bei nichtlinearem Zusammenhang eine Progression oder eine Degression maßgebend ist.

Von einer positiven oder direkten Regression spricht man, wenn gleichzeitig die Werte der beiden Veränderlichen zunehmen oder sich vermindern. Weil die beiden Variablen sich in gleicher Richtung bewegen, wird der positive Zusammenhang auch als Gleichläufigkeit oder Parallelismus bezeichnet.

Wird z. B. mit Hilfe des wissenschaftlich-technischen Fortschritts das technische Produktionsniveau in einem Betrieb erhöht und steigt als Folge dessen die Arbeitsproduktivität, dann liegt ein solcher positiver Zusammenhang vor. Dasselbe trifft auf die Erhöhung der Kosten bei der Zunahme des Ausschusses zu.

Eine negative oder indirekte Regression ist gegeben, wenn die eine Zufallsvariable in die entgegengesetzte Richtung als die andere strebt. Man bezeichnet diese Erscheinung auch als Gegenläufigkeit oder Antagonismus. Als Beispiel kann hierfür die Steigerung der Arbeitsproduktivität genannt werden, die zu sinkenden Selbstkosten führt.

Je nachdem, ob zwischen den in die Analyse einbezogenen Variablen lineare oder nichtlineare Beziehungen bestehen, wird zwischen linearer bzw. nichtlinearer Regression unterschieden.

Entsprechend den empirischen Verteilungen können verschiedene Funktionstypen in Betracht kommen, z.B. die lineare Funktion, Exponentialfunktion, Potenzfunktion, logistische Funktion, Hyperbelfunktion, Parabelfunktion, logarithmische Funktion u. a. (vgl. hierzu Kap. 9).

Für die Regressionsrechnung leistet die von Gauß entwickelte Methode der kleinsten Quadratsumme gute Dienste, weil mit ihrer Hilfe die Parameter für die Regressionsfunktion aus den empirischen Angaben relativ gut geschätzt werden können.

Mit dieser Methode werden die Abweichungen bzw. Differenzen zwischen den empirischen Werten und den Funktionswerten für die Zufallsvariable  $Y$  bestimmt, wobei für die Abweichung

$$y_i - Y_i \quad (24)$$

gilt.

Wir wollen mit Hilfe dieser Abweichungen die Regressionsgerade bestimmen, die der Regressionsfunktion entspricht. Dabei gehen wir theoretisch von der Überlegung aus, dass zur exakten Fixierung der Regressionsgeraden die Summe der positiven Abweichungen der empirischen Werte mit der Summe der negativen Abweichungen übereinstimmen sollte.

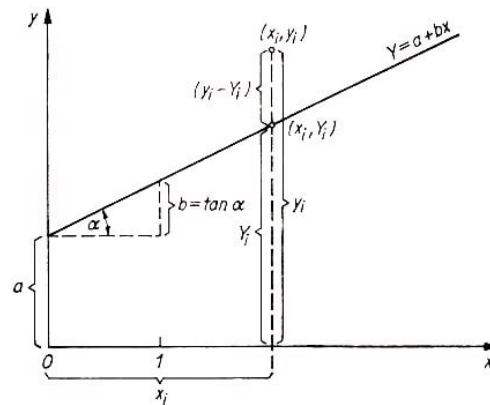


Abb. 21 Schema zur linearen Regression - Regressionsgerade und deren Parameter

Dann beträgt die Summe der Abweichungen aller empirischen Werte  $y_i$  von den theoretischen Werten  $Y_i$ , die sich - wie aus der Abbildung 21 ersichtlich ist - auf der Regressionsgeraden befinden, Null. Formelmäßig ausgedrückt bedeutet das:

$$\sum_{i=1}^n (y_i - Y_i) = 0 \quad (25)$$

Zur eindeutigen Bestimmung unserer Regressionsgeraden reicht diese Bedingung aber noch nicht aus. Es ist deshalb notwendig, eine weitere Forderung zu stellen: Wenn die Summe der Quadrate der Abweichungen zwischen den beobachteten  $y$ -Werten und den Werten der Regressionskurve  $Y_i$  ein Minimum beträgt, würde die Regressionsfunktion den gegebenen Verhältnissen entsprechend am besten abgeschätzt. Mit anderen Worten:

Es wird für die empirischen Werte diejenige Funktion gesucht, für die die Summe der Abweichungsquadrate der empirischen Werte  $y_i$  von den Funktionswerten  $Y_i$  - diese Summe wird hier mit  $A_y$  bezeichnet - ein Minimum ist, d.h.

$$A_y = \sum_{i=1}^n (y_i - Y_i)^2 = \text{Minimum} \quad (26)$$

Bei dieser Methode der kleinsten Quadratsumme wird vorausgesetzt, dass die zufälligen Abweichungen voneinander unabhängig sind und eine Normalverteilung gegeben ist. (Die Normalverteilung wird im Abschnitt 6.4. erläutert.)

Die Methode der kleinsten Quadratsumme hilft uns also, das Wesentliche des zu analysierenden Zusammenhangs herauszuarbeiten und ihn von den auf Nebenursachen beruhenden Störungen zu befreien. Da diese Methode die minimale Summe der quadratischen Abweichungen fordert, gewährleistet sie, dass sich die ermittelte Funktion von allen Funktionen dieser Art bestmöglich dem Verlaufe der empirischen Werte anpasst.

Diese Methode der kleinsten Quadratsumme oder kurz der kleinsten Quadrate gleicht alle vorliegenden empirischen Werte unter den gegebenen Bedingungen zu einer Regressionskurve aus.

Sind zwei Zufallsvariablen vorhanden, von denen die eine die Ursache und die andere

die Wirkung ist, handelt es sich um eine einfache Regression. Bestehen lineare Beziehungen, wie sie die Abbildung 21 veranschaulichen soll, so kann man die Funktion für die Regressionsgerade durch die allgemeine Gleichung einer Geraden auf der Grundlage der Bestimmung von zwei Punkten durch

$$Y = a + bx \quad (27)$$

charakterisieren.

In diesem Fall wird das Modell der linearen Regression zugrunde gelegt, bei dem von der Annahme ausgegangen wird, dass jeder Wert  $y_i$  eine Realisation der normalverteilten Zufallsvariablen  $Y$  mit der konstanten quadratischen Streuung  $\sigma^2$  und dem Mittelwert  $\alpha + \beta x_i$  ist.

Jedem der vorgegebenen  $x_i$  entspricht eine normalverteilte Gesamtheit der  $y_i$  und jedem  $x_i$  ein Mittelwert  $Y_i = f(x_i)$ . Wir nehmen noch an, dass die quadratische Streuung für jede unserer Verteilungen gleich groß ist und die Mittelwerte von  $Y_i(\bar{Y})$  auf der in Abbildung 21 eingezeichneten Geraden liegen, wobei für diesen Mittelwert gilt:

$$\bar{Y} = f(x) = \alpha + \beta x \quad (28)$$

Auf die Bedeutung von Parametern der Grundgesamtheit bzw. von theoretischen Verteilungen, die mit griechischen Buchstaben gekennzeichnet werden, wurde im 4. Kapitel schon hingewiesen.

Darüber hinaus wird in den Kapiteln 6 bis 8 dieses Buches noch näher darauf eingegangen.

Wir suchen  $A_y$  (die Summe der quadratischen Abweichungen), die bei gegebenen  $x_i$  und  $y_i$  von den Parametern  $a$  und  $b$  abhängt. Die Summe  $A_y$  erhöht oder vermindert sich in Abhängigkeit von der Größe dieser Parameter  $a$  und  $b$ . Damit ist erkennbar, dass eine Funktion von zwei Veränderlichen vorliegt. Wir können also auch formulieren

$$A_y = f(a, b) \quad (29)$$

Diese Funktion soll minimiert werden.

Von den vorstehenden theoretischen Betrachtungen ausgehend, gelangt man mit Hilfe der Differentialrechnung zu Gleichungen, die als "normale Regressionsgleichungen" oder kurz "Normalgleichungen" für die einfache lineare Regression bezeichnet werden, und zwar zu

$$\sum_{i=1}^n y_i = n \cdot a + b \sum_{i=1}^n x_i \quad (30)$$

$$\sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 \quad (31)$$

In diesen Gleichungen sind uns die  $x_i$  und die  $y_i$  schon bekannt;  $a$  und  $b$  sind Parameter oder Koeffizienten der Regressionsfunktion, d.h.

$a$ : Größe, die zusammenfassend den Einfluss aller weiteren Ursachen (außer der Ursache

$x$ ) auf die Wirkung ausdrückt und als Konstante den Schnittpunkt der Regressionsgeraden mit der Ordinate angibt.

$b$ : linearer Regressionskoeffizient, der die Veränderung der Größe  $y$  (Wirkung) charakterisiert, wenn die Ursache  $x$  um eine Einheit zu- oder abnimmt. Er bestimmt den Verlauf der Regressionsgeraden.

Bei einer zweidimensionalen Verteilung entspricht die Regressionsgerade der Bestimmung des Mittelwertes einer eindimensionalen Verteilung. Die Neigung (der Anstieg) der Regressionsgeraden wird durch den Regressionskoeffizienten zum Ausdruck gebracht. Ist der lineare Regressionskoeffizient positiv, so führt er zum Ansteigen der Geraden; ist er negativ, so fällt die Gerade ab. Anders ausgedrückt:

Ist der lineare Regressionskoeffizient positiv, so nimmt der Betrag der Funktion mit wachsendem  $x$  zu, während beim negativen Regressionskoeffizienten mit zunehmender Größe von  $x$  der Betrag der Funktion abnimmt.

Charakterisieren wir die Regression mit dem Regressionskoeffizienten  $b$ , so können wir für den Regressionskoeffizienten der Regression von  $y$  bezüglich  $x$  auch schreiben  $b_{y \cdot x}$ . Wir kennzeichnen damit eine Regressionsgerade. Stets gibt es aber theoretisch noch eine zweite Regressionsgerade. Was bringt sie zum Ausdruck ?

In der Praxis gibt es Fälle, in denen die Ursache-Wirkung-Beziehung nicht immer eindeutig ist. Nehmen wir z. B. den Zusammenhang zwischen der Körpergröße und dem Körpergewicht jedes Angehörigen einer Personengruppe oder zwischen dem Druck und dem Volumen eines Gases bei konstanter Temperatur oder zwischen dem eingesetzten Grundfonds und dem Produktionsvolumen, so wäre hier ein Vertauschen von Ursache und Wirkung denkbar.

Auf den ersten Blick mag man geneigt sein, zu entscheiden, dass der Umfang des eingesetzten Grundfonds ( $x$ ) die Ursache für die Höhe des Produktionsvolumens ( $y$ ) ist. Zweifellos ist das prinzipiell richtig. Bei der Planung steht allerdings auch die Frage: In welchem Umfang müssen Grundfonds eingesetzt werden, um ein bestimmtes Produktionsvolumen zu erhalten ? Damit wird die Frage nach der Abhängigkeit des Umfangs der Grundfonds ( $x$ ) von der Höhe des Produktionsvolumens ( $y$ ) gestellt.

Dann interessiert also die Regression von  $x$  bezüglich  $y$ , so dass der Regressionskoeffizient  $b_{x \cdot y}$  gefragt ist. Damit ist  $x$  die abhängige Veränderliche und  $y$  die unabhängige. Der Regressionskoeffizient  $b_{x \cdot y}$  entspricht der zweiten Regressionsgeraden.

Betrachten wir noch kurz den Zusammenhang zwischen Körpergröße und Körpergewicht. Gehen wir zunächst davon aus, dass mit zunehmender Körperlänge  $x$  sich auch das Körpergewicht  $y$  erhöht, so liegt hier der beschriebene Fall  $b_{y \cdot x}$  vor. Wir könnten aber auch für eine Personengruppe berechnen, um wieviel Zentimeter sich die Körperlänge im Durchschnitt erhöht, wenn das Gewicht um eine bestimmte Anzahl von Kilogramm zunimmt. Wir erhalten dann wiederum eine zweite Regressionsgerade, die dem Regressionskoeffizienten  $b_{x \cdot y}$  entspricht, der die Regression von  $x$  bezüglich  $y$  zum Ausdruck bringt.

Bei allen derartigen Beispielen ist von der Logik des Zusammenhangs auszugehen. Bei dem hier erörterten Beispiel für den Zusammenhang zwischen der Körperlänge und dem Körpergewicht wäre wahrscheinlich der Regressionskoeffizient  $b_{x \cdot y}$  weniger sinnvoll. Wir würden deshalb  $x$  als unabhängige Veränderliche bestimmen. Trotzdem bestehen in allen Fällen theoretisch zwei Regressionskoeffizienten und damit zwei Regressionsgeraden.

Wir wollen uns in den folgenden Ausführungen jedoch nur mit dem Regressionskoeffizienten  $b_{y \cdot x}$  beschäftigen. Die beiden erwähnten Normalgleichungen unserer Regression gelten als Eckpfeiler zur Bestimmung der Regressionskurve.

Trotzdem sind uns  $a$  und  $b$  immer noch unbekannt. Wir könnten sie einmal mit Hilfe von Determinanten und der sogenannten "Cramerschen Regel" bestimmen. Neben dieser Möglichkeit wird in der statistischen Fachliteratur jedoch oft der folgende und für praktische Berechnungen auch rationellere Weg beschrieben: Ausgehend von der ersten Normalgleichung (30), erhalten wir nach Division durch  $n$

$$a + \frac{b \sum_{i=1}^n x_i}{n} = \frac{\sum_{i=1}^n y_i}{n} \quad (32)$$

$$a + b\bar{x} = \bar{y} \quad (33)$$

Für  $a$  kann man hieraus folgern:

$$a = \bar{y} - b\bar{x} \quad (34)$$

Damit können wir für (27) auch schreiben

$$Y = \bar{y} + b(x - \bar{x}) \quad (35)$$

Der Regressionskoeffizient  $b$  kann durch Einsetzen von  $a$  in die zweite Normalgleichung (31) gewonnen werden. Diese wird dann nach  $b$  aufgelöst. Weiterhin sind auch noch folgende Überlegungen bedeutsam: Nach der Division einer mit Hilfe von Determinanten gewonnenen Formel durch  $n$  erhält man

$$b = \frac{\sum_{i=1}^n y_i x_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} \quad (36)$$

Da

$$\frac{\sum_{i=1}^n x_i}{n} = \bar{x} \quad \text{sowie} \quad \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$$

ist, können wir auch schreiben

$$b = \frac{\sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i} \quad (37)$$

oder

$$b = \frac{\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i} \quad \text{oder} \quad b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (38,39)$$

Multiplizieren wir die Klammern in dieser Formel aus und formen sie um, so ergibt sich die Übereinstimmung der Formel (39) mit den Formeln (37) und (38). Wenn in der Formel (39) Zähler und Nenner durch  $n - 1$  dividiert werden, so entsteht im Zähler die Kovarianz von  $x$  und  $y$  sowie im Nenner die Varianz von  $x$ . Daraus ist erkennbar, dass der Regressionskoeffizient  $b$  auch mit Hilfe der Kovarianz und der Varianz von  $x$  bestimmt werden kann als

$$b = \frac{s_{xy}}{s_x^2} \quad (40)$$

Der Regressionskoeffizient geht also aus der Gegenüberstellung von Varianzen hervor. Im Abschnitt 4.4. war bereits erläutert worden, was unter der Varianz zu verstehen ist. Die Kovarianz ist die Produktsumme aus den Abweichungen der Werte von zwei statistischen Verteilungen oder Reihen und ergibt sich als "Summe der für die sich entsprechenden Werte zweier Messreihen gebildeten Produkte ihrer linearen Abweichungen von ihrem arithmetischen Mittel, dividiert durch die um Eins verminderte Anzahl der Einzelwertepaare".<sup>25</sup>

Die Kovarianz bildet beim linearen Regressionskoeffizienten den Zähler, und zwar allgemein nach der Formel

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n-1} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n-1} \quad (41)$$

Aus den empirischen Wertepaaren  $x_i$  und  $y_i$  können zur Bestimmung der Kovarianz die arithmetischen Mittel  $\bar{x}$  und  $\bar{y}$  gebildet werden. Unter der Annahme von  $h_i = 1$ , wobei  $h_i$  absolute Häufigkeiten sind, ergeben sich als Varianz von  $x$

$$s_x^2 = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) \quad (42)$$

und als Varianz von  $y$

$$s_y^2 = \frac{1}{n-1} \left( \sum_{i=1}^n y_i^2 - \bar{y} \sum_{i=1}^n y_i \right) \quad (43)$$

Die Formeln (42) und (43) gehen dabei aus der Formel (19) im Abschnitt 4.4. hervor.

Um die Anpassung verschiedener Formen der Regressionsfunktion an die empirischen Wertepaare, also die Zuverlässigkeit einer Regressionsschätzung, beurteilen zu können, ist es vorteilhaft, den Standardfehler der Regressionsschätzung (45) - auch als Standardabweichung um die Regression bezeichnet - sowie die Varianz um die Regression

<sup>25</sup>DDR-Standard, „Statistische Qualitätskontrolle - Begriffe“. TGL 14449, Blatt 2, S. 18

(44) zu berechnen. Wir verstehen hierbei unter der Varianz die durchschnittliche Abweichung der empirischen Werte von den Funktionswerten:

$$s_{xy}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - Y_i)^2 \quad , \quad s_{xy} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - Y_i)^2} \quad (44,45)$$

Die ein Minimum darstellende Summe der Quadrate der Abweichungen der  $y_i$ -Werte von den Funktionswerten  $Y_i$ , die sich auf der Regressionsgeraden befinden, steht im Zähler der Formel (44).

Im Nenner dieser Formel taucht  $n - 2$  auf. Der Leser wird sich jetzt wahrscheinlich fragen, warum hier nicht auch  $n - 1$  steht. Das hängt mit den schon im Abschnitt 4.4. erläuterten Freiheitsgraden zusammen. Wir erkannten bereits, dass die Schätzungen von  $a$  und  $b$  von den Mess- bzw. Beobachtungswerten  $y_i$  abhängig sind. Das ist gleichbedeutend mit zwei linearen "Einschränkungen" der Beobachtungs- oder Messreihe.

Die  $\sum_{i=1}^n (y_i - Y_i)^2$  kann als Summe der Quadrate von  $n - 2$  unabhängigen linearen Kombinationen, die den Mittelwert 0 und die Varianz  $\sigma^2$  haben, betrachtet werden, so dass eine Verteilung mit zwei Freiheitsgraden vorliegt, bei der gilt:

$$\frac{(n-2)s_{yx}^2}{\sigma^2} \quad (46)$$

In Verbindung mit dem Prüfen von Verteilungen wird auf die Freiheitsgrade im 8. Kapitel nochmals eingegangen.

Seriengröße in 1000 Stück	Materialverbrauch je 1000 Stück in kg		
$x_i$	$y_i$	$x_i \cdot y_i$	$x_i^2$
20	189	3780	400
25	186	4650	625
30	191	5730	900
50	183	9150	2500
60	179	10740	3600
65	171	11115	4225
70	174	12180	4900
75	170	12750	5625
395	1443	70095	22775
$\bar{x} = 49,375$	$\bar{y} = 180,375$		

Tabelle 10 Materialverbrauch und Seriengröße - Rechentabelle für die Regressionsrechnung

Wir wollen jetzt das am Beginn dieses Unterabschnittes über die Regressionsrechnung erwähnte Beispiel der Untersuchung der Abhängigkeit des Materialverbrauchs von der Seriengröße etwas näher betrachten. Bei dieser Aufgabe geht es darum, den Einfluss der Seriengröße ( $x_i$ ) auf die Menge des verbrauchten Materials ( $y_i$ ) mit Hilfe der Regressionsrechnung zu untersuchen.



In den ersten beiden Spalten von Tabelle 10 sind die Ausgangsdaten für unsere Aufgabe enthalten, und zwar der Materialverbrauch ( $y_i$ ) je 1000 Stück in kg sowie die Seriengröße ( $x_i$ ), ebenfalls in 1000 Stück. Bei den Stückzahlen handelt es sich um die herzustellenden Produkte.

Aus dem Verlauf der beiden Reihen in der Tabelle 10 ist eine Gegenläufigkeit ersichtlich. Die grafische Darstellung dieser Zahlenangaben (vgl. Abb. 22) lässt erkennen, dass es sich um einen negativen linearen Verlauf handelt.

Da zwei Variablen vorhanden sind, haben wir es mit einer einfachen und nicht mit einer mehrfachen linearen Regression zu tun. Demnach ist von einer einfachen linearen Regressionsfunktion auszugehen.

Sie wird durch die zunächst empirisch in Abbildung 22 eingezeichnete Regressionsgerade charakterisiert.

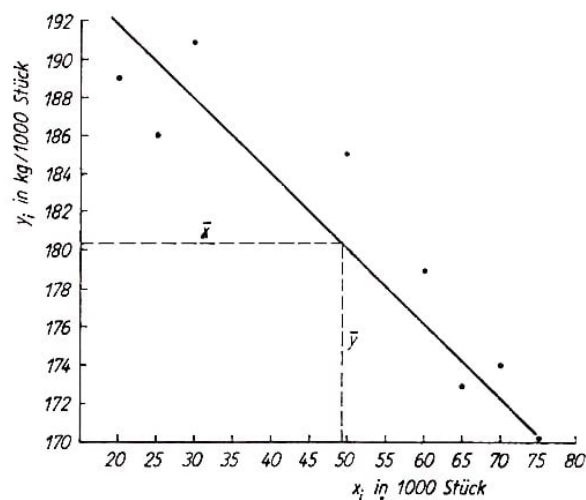


Abb. 22 Streuungsdiagramm deren Parameter

Ebenso wurden die Mittelwerte  $\bar{x}$  und  $\bar{y}$  aus der Tabelle 10 übernommen und eingezeichnet.

Sowohl die Regressionsfunktion als auch der Regressionskoeffizient sollen für unser Beispiel berechnet werden. Rechnerisch kann die Regressionsfunktion, die in der Regressionsgeraden zum Ausdruck kommt, mit Hilfe der Methode der kleinsten Quadratsumme aus den empirischen Werten ermittelt werden.

Die zur Bestimmung der Konstanten  $a$  und des Regressionskoeffizienten  $b$  notwendigen Werte werden nach Tabelle 10 gewonnen. Wir erhalten damit zunächst

$$\sum_{i=1}^n x_i, \quad \sum_{i=1}^n y_i, \quad \sum_{i=1}^n x_i y_i \quad \text{und} \quad \sum_{i=1}^n x_i^2$$

$\bar{x}$  und  $\bar{y}$  werden ermittelt als:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{395}{8} = 49,375 \quad , \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{1443}{8} = 180,375$$

Mit diesen Werten können  $a$  und  $b$  bestimmt werden, wobei wir mit  $b$  beginnen, da dieser Wert zur Berechnung von  $a$  benötigt wird. Nach der Formel (39) erhalten wir zunächst für  $b$ :

$$b = \frac{70095 - 180,375 \cdot 395}{22775 - 49,375 \cdot 1443} = -0,3524$$

Nachdem  $b$  berechnet wurde, kann dieser Wert in die Formel für  $a$  (34) eingesetzt werden. Somit ergibt sich

$$a = 180,375 - (-0,3524) \cdot 49,375 = 197,7765$$

Die Regressionsfunktion lautet demnach:

$$Y = 197,7765 - 0,3524x_i$$

$Y_1, Y_2, \dots, Y_8$  erhalten wir, indem für  $x_1 = 20, x_2 = 25, \dots, x_8 = 75$  gesetzt wird, d.h.

$$Y_1 = 197,77 - 0,3524 \cdot 20 = 190,7277 \text{ kg/1000 Stück usw.}$$

Diese Funktionswerte sind in Tabelle 11 enthalten, die jedoch erst im nächsten Abschnitt erörtert werden soll.

Seriengröße in 1000 Stück	Materialverbrauch je 1000 Stück in kg	Funktionswerte		
$x_i$	$y_i$	$Y_i$	$(y_i - Y_i)^2$	$(y_i - \bar{y})^2$
20	189	190,73	2,985	74,39
25	186	188,96	8,795	31,64
30	191	187,20	14,414	112,89
50	183	180,15	8,096	6,89
60	179	176,63	5,615	1,89
65	171	174,87	14,962	87,89
70	174	173,11	0,799	40,64
75	170	171,34	1,806	107,64
395	1443	1442,99	57,472	463,87
$\bar{x} = 49,375$	$\bar{y} = 180,375$			

Tabelle 11 Rechentabelle zur Ermittlung des Maßes der Bestimmtheit bzw. Unbestimmtheit

Der konstante, von der Seriengröße nicht abhängige Materialverbrauch beträgt pro 1000 Stück 197,776 kg (Wert von  $a$ ).

Wird die Seriengröße jeweils um 1000 Stück erhöht, dann sinkt der Verbrauch an Material jeweils um 0,325 kg (Wert von  $b$ ). Dieses Sinken kommt durch das negative Vorzeichen des Regressionskoeffizienten zum Ausdruck.

Mit Hilfe der berechneten Werte wird noch folgendes ermöglicht:

- Innerhalb der Variationsbreite von  $x$  kann durch Interpolation für die nicht vorliegenden Werte der Seriengröße (z. B. für 40000 Stück) berechnet werden, wie groß der Materialverbrauch ist.

Für außerhalb der Variationsbreite liegende Werte der Seriengröße, z. B. für 85000 Stück, könnte ebenfalls ermittelt werden, wie hoch der Materialverbrauch wäre. Dabei müsste allerdings vorausgesetzt werden, dass die Seriengröße  $x$  als hauptsächlicher Beeinflussungsfaktor ihre Stellung im Ursachenkomplex nicht verändert und auch weiterhin die Hauptursache für die Menge des verbrauchten Materials ist und dass die Form des Zusammenhangs (negative Linearität und Proportionalität zwischen der Veränderung von  $x$  und  $y$ ) ebenfalls über den Wert von 75000 Stück hinaus bestehen bleibt (Extrapolation).

Im allgemeinen besitzt die Regressionsfunktion nur Gültigkeit im Bereich der empirischen Wertepaare.

Unter der Bestimmung der Art oder der Form des Zusammenhangs versteht man also die Gewinnung der Regressionsfunktion aus dem empirischen Zahlenmaterial. Man spricht von der einfachen Form des Zusammenhangs, wenn sich die Verhaltensweise der Wirkung durch die Veränderung einer der sie beeinflussenden Ursachen wandelt, und von der komplexen Form des Zusammenhangs, wenn sich die Verhaltensweise der Wirkung auf Grund der Veränderung mehrerer der sie beeinflussenden Ursachen wandelt. Wir wollen damit zur Korrelation überleiten.

## 5.4 Einfache, multiple und partielle Korrelation

Unter Beachtung der Zahl der Zufallsvariablen, die in eine solche Untersuchung einbezogen werden, ist zu unterscheiden zwischen der einfachen Korrelation, wenn zwei Veränderliche zueinander in Beziehung gesetzt werden, und der mehrfachen oder multiplen Korrelation, wenn mehr als zwei Variablen einbezogen werden.

Beispiele für einfache Korrelationen wurden schon im Abschnitt 5.2. angeführt. Eine mehrfache Korrelation liegt vor, wenn z. B. der Zusammenhang zwischen dem Mechanisierungsgrad, dem Stand der Arbeitsproduktivität und der Qualifikation der Werktätigen oder der Außentemperatur, dem Verbrauch an Heizmaterial und der Innentemperatur bestimmt werden soll.

Diese Beispiele zeigen zugleich, dass die multiple Korrelation mit der Wiedergabe der erwähnten objektiv bestehenden stochastischen Mehrfachbeziehungen die Erfassung eines größeren Ursache-Wirkung-Komplexes ermöglicht.

Will man die innere Beziehungsstruktur bei der Untersuchung eines solchen größeren Ursache-Wirkung-Komplexes analysieren, z. B. zwischen zwei Veränderlichen bei gleichzeitiger Ausschaltung der Einflüsse der dritten und - soweit vorhanden - der vierten Variablen, dann handelt es sich um eine partielle Korrelation.

Auf den verschiedensten Gebieten treten solche Fälle auf, bei denen mehrere Ursachen gleichzeitig wirken und gemeinsam zu einer Folge führen. Um den Grad des Einflusses einzelner Faktoren auf das Gesamtergebnis zu kennen und eventuelle Veränderungen vornehmen zu können, ist es notwendig, zu wissen, wie sich jede Veränderliche, jeder Faktor auswirkt.

Zunächst wollen wir uns mit dem Bestimmtheitsmaß und dem einfachen Korrelations-

koeffizienten vertraut machen. Dieser Koeffizient wird deshalb als einfach bezeichnet, weil es sich um eine einfache, nur zwischen zwei Variablen bestehende Korrelation und nicht um eine mehrfache handelt.

Das Bestimmtheitsmaß, eine seiner Dimension nach quadratische Kennziffer, wird - ausgehend von der Streuung der Funktionswerte - um den arithmetischen Mittelwert gebildet, wobei im Nenner die Gesamtstreuung enthalten ist.

Sie setzt sich zusammen aus der Streuung der Einzelwerte von  $y$  um die Funktionswerte und der Streuung der Funktionswerte um den arithmetischen Mittelwert. Die Formel (47) gibt diese Beziehung wieder:

$$B = \frac{\sum_{i=1}^n (Y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (47)$$

wobei  $0 \leq B \leq 1$  gilt. (48)

Ermittelt man dieses Maß auf der Grundlage der Streuung der Einzelwerte um die Funktionswerte - im Gegensatz zur vorstehenden Formel (47) -, so ergibt sich das Maß der Unbestimmtheit nach Formel (49):

$$U = \frac{\sum_{i=1}^n (y_i - Y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (49)$$

Die Werte beider Maße ergeben zusammen 1, d.h.

$$B + U = 1 \quad (50)$$

Diese Tatsache gestattet es,  $B$  als Differenz von  $1 - U$  zu ermitteln, also

$$B = 1 - U \quad \text{bzw.} \quad B = 1 - \frac{\sum_{i=1}^n (y_i - Y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (51)$$

Ziehen wir aus diesem Ausdruck die positive Wurzel, so erhalten wir den allgemeinen einfachen Korrelationskoeffizienten  $R$ :

$$R = \sqrt{B} \quad , \quad R = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - Y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (52,53)$$

Da dieser Korrelationskoeffizient die Wurzel aus dem Bestimmtheitsmaß darstellt, ist offensichtlich, dass er in enger Beziehung zu  $B$  steht.

Wie aus (48) hervorgeht, kann das Bestimmtheitsmaß nur positive Werte annehmen. Um auch die Form des Zusammenhangs zum Ausdruck zu bringen, erhält deshalb der mit Hilfe von  $B$  ermittelte Korrelationskoeffizient  $R$  das Vorzeichen des Regressionskoeffizienten.

Das Bestimmtheitsmaß kann zur Prüfung und Beurteilung linearer und nichtlinearer Zusammenhänge genutzt werden. Es wird insbesondere dann verwendet, wenn einerseits aus dem vorhandenen empirischen Zahlenmaterial der Funktionstyp nicht eindeutig bestimmt werden kann, aber andererseits eine weitgehende Anpassung gewünscht wird. In diesem Falle würde der Zusammenhang mit den in Betracht kommenden Funktionstypen berechnet, jeweils das Bestimmtheitsmaß ermittelt und die gewonnenen Werte des Bestimmtheitsmaßes miteinander verglichen. Jener Funktionstyp ist als der angemessenste anzusehen, bei dem das Bestimmtheitsmaß dem Wert 1 am nächsten kommt.

In der praktischen statistischen Arbeit ist jedoch der Korrelationskoeffizient häufiger anzutreffen als das Bestimmtheitsmaß.

Der erwähnte Korrelationskoeffizient  $R$  gilt allgemein. Er ist für die lineare wie für die nichtlineare Korrelation anwendbar. Den speziellen linearen Korrelationskoeffizienten wollen wir mit  $r$  symbolisieren. Er geht aus den nachstehenden Formeln hervor:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (54)$$

oder aus

$$B = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} \quad \text{mit} \quad r = \sqrt{B} \quad \text{wird} \quad r = \frac{s_{xy}}{s_x \cdot s_y} \quad (55,56,57)$$

Das Vorzeichen für  $r$  entspricht dem des Regressionskoeffizienten  $b$  bzw. ergibt sich in Formel (54) aus dem Zähler von  $r$ .

Für den Korrelationskoeffizienten  $r$  gilt allgemein:

$$-1 \leq r \leq +1 \quad (58)$$

Bei einem Wert von nahe -1 liegen stark ausgeprägte gegensätzliche Beziehungen, also eine negative Korrelation vor (hohen  $x_i$  entsprechen niedrige  $y_i$ ). Bewegt sich der Korrelationskoeffizient nahe +1, so zeigt sich eine stark ausgeprägte positive Verbindung zwischen beiden Variablen (hohen  $x_i$  entsprechen hohe  $y_i$ ). Das Vorzeichen des Korrelationskoeffizienten drückt die Richtung des Zusammenhangs aus.

Die Stärke oder Strenge des Zusammenhangs kommt dagegen in der Größe von  $r$  zum Ausdruck. Je näher der Koeffizient bei 1 liegt, um so stärker ist der Zusammenhang. Bei einem Korrelationskoeffizienten von  $r = 0$  oder nahe Null kann auf Grund des vorliegenden Zahlenmaterials keine Korrelation nachgewiesen werden.

Damit kann aber nicht unbedingt behauptet werden, dass zwischen  $x$  und  $y$  keine Korrelation besteht, zumal am Anfang einer Korrelationsanalyse eine Prüfung des logischen Zusammenhangs stehen soll.

Wie die vorstehenden Darlegungen auch erkennen lassen, sind verschiedene Korrelationskoeffizienten entwickelt worden, die unter bestimmten Umständen jeweils Vor- und

Nachteile aufweisen.

Für das im Abschnitt 5.3. durchgerechnete Beispiel fertigen wir zur Berechnung des Bestimmtheitsmaßes und des allgemeinen Korrelationskoeffizienten  $R$  wieder eine Rechentabelle an (s. Tab. 11).

Danach ergeben sich gemäß den Formeln (51), (49) und (52):

$$B = 1 - \frac{57,472}{463,87} = 0,876$$

$$U = \frac{57,472}{463,87} = 0,124$$

$$R = \sqrt{0,876} = 0,946 \hat{=} -93,60\%$$

Wir wollen prüfen, ob für unser Beispiel beide Korrelationskoeffizienten übereinstimmen, und berechnen deshalb nach dem allgemeinen noch den linearen Korrelationskoeffizienten  $r$ . Hierzu fertigen wir wieder eine Rechentabelle (Tab. 12) an, für die ein Teil der Angaben aus den vorangegangenen Rechentabellen 10 und 11 übernommen werden kann.

Seriengröße	Materialverbrauch					
$x_i$	$y_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
20	189	8,625	74,39	-29,375	862,89	-253,3593
25	186	5,625	31,64	-24,375	594,14	-137,1093
30	191	10,625	112,89	-19,375	375,39	-205,8593
50	183	2,625	6,89	0,625	0,39	1,6406
60	179	-1,375	1,89	10,625	112,89	-14,6093
65	171	-9,375	87,89	15,625	244,14	-146,4843
70	174	-6,375	40,64	20,625	425,39	-131,4843
75	170	-10,375	107,64	25,625	656,64	-265,8593
395	1443	0	463,87	0	3271,87	-1153,1235
$\bar{x} = 49,375$	$\bar{y} = 180,375$					

Tabelle 12 Rechentabelle zur Bestimmung des Korrelationskoeffizienten

Auf der Grundlage der Tabelle 12 ergibt sich nach Formel (54)

$$r = \frac{-1153,12}{\sqrt{3271,87 \cdot 463,87}} = -0,9360 \hat{=} -93,60\%$$

Die Übereinstimmung beider Korrelationskoeffizienten ist somit gegeben.

## 5.5 Rangkorrelation

Bisher wurde die Korrelation unter dem Aspekt der Berechnung quantitativer Zusammenhänge auf der Grundlage von Beobachtungs- bzw. Messwerten betrachtet. Ein Ziel war es dabei, den Korrelationskoeffizienten als Maß dieses Zusammenhangs zu berechnen. In der Praxis treten aber auch Fälle auf, in denen keine Messwerte, sondern Rangplätze für die einzelnen Elemente vorliegen oder bei denen qualitative Merkmale hinsichtlich ihres Zusammenhangs zu beurteilen sind (vgl. Kap. 2).

		Merkmalsvariation bzw. Skalierung der Zufallsvariablen		
		$X$		
		nominal	ordinal	metrisch, in Intervallen
$Y$	nominal	Kontingenzkoeffizient Koeffizient $\Phi$ tetrachorischer Koeffizient Assoziationskoeffizient		
	ordinal	Rangkorrelationskoeffizient		
	metrisch in Intervallen	Maßkorrelationskoeffizient		

Tabelle 13 Übersicht über Koeffizienten zur Prüfung des Zusammenhangs zwischen zwei Zufallsvariablen in Abhängigkeit von der Merkmalsvariation bzw. Skalierung der Zufallsvariablen

Je nachdem in welcher Form die Daten für eine Untersuchung der Abhängigkeit zwischen zwei Zufallsvariablen vorliegen, ist derjenige Koeffizient auszuwählen, der dieser Form der Daten angemessen ist. Dabei kann die Tabelle 13 eine Hilfe sein.<sup>26</sup>

Sie gibt einen Überblick über verschiedenartige Koeffizienten zur Charakterisierung von Zusammenhängen, von Abhängigkeiten entsprechend der Merkmalsvariation bzw. der Skalierung der Zufallsvariablen (nominal, ordinal, metrisch oder in Intervallen) und ist damit zugleich eine Konkretisierung der letzten Zeile der Tabelle 4.

Den Zusammenhang zwischen zwei ordinalen Verteilungen, z. B. zwischen zwei Reihen von Rangplätzen, berechnet man mit Hilfe des Rangkorrelationskoeffizienten. Es handelt sich dabei um einen Sonderfall der Maßkorrelation, bei dem die Variablen  $X$  und  $Y$  nur durch die ersten positiven Zahlen der Zahlenreihe  $1...n$  gebildet werden. Der Rangkorrelationskoeffizient wird dort berechnet, wo Sachverhalte durch Rangzahlen ausgedrückt werden müssen, weil keine Möglichkeit zur Aufstellung einer Messreihe besteht.

Wenn innerhalb der Verteilungen jede Rangzahl nur einmal vorkommt, dann sind die Rangzahlen mit den Ordnungszahlen der Einheiten identisch. Haben zwei Einheiten einer ordinalen Verteilung dieselbe Rangzahl erhalten, so bestimmt man das arithmetische Mittel aus den Ordnungszahlen der Einheiten mit derselben Rangzahl und verwendet diesen Mittelwert als Rangzahl.

Haben mehr als zwei Einheiten einer ordinalen Verteilung dieselbe Rangzahl, kann man auch mit dem Mittelwert der Ordnungszahlen arbeiten. Da das aber manchmal etwas umständlich ist, empfehlen die Autoren einschlägiger Statistik-Fachbücher die Benutzung spezieller Formeln, deren Behandlung hier jedoch zu weit führen würde.

Der Rangkorrelationskoeffizient liegt ebenso wie der schon erörterte Maßkorrelationskoeffizient zwischen  $-1$  und  $+1$ . Er wird - wie wir an dem noch zu berechnenden Beispiel im einzelnen sehen werden - unter Verwendung der ins Quadrat gesetzten Differenzen der Rangverteilung für die  $x_i$  und die  $y_i$  ermittelt. Sind die Rangzahlen der  $x$ -Verteilung gleich denen der  $y$ -Verteilung, betragen diese Differenzen 0. In diesem Falle ist der Rangkorrelationskoeffizient  $+1$ .

Sind die Rangzahlen so verteilt, dass die eine die Umkehrung der anderen darstellt, ergibt sich

<sup>26</sup>Vgl. Clauß, G., und Ebner, H.: Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Volk und Wissen — Volkseigener Verlag, Berlin 1967, S. 103

als Rangkorrelationskoeffizient  $-1$ . In beiden Fällen besteht ein enger Zusammenhang zwischen den Verteilungen.

Das Vorzeichen drückt die Richtung des Zusammenhangs, der Abhängigkeit aus. Wenn der Rangkorrelationskoeffizient gleich Null oder nahe Null ist, dann sind die Verteilungen unabhängig voneinander, bzw. es besteht nur ein sehr loser Zusammenhang.

Besonders in der Soziologie, Psychologie, Pädagogik und Ökonomie gibt es solche Rangfolgen, für die die Rangkorrelationskoeffizienten bestimmt werden müssen. Der Rangkorrelationskoeffizient gestattet im allgemeinen nur eine grobe Aussage.

Mit solchen Resultaten ist vorsichtig zu verfahren. Sie sind im großen und ganzen nur in Verbindung mit zusätzlichen Angaben und einer ausführlichen Kommentierung nützlich. Prinzipiell gilt, dass weder Rang- noch Maßkorrelationskoeffizienten unkritisch berechnet und angewendet werden dürfen.

Beispiele für die Berechnung des Rangkorrelationskoeffizienten sind u.a.

- die Prüfung der Korrelation zwischen der Vorliebe Jugendlicher für ausgewählte Hobbys und ihren Leistungen in bestimmten Fächern
  - die Feststellung der Übereinstimmung bzw. Abweichung bei Beurteilungen über Kollektivangehörige durch verschiedene Personen, z. B. bei der Beurteilung einer Gruppe von Lehrlingen oder einiger bestimmter Lehrlinge durch den Lehrausbilder und den FDJ-Sekretär
  - die Überprüfung der Wirksamkeit eines psychologischen Eignungstests in der Form, dass die Korrelation der sich auf der Grundlage des Eignungstests bei einer Gruppe von Personen ergebenden Rangfolge mit jener nach dem praktischen Einsatz tatsächlich erreichten Rangfolge hergestellt wird
  - die Kontrolle der Übereinstimmung beim Begutachten schriftlicher Schüler- oder Studentearbeiten durch zwei Lehrer
  - die Ermittlung des Zusammenhangs zwischen zwei verschiedenen Sportarten beim Training
  - die Prüfung des Zusammenhangs zwischen den Qualitätsstufen eines Erzeugnisses und seiner Haltbarkeit
  - der Vergleich von Betrieben einer VVB im Hinblick auf ihren Rang bezüglich ihres Mechanisierungsgrades der Arbeit und der Arbeitsproduktivität.
- Weitere Beispiele veranschaulicht Abbildung 23.

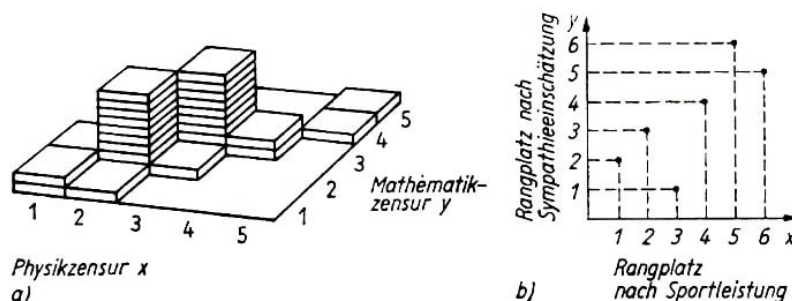


Abb. 23 Verteilung von Rangplätzen

Der linke Bildteil gibt einen Rangvergleich von Schülerleistungen in den Fächern Mathematik und Physik wieder, der rechte Bildteil zeigt die tatsächlich erreichte Leistung im Sportunterricht (ausgedrückt in einem Rangplatz des betreffenden Schülers) und die subjektive Bewertung des



Sportunterrichtes im Vergleich zu anderen Fächern durch diesen Schüler.

Wir wollen ein kleines Beispiel zur Rangkorrelation berechnen.

Auf einer Tieraussstellung sollen die 10 besten Tiere Preise bzw. Diplome erhalten. Es sind zwei Preisrichter vorhanden, die die Tiere in eine Rangfolge von 1...10 einstufen. Die Aufgabe besteht darin, zu prüfen, wie groß der Grad des Zusammenhangs zwischen diesen beiden Urteilen der Preisrichter ist.

Dazu wird der Rangkorrelationskoeffizient nach folgender Formel berechnet:

$$r^* = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (59)$$

Es bedeuten dabei

$r^*$  : Rangkorrelationskoeffizient

$x_i, y_i$ : Rangplatz, der durch den jeweiligen Preisrichter vergeben wurde

$d_i$ : Differenz zwischen den Rangplätzen ( $x_i - y_i$ )

$n$ : Anzahl der Rangplätze.

Beurteilte Tiere	Rangplatz durch Preisrichter 1	Rangplatz durch durch Preisrichter 2	Differenz zwischen den Rangplätzen	
	$x_i$	$y_i$	$d_i = x_i - y_i$	$d_i^2$
A	3	1	2	4
B	2	4	-2	4
C	6	2	4	16
D	7	10	-3	9
E	9	8	1	1
F	4	6	-2	4
G	8	9	-1	1
H	1	3	-2	4
I	5	5	0	0
J	10	7	3	9
$\Sigma$	55	55	0	52

Tabelle 14 Verteilung von Rangplätzen

Die Verteilung der Rangplätze 1...10 sowie die tabellarischen Vorarbeiten für die Berechnung von  $r^*$  sind aus Tabelle 14 zu ersehen. Nach (59) ergibt sich

$$r^* = 1 - \frac{6 \cdot 52}{10(100 - 1)} = 1 - \frac{52}{165} = +0,68$$

Bei einem Rangkorrelationskoeffizienten von +0,68 kann eingeschätzt werden, dass die Urteile der beiden Preisrichter noch relativ gut übereinstimmen. Allerdings weichen die für das Tier C zugeteilten Rangplätze stark voneinander ab.

Bei der Berechnung des Koeffizienten sollte beachtet werden, dass bei  $n < 5$  dieser Koeffizient kaum aussagefähig ist. Wird  $n = 30$  überschritten, so ist es infolge der sich vermindernden Übersicht schwierig, die Rangplätze ordnungsgemäß auf die Objekte zu verteilen. Allerdings kann man sich dann als Beurteilender dadurch helfen, dass z.B. bei der Vergabe der Plätze 1...50 erst einmal die einzelnen Objekte jeweils einer von fünf Gruppen zugeordnet werden und danach innerhalb dieser Gruppen eine Rangfolge festgelegt wird.

## 5.6 Korrelation qualitativer Merkmale

In allen jenen Fällen, in denen lediglich Merkmale aus qualitativ unterschiedlichen Gegenstandsklassen vorliegen, wird der Zusammenhang durch einen Kontingenzkoeffizienten oder einen Assoziativkoeffizienten ermittelt.

Das trifft besonders auf Aufgaben zu, die bei psychologischen, pädagogischen und soziologischen Forschungen zu lösen sind.

Der Assoziativkoeffizient  $Q$  wird für die Ermittlung des Zusammenhangs zwischen Zufallsvariablen mit alternativen Merkmalen – die in einer Vierfeldertabelle zusammengestellt sind – verwendet. Er soll hier nicht weiter betrachtet werden.

Der Berechnung des Kontingenzkoeffizienten  $C$  liegen qualitative Merkmale zugrunde, die in einer  $m \cdot n$ -Tabelle zusammengefasst werden. Mit diesem Kontingenzkoeffizienten wollen wir uns jetzt noch kurz beschäftigen.

Der Kontingenzkoeffizient  $C$  drückt als Maßzahl den Zusammenhang zwischen zwei oder mehr nominalen Verteilungen aus. Er kann alle Werte zwischen 0 und 1 annehmen und demzufolge auch nur die Stärke des Zusammenhangs zwischen den untersuchten Verteilungen und nicht wie der schon erwähnte Maß- und auch der Rangkorrelationskoeffizient mit Hilfe des Vorzeichens die Richtung der Abhängigkeit zum Ausdruck bringen.

Sind die Verteilungen unabhängig voneinander, hat der Kontingenzkoeffizient den Wert 0 und, wenn sie völlig voneinander abhängig sind, den Wert 1. Die Zahl der Felder wirkt sich bei seiner Berechnung aus. Der maximal erreichbare Wert des Koeffizienten  $C$  ist um so niedriger, je kleiner die Zahl der Tabellenfelder ist.

Nach einer Filmvorführung wurden z. B. 500 Besucher nach verschiedenen Gesichtspunkten befragt. Auf der Grundlage der allgemeinen Aussage, ob der Film den Besuchern gefallen oder nicht gefallen hat, sollte u. a. geprüft werden, ob zwischen diesen Urteilen und dem Alter der Teilnehmer an der Filmvorführung ein Zusammenhang bestand. Der berechnete Kontingenzkoeffizient lag mit  $C = 0,875$  nahe 1. Demzufolge konnte daraus geschlossen werden, dass zwischen der Beurteilung des Films und dem Alter der Besucher ein starker Zusammenhang bestand.

Wir wollen hier noch auf ein anderes Beispiel eingehen. Von Clauß und Ebner<sup>27</sup> wurden 446 jugendliche Lehrlinge gefragt

- ob ihr Berufswunsch in Erfüllung ging oder nicht ( $X$ ) und
- ob sie mit ihrem gegenwärtigen Lehrberuf zufrieden sind oder nicht ( $Y$ ).

Die Angaben hierzu sind aus der Tabelle 15 zu ersehen. Von diesen Daten ausgehend, wurde geprüft, ob ein Zusammenhang besteht zwischen der Erfüllung (bzw. Nichterfüllung) des Berufswunsches und der Zufriedenheit (bzw. Unzufriedenheit) mit dem Beruf.

Dazu haben die genannten Autoren Koeffizienten ermittelt, und zwar

- den Koeffizienten  $\Phi$ , der als Vierfelderkoeffizient berechnet wird, wenn die Merkmalsklassen echte Alternativen darstellen
- und den Kontingenzkoeffizienten  $C$ , der im allgemeinen nur verwendet wird, wenn jede Variable mehr als zwei Klassen besitzt.

---

<sup>27</sup>Vgl. ebenda, S. 264ff.

		Berufswunsch $X$		
		unerfüllt	erfüllt	
Einstellung	zufrieden	217	114	331
zum Lehrberuf $Y$	zufrieden	101	14	115
		318	128	446

Tabelle 15 Berufswunscherfüllung und Zufriedenheit mit dem Lehrberuf bei 446 jugendlichen Lehrlingen

Der Vergleich der Ergebnisse zeigte, dass die Koeffizienten  $\Phi$  und  $C$  der Aufgabenstellung und der Ausgangssituation angemessene Informationen brachten, und zwar  $\Phi \approx 0,23$  und  $C = 0,21$ . Diese Resultate sagen aus, dass Berufswunsch und Zufriedenheit mit dem Lehrberuf - bezogen auf diesen konkreten Fall - nur schwach miteinander in Verbindung stehen.

Allerdings sei an diesem Beispiel nochmals darauf verwiesen, dass sich bei derartigen soziologischen oder pädagogischen Untersuchungen erst durch den Vergleich mehrerer Kennzahlen — die sich im Hinblick auf definierbare Eigenschaften unterscheiden - gesicherte und aufschlussreiche Aussagen gewinnen lassen.

In der Statistik werden darüber hinaus noch weitere Kontingenzkoeffizienten verwendet, z. B.

- der tetrachorische Koeffizient  $r_{tet}$ , der berechnet werden kann, wenn die Variablenklassen normalverteilten Gesamtheiten entstammen

- die biserialen Korrelationskoeffizienten, die verwendet werden, wenn eine Veränderliche zweifach und die andere mehrfach gestaffelt ist.

Jeder dieser Kontingenzkoeffizienten liefert jedoch nur sinnvolle Aussagen zur Beschreibung von Zusammenhängen zwischen qualitativen Variablen, wenn spezielle Bedingungen erfüllt worden sind und die Aussagefähigkeit dieser Koeffizienten entsprechend beachtet wird.

Einige weitere Fälle, bei denen derartige Koeffizienten berechnet werden können, sind Untersuchungen zur Ermittlung der Stärke des Zusammenhangs zwischen

- den Leistungen eines Kollektivmitglieds und der Häufigkeit der Anerkennung dieses Mitarbeiters innerhalb des Kollektivs
- dem Geschlecht von Schülern und ihrer Zugehörigkeit zu einer naturwissenschaftlichen oder technischen Arbeitsgemeinschaft
- der Einstellung männlicher und weiblicher Jugendlicher zur Lyrik
- der Zufriedenheit mit der Familiensituation und der Zufriedenheit mit dem eigenen Charakter der befragten Jugendlichen
- dem Lebensalter von Versuchspersonen und dem Zeitaufwand zur Lösung einer genormten Aufgabe
- der Vorbildung von Prüflingen (z. B. unterschieden nach 10jähriger und 12jähriger Schulbildung) und der Anzahl der von jedem Prüfling in einer Prüfung richtig gelösten Aufgaben
- dem Interesse von Schülern für ein bestimmtes Schulfach und die erreichte Punktzahl bei einer Vergleichsarbeit in diesem Fach usw.

## 6 Was sind theoretische Verteilungen?

Im Kapitel 4 hatten wir uns anhand praktischer Beispiele mit den wesentlichen Eigenschaften und den Maßzahlen der empirischen Verteilungen befasst und dabei festgestellt, dass empirische Verteilungen immer im Ergebnis statistischer Erhebungen entstehen. Die Beobachtungswerte werden also durch praktische Untersuchungen gewonnen.

So waren es beispielsweise Messergebnisse, die über bestimmte Zusammenhänge und charakteristische Merkmale Aufschluss gaben.

Die Statistik beschränkt sich jedoch nicht auf das Erfassen von Einzelwerten, deren Zusammenfassung in Gruppen bzw. Klassen, die tabellarische und grafische Darstellung sowie das Berechnen statistischer Maßzahlen, wie Mittelwert und Streuung. Oft sind für statistische Analysen noch tiefergehende Betrachtungen notwendig. Das wurde bereits im Kapitel 5 deutlich.

Weitere Aspekte werden in den folgenden Kapiteln aufgezeigt. So ist z. B. jede Stichprobe bezüglich ihrer Aussage für die Grundgesamtheit mit einer gewissen Unsicherheit behaftet.

In der Praxis soll aber durch Stichproben eine den tatsächlichen Gegebenheiten entsprechende Einschätzung getroffen werden können, woraus sich die Notwendigkeit ergibt, diese Unsicherheit quantitativ zu erfassen, was wiederum die Kenntnis der in diesem Zusammenhang gültigen statistischen Gesetzmäßigkeiten voraussetzt. Dabei spielen theoretische Verteilungen eine sehr wesentliche Rolle. Ihren Inhalt und ihre praktische Bedeutung wollen wir in diesem Kapitel darlegen.

Die theoretischen Verteilungen beruhen im Unterschied zu den empirischen Verteilungen, bei denen durch statistische Erhebungen Beobachtungswerte gewonnen werden, auf den Gesetzen der Wahrscheinlichkeitsrechnung. Sie sind Wahrscheinlichkeitsverteilungen und geben die Aufteilung der Wahrscheinlichkeitssumme 1 (oder 100%) auf einzelne Merkmalsvariationen entsprechend bestimmten Ausgangsbedingungen in einer mathematischen Funktion - der Verteilungsfunktion - an.

Bevor wir aber auf prinzipielle Probleme der theoretischen Verteilungen eingehen, sei wegen seiner Wichtigkeit für die folgenden Untersuchungen noch einmal kurz auf den Begriff "Zufallsvariable" zurückgekommen, den wir bereits im Kapitel 2 kennenlernten.

Wir hatten dort u. a. gesagt, dass wir unter einer Zufallsvariablen  $X$  eine veränderliche Größe verstehen, die Werte in bestimmten vorgegebenen Intervallen der Zahlengeraden mit bestimmten Wahrscheinlichkeiten annimmt. Eine Zufallsvariable kann durch eine Verteilungsfunktion eindeutig beschrieben werden.

Wir erinnern uns, dass wir als eine diskrete Zufallsvariable eine Variable bezeichnen, die nur abzählbar viele Werte annehmen kann (Augenzahl beim Wurf mit einem Würfel, die Anzahl der fehlerhaften Stücke in einem Warenposten, die Anzahl der Kunden an einem Schalter, die Geburt eines Jungen oder eines Mädchens).

Eine stetige Zufallsvariable kann dagegen unendlich viele Werte - z. B. alle Werte in einem bestimmten Intervall - annehmen. Man spricht in diesem Zusammenhang auch von überabzählbar vielen Werten. Als Beispiele für stetige zufällige Variablen seien der Ernteertrag pro Hektar, die Abweichungen vom Sollmaß bei hergestellten Werkstücken, das Körpergewicht von Schülern oder die Geschwindigkeit von Fahrzeugen aufgeführt.

Die Zufallsvariable kann also nicht nur diskrete Werte, sondern jeden beliebigen Zwischenwert annehmen.

Entsprechend diesen beiden Gruppen von Zufallsvariablen existieren auch zwei große Gruppen

von theoretischen Verteilungen:

- die diskreten Verteilungen und
- die stetigen Verteilungen.

Theoretische Verteilungen, die in der praktischen Arbeit, insbesondere auch in den Naturwissenschaften und in der Technik, eine große Bedeutung haben, sind:

Diskrete Verteilungen

- die Binomialverteilung
- die Poisson-Verteilung
- die hypergeometrische Verteilung

Stetige Verteilungen

- die Normalverteilung
- die negative Exponentialverteilung.

Hinzu kommt noch die Gruppe der sogenannten Prüfverteilungen, die für das im Kapitel 8 erläuterte Prüfen von Hypothesen bedeutsam sind. Prüfverteilungen sind z. B. die  $t$ -Verteilung und die  $\chi^2$ -Verteilung, auf denen die dort erläuterten Prüfverfahren  $t$ -Test und  $\chi^2$ -Test beruhen.

Wir wollen im folgenden drei theoretische Verteilungen, und zwar die Binomialverteilung, die Poisson-Verteilung und die Normalverteilung, etwas näher vorstellen und an ihnen grundsätzliche Probleme der Arbeit mit theoretischen Verteilungen aufzeigen.

Da die theoretischen Verteilungen auf den Gesetzen der Wahrscheinlichkeitsrechnung beruhen, wollen wir uns aber zunächst mit einigen grundsätzlichen Fragen der Wahrscheinlichkeitsrechnung, insbesondere mit der Definition des Begriffes Wahrscheinlichkeit, befassen, ohne jedoch dieses vielseitige Teilgebiet der Mathematik erschöpfend behandeln zu können.

## 6.1 Einige Aspekte der Wahrscheinlichkeitsrechnung

Ein kurzer historischer Abriss über die Entwicklung der Wahrscheinlichkeitsrechnung wurde schon im Kapitel 1 gegeben. Wir wollen uns jetzt mit einigen inhaltlichen Fragen beschäftigen. Verdeutlichen wir uns den Inhalt der Wahrscheinlichkeitsrechnung an einem einfachen Beispiel.

Beim Wurf mit einem Würfel können die Augenzahlen 1, 2, 3, 4, 5, 6 auftreten. Bei jedem Wurf wird dabei eine dieser Zahlen eintreffen, wobei jedoch nicht vorausgesagt werden kann, welche es sein wird. Das Eintreffen der Zahlen ist dem Zufall unterworfen.

Deshalb sprechen wir hier von einer zufälligen Erscheinung und bezeichnen das Eintreffen der Zahl  $i$  ( $i$  ist in unserem Falle 1, 2, 3, 4, 5 oder 6) als ein zufälliges Ereignis bzw. Zufallsereignis.

Allgemein kann ein zufälliges Ereignis bei einem Versuch eintreten (man spricht in diesem Zusammenhang auch von einem zufälligen Versuch), der bestimmten Bedingungen genügt; es muss aber nicht eintreten. Es lässt sich also nicht voraussagen, ob das zufällige Ereignis eintreten wird oder nicht.<sup>28</sup>

Wir hatten einen solchen Sachverhalt bei der Definition der Zufallsvariablen im Kapitel 2 bereits gestreift und erinnern uns auch daran, dass Ereignisse die Elemente einer Menge sind.

Die Wahrscheinlichkeitsrechnung befasst sich mit den Gesetzmäßigkeiten derartiger zufälliger

---

<sup>28</sup>Vgl. Storm, R.: Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle. VEB Fachbuchverlag, Leipzig 1972, S. 19

Ereignisse. Die Gesetzmäßigkeiten lassen sich aber erst bei einer großen Anzahl unabhängig voneinander durchgeführter Wiederholungen eines entsprechenden Versuches erkennen. Deshalb muss sich der betreffende Versuch (zumindest theoretisch) unter Einhaltung der ihn kennzeichnenden äußeren Bedingungen beliebig oft wiederholen lassen. Die im Hinblick auf solche Gesetzmäßigkeiten zu untersuchenden Erscheinungen müssen also Massenerscheinungen sein.

In unserem Beispiel ist die Zufälligkeit der Ereignisse gegeben, denn das Ergebnis eines Wurfes (z. B. das Würfeln einer 1) lässt sich nicht voraussagen. Außerdem sind die Voraussetzungen für eine Massenerscheinung erfüllt, da sich der Wurf eines Würfels beliebig oft wiederholen lässt.

An solchen einfachen Beispielen, denen man auf den ersten Blick keine große praktische Bedeutung beimessen mag, werden in der Literatur zumeist die grundsätzlichen Probleme der Wahrscheinlichkeitsrechnung behandelt, denn ihre Verwendung hat einige wesentliche Vorteile. Sie sind jedem Leser gleichermaßen geläufig, an ihnen lassen sich viele Aussagen der Wahrscheinlichkeitsrechnung anschaulich erklären, und sie dienen als Modelle für zahlreiche praktische Fälle. Es sollen noch einige Beispiele für zufällige Erscheinungen im Sinne der Wahrscheinlichkeitsrechnung angegeben werden:

- das Erreichen von Zahl oder Wappen beim Werfen einer Münze
- die Anzahl der Kunden, die pro Stunde ein Warenhaus betreten
- die Anzahl der Ferkel je Wurf
- die Lebensdauer von Glühlampen
- die Wartezeit der Patienten in einem Wartezimmer
- das Körpergewicht Neugeborener usw.

Aus den bisherigen Kapiteln ist uns schon bekannt, dass hier Zufallsvariablen vorliegen, und zwar bei den ersten drei Beispielen diskrete und bei den übrigen stetige Zufallsvariablen.

Unter einem Versuch ist dabei z. B. das Werfen einer Münze oder das Ermitteln der Anzahl der Ferkel eines Wurfes oder das Bestimmen der Lebensdauer einer Glühlampe zu verstehen. Sein Ergebnis ist ein Zufallsergebnis.

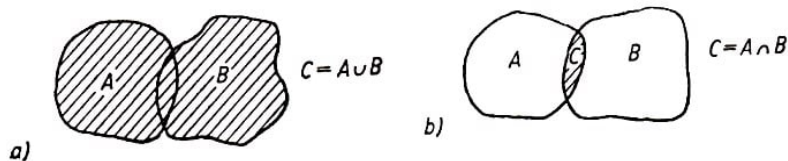
Es kann nicht Aufgabe dieses Buches sein, alle Beziehungen zwischen zufälligen Ereignissen darzulegen. In kurzer Form sei aber zumindest auf die Summe und auf das Produkt von zufälligen Ereignissen eingegangen.<sup>29</sup>

Prinzipiell gelten für zufällige Ereignisse die gleichen Beziehungen wie für Mengen. Deshalb kann der Leser, der Kenntnisse auf dem Gebiet der Mengenlehre besitzt, leicht entsprechende Analogien herstellen.

Das Ereignis

$$C = A \cup B \quad (60)$$

(lies: " $C = A$  vereinigt mit  $B$ " oder " $C =$  Summe von  $A$  und  $B$ ") ist die (logische) Summe der Ereignisse  $A$  und  $B$ . Es tritt ein, wenn wenigstens eines der beiden Ereignisse (das Ereignis  $A$  oder das Ereignis  $B$ ) eingetreten ist. Bei Mengen heißt  $A \cup B$  die Vereinigung von  $A$  und  $B$ ; die Vereinigungsmenge zweier Mengen ist die Menge aller Elemente, die mindestens einer dieser Mengen angehören (vgl. Abb. 24a).



<sup>29</sup>Vgl. ebenda, S. 20

Abb. 24 Vereinigung und Durchschnitt zweier Mengen  
a) Vereinigung von  $A$  und  $B$ ; b) Durchschnitt von  $A$  und  $B$

Ist beim Würfeln mit einem Würfel das Ereignis  $A$  das Werfen einer 1, wofür die abkürzende Schreibweise  $A = \{1\}$  gebraucht wird, und das Ereignis  $B$  das Werfen einer geraden Zahl, also  $B = \{2, 4, 6\}$ , dann tritt das Ereignis  $C = A \cup B$  ein, wenn entweder 1, 2, 4 oder 6 geworfen wird.

Das Ereignis

$$C = A \cap B \quad (61)$$

(lies: " $C$  = Produkt von  $A$  und  $B$ " oder " $C$  =  $A$  geschnitten mit  $B$ ") ist das (logische) Produkt der Ereignisse  $A$  und  $B$ . Es tritt dann ein, wenn sowohl das Ereignis  $A$  als auch das Ereignis  $B$  eingetreten ist. Bei Mengen heißt  $A \cap B$  der Durchschnitt von  $A$  und  $B$ ; die Durchschnittsmenge zweier Mengen ist die Menge aller Elemente, die zu beiden Mengen gehören (vgl. Abb. 24b).

Ist beim Wurf mit einem Würfel das Ereignis  $A = \{1, 2, 3\}$  und das Ereignis  $B = \{1, 3, 5\}$ , dann tritt das Ereignis  $C = A \cap B$  ein, wenn eine 1 oder eine 3 geworfen wird.

Für die weiteren Betrachtungen sind vor allem noch folgende Begriffe von Bedeutung:

- Sicheres Ereignis ( $E$ )

Ein Ereignis heißt sicheres Ereignis, wenn es bei jedem Versuch eintritt. So ergibt die Summe der sechs möglichen Ereignisse beim Würfeln ein sicheres Ereignis  $E$ , denn bei jedem Wurf tritt eines der sechs Ereignisse, eine 1, 2, 3, 4, 5 oder 6, mit Sicherheit ein. Damit ist in diesem Fall

$$A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 \cup A_6 = E$$

- Unmögliches Ereignis ( $\emptyset$ )

Unter einem unmöglichen Ereignis versteht man ein Ereignis, das bei keinem Versuch eintreten kann. Bei einem Wurf mit einem Würfel z. B. ist das Erreichen einer 7 ein unmögliches Ereignis. Man schreibt dafür  $\{7\} = \emptyset$

- Komplementäres Ereignis  $\bar{A}$  (lies:  $A$  quer)

Das Ereignis  $\bar{A}$  heißt das zum Ereignis  $A$  komplementäre Ereignis. Es tritt ein, wenn Ereignis  $A$  nicht eintritt. Ist beim Wurf mit einem Würfel das Werfen einer geraden Zahl das Ereignis  $A$ , dann ist das hierzu komplementäre Ereignis  $\bar{A}$  das Werfen einer ungeraden Zahl. Man schreibt:  $A = \{2, 4, 6\}$ ,  $\bar{A} = \{1, 3, 5\}$ .

Es gilt immer  $A \cup \bar{A} = E$  und  $A \cap \bar{A} = \emptyset$ .

- Unvereinbare bzw. disjunkte Ereignisse

Ereignisse heißen unvereinbar bzw. disjunkt, wenn ihr gleichzeitiges Eintreten unmöglich ist. Schließen sich zwei Ereignisse,  $A$  und  $B$ , gegenseitig aus, sind sie also disjunkt, so wird formuliert:  $A \cap B = \emptyset$ .

Setzt man als Ereignis  $A$  das Werfen einer 1, also  $A = \{1\}$ , und als Ereignis  $B$  das Werfen einer 6, also  $B = \{6\}$ , dann sind  $A$  und  $B$  disjunkte Ereignisse, d.h.  $A \cap B = \emptyset$ .

- Ereignisfeld

Die Menge aller möglichen Zufallsereignisse, die bei einem Versuch auftreten können, bezeichnet man als Ereignisfeld  $\mathfrak{A}$ , d.h., jedem Versuch kann ein Ereignisfeld zugeordnet werden, dessen Elemente die Zufallsereignisse sind.

Beim Versuch "Wurf mit einem Würfel" gehören u. a. folgende Ereignisse zum Ereignisfeld:  
 $A_i$ : die Augenzahl  $i$  würfeln ( $i = 1, 2, 3, 4, 5, 6$ )

$B$ : eine gerade Augenzahl würfeln  
 $C$ : eine ungerade Augenzahl würfeln  
 $D$ : mindestens eine 3 würfeln  
 $E$ : höchstens eine 4 würfeln  
 $F$ : eine Augenzahl würfeln, die eine Primzahl ist, usw.

Damit ist nur eine Auswahl von Ereignissen aufgeführt, die zum Ereignisfeld des Versuches "Wurf mit einem Würfel" gehören. Wir erkennen aber, was unter dem Ereignisfeld zu verstehen ist, eben die Menge (oder die Gesamtheit) aller möglichen Ereignisse, die bei einem Versuch auftreten können.

Zum Ereignisfeld gehören auch das sichere Ereignis  $E$  und das unmögliche Ereignis  $\emptyset$ . In der Regel ist also ein Ereignisfeld umfangreicher als die Menge der im Zusammenhang mit einem Versuch unmittelbar interessierenden Ereignisse.

- Elementarereignis

Ein Ereignis, das nicht zerlegbar ist, heißt Elementarereignis. Zum Beispiel sind die Ereignisse  $A_i$  ( $i = 1, 2, \dots, 6$ ) beim Wurf mit einem Würfel Elementarereignisse. Dagegen sind die bei der Erläuterung des Ereignisfeldes angegebenen Ereignisse  $B$  bis  $F$  zerlegbare Ereignisse (z. B. lässt sich das Werfen einer geraden Augenzahl  $B = \{2, 4, 6\}$  als die Summe  $2 \cup 4 \cup 6$  darstellen).

Nach diesen Erläuterungen zu grundlegenden Begriffen und Regeln der Wahrscheinlichkeitsrechnung wollen wir uns nun dem Begriff Wahrscheinlichkeit selbst zuwenden. Für den Wahrscheinlichkeitsbegriff existieren mehrere Definitionen, die von unterschiedlichen Betrachtungsweisen ausgehen. Drei dieser Definitionen sollen hier angegeben werden.

Der älteste Wahrscheinlichkeitsbegriff ist der sogenannte klassische Wahrscheinlichkeitsbegriff. Er wurde von P. Laplace bei Untersuchungen über Glücksspiele entwickelt und geht von der Gleichmäßigkeit zufälliger Ereignisse aus.

Beim Wurf mit einem Würfel repräsentieren z. B. die Elementarereignisse  $A_1, A_2, \dots, A_6$  (das Werfen einer 1, 2, 3, 4, 5 oder 6) die Elementarereignisse des betreffenden Ereignisfeldes.

Für das Eintreffen der einzelnen Elementarereignisse besteht die gleiche Wahrscheinlichkeit, nämlich  $\frac{1}{6}$ .

Wird dieses Beispiel verallgemeinert, kommt man zu folgender Beziehung für die Wahrscheinlichkeit  $P$  eines Ereignisses  $A$ :

$$P(A) = \frac{g}{k} \quad (62)$$

bzw.

$$P(A) = \frac{\text{für } A \text{ günstige Elementarereignisse}}{\text{mögliche Elementarereignisse}}$$

Die klassische Definition des Wahrscheinlichkeitsbegriffes besagt:

Die Wahrscheinlichkeit  $P$  für das Eintreffen des Ereignisses  $A$  berechnen wir aus dem Quotienten der Anzahl der für das Eintreten von  $A$  günstigen Elementarereignisse ( $g$ ) und der Anzahl aller überhaupt möglichen Elementarereignisse des betreffenden Ereignisfeldes  $\mathfrak{A}$ , die wir durch das Symbol  $k$  ausdrücken.

Weil immer die Beziehung  $0 \leq g \leq k$  gilt, ergibt  $P(A)$  immer eine Zahl zwischen 0 und 1. Für  $g = k$  (d.h.  $A = E$ ) ist  $P(A) = 1$ , für  $g = 0$  (d.h.  $A = \emptyset$ ) ist  $P(A) = 0$ .

Der klassische Wahrscheinlichkeitsbegriff hat allerdings den Nachteil, dass er nur auf Ereignisfelder anwendbar ist, die eine endliche Zahl gleichmöglicher Elementarereignisse enthalten.

In zahlreichen praktischen Fällen bereitet es jedoch auch bei einer endlichen Zahl von Elementarereignissen große Schwierigkeiten, alle gleichmöglichen Elementarereignisse herauszufinden.



Das Ermitteln der Größen  $g$  und  $k$  erfolgt dabei in der Regel nach den Gesetzen der Kombinatorik. Die Kombinatorik ist ein Teilgebiet der Mathematik, das die verschiedenen Möglichkeiten der Anordnung und Zusammenstellung von Elementen einer Menge untersucht.

Von anderen Überlegungen geht die statistische Definition der Wahrscheinlichkeit aus. Wir haben bereits ausgeführt, dass nicht vorausgesagt werden kann, mit einem bestimmten Wurf beispielsweise eine 1 zu werfen. Wenn wir den Würfel aber sehr oft werfen und das jeweilige Auftreten einer 1 vermerken, werden wir bald feststellen, dass die relative Häufigkeit für das Auftreten einer 1 ( $f_1$ ) um den Wert  $\frac{1}{6} = 0,1666$  schwankt. Sie nähert sich dem Wert - um so mehr, je mehr Würfe wir durchführen.

Die relative Häufigkeit (vgl. hierzu auch Abschnitt 4.1.) für das Würfeln einer 1 berechnen wir dabei nach Formel

$$f_1 = \frac{h_1}{n} \quad (63)$$

Hierin ist  $f_1$  die relative Häufigkeit für das Eintreffen einer 1,  $h_1$  die Zahl der erreichten Einsen (die absolute Häufigkeit) bei einer bestimmten Anzahl von Würfeln und  $n$  die Zahl der ausgeführten Würfe.

Die Tabelle 16 zeigt einige Werte eines derartigen Versuches für das Erreichen der Zahl 1.

Anzahl der Würfe $n$	Anzahl der vorkommenden Einsen $h_1$	relative Häufigkeit $f_1 = \frac{h_1}{n}$
10	3	0,3000
100	18	0,1800
1000	161	0,1610
10000	1656	0,1656

Tabelle 16 Erreichte Augenzahl beim Würfeln bei unterschiedlicher Anzahl von Versuchen

Die aufgezeigte Eigenschaft wird als Stabilität der relativen Häufigkeit bezeichnet und drückt aus, dass bei einer genügend großen Zahl von Versuchen die relative Häufigkeit eine gewisse Stabilität zeigt. Die relative Häufigkeit eines zufälligen Ereignisses schwankt dabei um einen festen Wert. Es ist mathematisch bewiesen, dass sie sich diesem Wert um so mehr nähert, je größer die Anzahl der Versuche ist.

Diese Konstante (sie liegt immer zwischen 0 und 1) wird als die statistische Wahrscheinlichkeit für das Auftreten eines zufälligen Ereignisses bezeichnet und ebenfalls durch den Buchstaben  $P$  ausgedrückt, der allgemein den Begriff Wahrscheinlichkeit symbolisiert. Der statistische Wahrscheinlichkeitsbegriff ist bei praktischen Anwendungen ausreichend.

Aus mathematischer Sicht können jedoch sowohl der klassische als auch der statistische Wahrscheinlichkeitsbegriff nicht voll befriedigen. Eine mathematisch exakte Definition gibt der axiomatische Wahrscheinlichkeitsbegriff, der abschließend erläutert werden soll. Die vom sowjetischen Mathematiker A. N. Kolmogorow 1933 veröffentlichte Axiomatik hat folgenden Inhalt:

#### Axiom 1

Jedem zufälligen Ereignis  $A$  aus dem Ereignisfeld  $\mathfrak{A}$  wird eine nichtnegative Zahl  $P(A)$  zugeordnet. Diese Zahl heißt die Wahrscheinlichkeit des Ereignisses  $A$ . Für sie gilt

$$0 \leq P(A) \leq 1 \quad (64)$$

#### Axiom 2

Die Wahrscheinlichkeit des sicheren Ereignisses  $E$  ist gleich 1, also

$$P(E) = 1 \quad (65)$$

Axiom 3

Die Wahrscheinlichkeit der Summe zweier disjunkter Ereignisse  $A$  und  $B$  (also  $A \cap B = \emptyset$ ) ist die Summe der Wahrscheinlichkeiten  $P(A)$  und  $P(B)$  dieser Ereignisse. Es gilt somit:

$$P(A \cup B) = P(A) + P(B) \quad (66)$$

Das bedeutet, die Wahrscheinlichkeit für das Eintreten entweder des Ereignisses  $A$  oder des Ereignisses  $B$  ist gleich der Summe der Wahrscheinlichkeiten der einzelnen Ereignisse, wenn sich diese gegenseitig ausschließen.

Unter einem Axiom ist eine bestimmte Aussage oder Grundeigenschaft zu verstehen, die als erwiesen gilt und aus der dann weitere Eigenschaften als Sätze abgeleitet werden. Die axiomatische Definition erklärt die Wahrscheinlichkeit durch bestimmte Eigenschaften der Wahrscheinlichkeit. Die Axiome bilden das Minimum von Forderungen an den Wahrscheinlichkeitsbegriff. Wird eines der Axiome weggelassen, ist der Wahrscheinlichkeitsbegriff als objektive Bewertungsgröße eines Ereignisses nur noch unvollkommen definiert. Alle weiteren Eigenschaften der Wahrscheinlichkeit lassen sich aus den drei Axiomen ableiten.

## 6.2 Die Binomialverteilung - eine Verteilung für alternative Merkmalsvariationen

Am Beispiel der Binomialverteilung sollen außer dem Inhalt dieser Verteilung auch einige grundsätzliche Probleme theoretischer Verteilungen aufgezeigt werden.

Zur Ableitung der Binomialverteilung verwendet man in der Literatur wegen seiner Anschaulichkeit oft ein Urnenmodell. In einer Urne sind schwarze und weiße Kugeln enthalten.

Die Wahrscheinlichkeit für das Ziehen einer schwarzen Kugel (Ereignis  $A$ ) ist  $p$ , die für das Ziehen einer weißen Kugel (Ereignis  $\bar{A}$ ) ist  $q$ . Beide Ereignisse sind die einzig möglichen und schließen somit einander aus, sind also Komplementärereignisse.

Da das Erhebungsmerkmal nur zwei qualitativ verschiedene Merkmalsvariationen aufweist, liegt ein alternatives Merkmal vor. Diesem Modell entsprechen zahlreiche praktische Aufgaben. Dabei gilt:

$$p + q = 1 \quad (67)$$

Der Urne entnimmt man nacheinander  $n$  Kugeln, wobei man sich für die Wahrscheinlichkeit interessiert, unter  $n$  Kugeln genau  $k$  schwarze Kugeln zu ziehen. Um eine unendliche Grundgesamtheit nachzubilden, wird jede Kugel nach der Ziehung wieder in die Urne zurückgelegt. Die Zufallsvariable  $X$  kann als die Anzahl der schwarzen Kugeln unter  $n$  gezogenen Kugeln definiert werden.

$X$  ist diskret und kann die ganzzahligen Werte  $k = 0, 1, \dots, n$  annehmen. Gesucht wird die Verteilung von  $X$ , d.h., die Wahrscheinlichkeiten  $P(X = k) = p_k$  ( $k = 0, 1, \dots, n$ ) sind zu bestimmen.

Ziehen wir zwei Kugeln, dann sind folgende Ereignisse möglich:

$$AA, A\bar{A}, \bar{A}A, \bar{A}\bar{A}$$

Beachten wir die Reihenfolge nicht, können wir auch schreiben

$$AA, 2A\bar{A}, \bar{A}\bar{A}$$

Das entspricht den Wahrscheinlichkeiten  $p^2, 2pq, q^2$ , so dass wir als Wahrscheinlichkeitsverteilung der Zufallsgröße  $X$  (Anzahl der schwarzen Kugeln unter  $n$  gezogenen Kugeln) für  $n = 2$  erhalten

$$P(X = 0) = q^2, \quad P(X = 1) = 2pq, \quad P(X = 2) = p^2$$

mit

$$\sum_{k=0}^2 P(X = k) = (p + q)^2 = 1$$

Für drei gezogene Kugeln erhalten wir bei ähnlichen Überlegungen die Wahrscheinlichkeitsverteilung

$$P(X = 0) = q^3, \quad P(X = 1) = 3pq^2, \quad P(X = 2) = 3p^2q, \quad P(X = 3) = p^3$$

mit

$$\sum_{k=0}^3 P(X = k) = (p + q)^3 = 1$$

Erweitern wir unsere Betrachtungen für  $n$  gezogene Kugeln, ergibt sich als Wahrscheinlichkeitsverteilung von  $X$ :<sup>30</sup>

$$\begin{aligned} P(X = 0) &= q^n \\ P(X = 1) &= nq^{n-1}p = \frac{n}{1}q^{n-1}p \\ P(X = 2) &= \frac{n(n-1)}{1 \cdot 2}q^{n-2}p^2 \\ &\dots \\ P(X = k) &= \frac{n(n-1)(n-2)\dots(n-k+1)}{1 \cdot 2 \cdot \dots \cdot k}q^{n-k}p^k \\ &\dots \\ P(X = n) &= p^n \end{aligned}$$

Wir führen jetzt die abkürzende Schreibweise

$$\frac{n}{1} = \binom{n}{1} \text{ (lies: } n \text{ über } 1), \quad \frac{n(n-1)}{1 \cdot 2} = \frac{n}{2}$$

und allgemein

$$\frac{n(n-1)(n-2)\dots(n-k+1)}{1 \cdot 2 \cdot 3 \cdot \dots \cdot k} = \binom{n}{k} \quad (68)$$

ein. Die gesuchte Wahrscheinlichkeit, unter  $n$  gezogenen Kugeln genau  $k$  schwarze zu finden, ist dann

$$p_k = P(X_k) = \binom{n}{k} p^k q^{n-k} \quad (k = 0, 1, \dots, n) \quad (69)$$

Das ist die Wahrscheinlichkeitsfunktion  $f(x)$  der Binomialverteilung.

Die Wahrscheinlichkeitsfunktion einer diskreten Zufallsvariablen  $X$  - für sie ist auch die Bezeichnung Häufigkeitsfunktion gebräuchlich - gibt die Wahrscheinlichkeiten  $p_k$  an, mit denen die diskrete Zufallsvariable ihre möglichen Werte  $k$  annimmt.

<sup>30</sup>Vgl. ebenda, S. 46f.

Ihr entspricht im übertragenen Sinn die Häufigkeitsverteilung empirischer Verteilungen bei diskreten Beobachtungswerten.

Außer der Wahrscheinlichkeitsfunktion  $f(x)$  können wir für eine Zufallsvariable  $X$  die Verteilungsfunktion  $F(x)$  angeben. Wenn  $X$  eine Zufallsvariable ist, dann existiert für jede reelle Zahl  $x$  eine Wahrscheinlichkeit dafür, dass der Wert, der von  $X$  angenommen wird, kleiner als  $x$  ist. Das wird wie folgt ausgedrückt:

$$P(X < x) = F(x) \quad (70)$$

Die Funktion  $F(x)$ , die über alle reellen Zahlen  $x$  definiert ist, bezeichnet man als die Verteilungsfunktion der Zufallsvariablen  $X$ . Die Verteilungsfunktion theoretischer Verteilungen entspricht im übertragenen Sinn der Summenverteilung bei empirischen Verteilungen.

Für die Binomialverteilung lautet die Verteilungsfunktion

$$F(x) = \sum_{k < x} p_k = P(X < x) = \begin{cases} 0 & \text{für } x \leq 0 \\ \sum_{k < x} \binom{n}{k} p^k q^{n-k} & \text{für } 0 < x \leq n \\ 1 & \text{für } x > n \end{cases} \quad (71)$$

Dabei gilt

$$\sum_{k=0}^n P(X = k) = \sum_{k=0}^n p_k = q^n + \binom{n}{1} q^{n-1} p + \dots + \binom{n}{n-1} q p^{n-1} + p^n = (p + q)^n = 1 \quad (72)$$

Wenn  $x \leq 0$ , d.h.  $k < 0$ , was praktisch nicht realisierbar ist, dann beträgt  $\sum_{k < x} p_k = 0$ . Für  $x > n$  ergibt  $\sum_{k < x} p_k = 1$ .

Für die anderen Werte von  $x$  sind die Wahrscheinlichkeiten nach der angegebenen Funktion zu berechnen. Der Ausdruck  $(p + q)^n$  wird in der Mathematik als Binom bezeichnet. Daraus leitet sich die Bezeichnung Binomialverteilung ab.

Die Symbole bedeuten in allgemeiner Darstellung

$f(x)$ : Wahrscheinlichkeitsfunktion

$F(x)$ : Verteilungsfunktion

$p_k = P(X = k)$ : Wahrscheinlichkeit dafür, dass die Zufallsgröße  $X$  den Wert  $k$  annimmt

$P(X < x)$ : Wahrscheinlichkeit dafür, dass die Zufallsgröße  $X$  einen Wert annimmt, der kleiner als der Wert  $x$  ist

$k$ : Wert der Zufallsgröße  $X$

$n$ : Stichprobenumfang

$p$ : Anteilswert mit der Eigenschaft  $A$

$q$ : Anteilswert mit der Eigenschaft  $\bar{A}$

$x$ : Schranke; reelle Zahl.

Die Größen  $p$  und  $n$  sind die Parameter der Binomialverteilung. Mittelwert und Varianz der Binomialverteilung werden nach den Formeln

$$\mu = np \quad \text{bzw.} \quad \sigma^2 = npq \quad (73,74)$$

berechnet. (Griechische Buchstaben werden zur Kennzeichnung statistischer Maßzahlen verwendet, wenn sich diese auf Grundgesamtheiten beziehen.)

Die Binomialverteilung gibt die Wahrscheinlichkeit dafür an, dass ein Merkmal, das in der Grundgesamtheit (z. B. in einem Warenposten) mit der Wahrscheinlichkeit  $p$  vorhanden ist, in

einer Stichprobe vom Umfang  $n$  gerade  $k$ -mal gefunden wird, während es  $(n - k)$ -mal nicht auftritt.

Eine Vorstellung von der Binomialverteilung kann durch das Ermitteln der Wahrscheinlichkeiten nach der Wahrscheinlichkeitsfunktion für verschiedene Werte von  $n$  und  $p$  gegeben werden. Die Tabellen 17 und 18 sowie die grafischen Darstellungen der darin enthaltenen Werte in den Abbildungen 25 und 26 zeigen hierfür zwei Beispiele.

$k$	$p_k$	$k$	$p_k$
0	0,5904	0	0,0078
4	0,3280	4	0,0547
2	0,0729	2	0,1640
3	0,0081	3	0,2734
4	0,0005	4	0,2734
5	0,0000	5	0,1640
$\Sigma$	0,9999	6	0,0547
		7	0,0078
		$\Sigma$	0,9998

Tabelle 17 (links) Wahrscheinlichkeiten  $p_k$  der Binomialverteilung für  $n = 5$ ,  $p = 0,1$

Tabelle 18 (rechts) Wahrscheinlichkeiten  $p_k$  der Binomialverteilung für  $n = 7$ ,  $p = 0,5$

Die Binomialverteilung ist im allgemeinen unsymmetrisch. Sie wird um so symmetrischer, je mehr sich  $p$  dem Wert 0,5 nähert und erreicht bei  $p = 0,5$  Symmetrie.

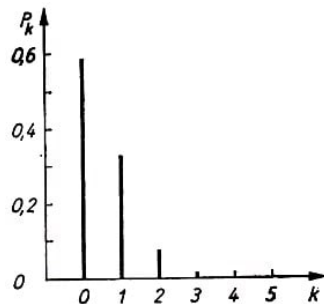


Abb. 25 Wahrscheinlichkeitsfunktion der Binomialverteilung für  $n = 5$  und  $p = 0,1$ ,

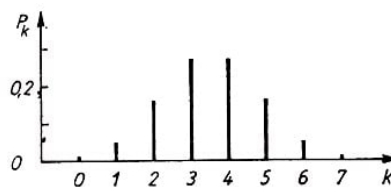


Abb. 26 Wahrscheinlichkeitsfunktion der Binomialverteilung für  $n = 7$  und  $p = 0,5$

Die Wahrscheinlichkeiten  $p_k$  für unterschiedliche  $p$  und  $n$  sind in zahlreichen statistischen Büchern in tabellierter Form angegeben.<sup>31</sup> Somit erfordert das Ermitteln dieser Werte keinen Rechenaufwand, sondern nur das Nachschlagen in einer Tabelle.

An einem Beispiel wollen wir nunmehr einen Einblick in die praktische Bedeutung der Binomialverteilung geben.

In einem Betrieb wird die Prüfung hergestellter Erzeugnisse nach dem Prinzip "gut - schlecht" durchgeführt. Der Anteil der Produkte, die Qualitätsmängel aufweisen, soll 10 % nicht überschreiten (Anteilswert  $p = 0,10$ ).

Es interessiert uns die Wahrscheinlichkeit dafür, dass sich unter 20 geprüften Erzeugnissen genau 2 mit Qualitätsmängeln befinden. Gegeben sind somit folgende Größen:

$$p = 0,10; \quad q = 0,90; \quad n = 20$$

<sup>31</sup>Vgl. z. B. Maibaum, G.: Wahrscheinlichkeitsrechnung. Volk und Wissen - Volkseigener Verlag, Berlin 1971, S. 214ff.

Gesucht wird die Wahrscheinlichkeit  $P(X = 2)$ . Wir berechnen nach der Wahrscheinlichkeitsfunktion (69)

$$P(X = 2) = \binom{20}{2} \cdot 0,10^2 \cdot 0,90^{18} = 0,2848 \hat{=} 28,48\%$$

Somit hat die Wahrscheinlichkeit, bei 20 geprüften Erzeugnissen genau 2 Erzeugnisse mit Qualitätsmängeln zu finden, den Wert 0,2848. Mit der Wahrscheinlichkeitsfunktion kann also die Wahrscheinlichkeit für das Eintreffen eines ganz bestimmten Ereignisses (in unserem Fall unter 20 Erzeugnissen 2 fehlerhafte zu finden) berechnet werden.

Wenn wir dagegen in unserem Beispiel nach der Wahrscheinlichkeit fragen, unter 20 Erzeugnissen höchstens 2 Erzeugnisse mit Qualitätsmängeln zu erhalten, müssen wir mit der Verteilungsfunktion arbeiten und rechnen

$$\begin{aligned} P(X \leq 2) &= \sum_{k=0}^2 P(X = k) = P(X = 0) + P(X = 1) + P(X = 2) \\ &= \binom{20}{0} \cdot 0,10^0 \cdot 0,90^{20} + \binom{20}{1} \cdot 0,10^1 \cdot 0,90^{19} \\ &\quad + \binom{20}{2} \cdot 0,10^2 \cdot 0,90^{18} = 0,1214 + 0,2698 + 0,2848 = 0,6740 \hat{=} 67,60\% \end{aligned}$$

Wir müssen in diesem Fall also die Summe aus den jeweiligen Einzelwahrscheinlichkeiten bilden. Die Größen  $\mu$  und  $\sigma^2$  betragen in diesem Beispiel

$$\begin{aligned} \mu &= np = 20 \cdot 0,1 = 2 \\ \sigma^2 &= npq = 20 \cdot 0,1 \cdot 0,9 = 1,8 \end{aligned}$$

Die Binomialverteilung ist in der Praxis u. a. für die statistische Qualitätskontrolle bei nicht-messender Prüfung (Prinzip gut-schlecht) von Bedeutung. Eine Stichprobe vom Umfang  $n$ , in der eine bestimmte Anzahl fehlerhafter Stücke gefunden wird, erlaubt z. B. Schlüsse, ob der gesamte Posten hergestellter Erzeugnisse den Qualitätsanforderungen genügt oder nicht.

Darauf wird im Kapitel 8 noch ausführlicher eingegangen. Die mathematische Grundlage hierfür gibt uns die Binomialverteilung. Streng genommen gilt die Binomialverteilung allerdings nur, wenn eine unendlich große Grundgesamtheit vorliegt. Das können wir in dem eben erläuterten Beispiel dadurch erreichen, dass wir dem Posten jeweils nur ein Erzeugnis entnehmen, es prüfen und nach erfolgter Prüfung dem Posten wieder zuführen, dann dem Posten das nächste Erzeugnis entnehmen, es prüfen und wieder dem Posten zuführen usw. Jedes Erzeugnis hat somit die Chance, mehrfach ausgewählt zu werden.

In der Regel verfährt man praktisch natürlich nicht so, sondern wird jeweils ein anderes Erzeugnis prüfen. Aber auch dann treffen bei genügend großem Umfang der Grundgesamtheit die Gesetzmäßigkeiten der Binomialverteilung näherungsweise zu, weil der Fehler, der durch das Nichtbeachten des Umfangs der Grundgesamtheit entsteht, in diesem Fall vernachlässigbar klein wird. Wir kommen darauf im Kapitel 7 zurück.

Liegen keine "großen Grundgesamtheiten" vor und wird mit Stichproben "ohne Zurücklegen" gearbeitet, dann haben die Gesetzmäßigkeiten der hypergeometrischen Verteilung Gültigkeit. Diese Verteilung beruht auf ähnlichen Prinzipien wie die Binomialverteilung, wobei aber der Umfang der Grundgesamtheit berücksichtigt wird.

### 6.3 Die Poisson-Verteilung und ihre praktische Anwendung

Die Poisson-Verteilung ist ebenso wie die Binomialverteilung eine diskrete Verteilung. Sie lässt sich aus der Binomialverteilung ableiten, wenn dort  $n$  gegen unendlich strebt und  $np = \lambda$  konstant bleibt.

Die Wahrscheinlichkeitsfunktion der Poisson-Verteilung ist durch die Beziehung

$$f(x) = p_k = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad (k = 0, 1, 2, \dots) \quad (75)$$

und ihre Verteilungsfunktion durch die Beziehung

$$F(x) = \sum_{k < x} p_k = P(X < x) = \begin{cases} 0 & \text{für } x \leq 0 \\ \sum_{k < x} \frac{\lambda^k}{k!} e^{-\lambda} & \text{für } x > 0 \\ 23 & \text{sonst} \end{cases} \quad (76)$$

gegeben. Es gilt

$$\sum_{k=0}^{\infty} p_k = 1 \quad (77)$$

Dabei bedeuten

$f(x)$ : Wahrscheinlichkeitsfunktion

$F(x)$ : Verteilungsfunktion

$p_k = P(X = k)$ : Wahrscheinlichkeit dafür, dass die Zufallsgröße  $X$  den Wert  $k$  annimmt

$P(X < x)$ : Wahrscheinlichkeit dafür, dass die Zufallsgröße  $X$  einen Wert annimmt, der kleiner als der Wert  $x$  ist

$k$ : Wert der Zufallsgröße  $X$

$\lambda$ : Konstante (Parameter der Poisson-Verteilung)

$e$ : Konstante (Wert: 2,718)

$x$ : Schranke; reelle Zahl.

Für den Mittelwert und die Varianz gelten folgende Beziehungen:

$$\mu = \lambda, \quad \sigma^2 = \lambda \quad (78)$$

Mittelwert und Varianz sind also gleich dem Parameter  $\lambda$ .

Wir ermitteln jetzt die Wahrscheinlichkeiten  $p_k$  nach der Wahrscheinlichkeitsfunktion für unterschiedliche  $\lambda$ . Dabei wählen wir als Beispiele  $\lambda = 0,5$  und  $\lambda = 2,0$ . Für  $\lambda = 0,5$  ergibt sich nach Formel (75) die Tabelle 19.

$k$	$p_k$	$k$	$p_k$
0	0,6065	0	0,1353
1	0,3033	1	0,2707
2	0,0758	2	0,2707
3	0,0126	3	0,1804
4	0,0016	4	0,0902
5	0,0002	5	0,0361
6	0,0000	6	0,0420
$\Sigma$	1,0000	7	0,0034
		8	0,0009
		9	0,0002
		10	0,0000
		$\Sigma$	0,9999

Tabelle 19 (links) Wahrscheinlichkeiten  $p_k$  der Poisson-Verteilung für  $\lambda = 0,5$

Tabelle 20 (links) Wahrscheinlichkeiten  $p_k$  der Poisson-Verteilung für  $\lambda = 2,0$

Die Werte der Tabelle 19 sagen u. a. aus, dass bei  $\lambda = 0,5$  die Wahrscheinlichkeit dafür, dass eine Zufallsgröße den Wert 0 annimmt, 0,6065 bzw. 60,65% beträgt; dass die Zufallsgröße den Wert 1 annimmt, 0,3033 bzw. 30,33% beträgt usw.

Die Wahrscheinlichkeiten sind in Abbildung 27 grafisch dargestellt.

Für  $\lambda = 2,0$  ergeben sich die in Tabelle 20 und Abbildung 28 enthaltenen Wahrscheinlichkeiten.

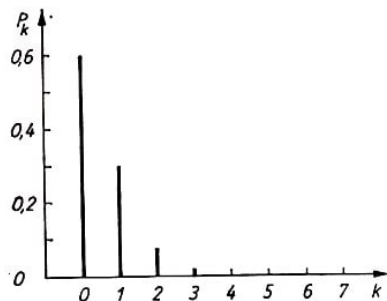


Abb. 27 Wahrscheinlichkeitsfunktion der Poisson-Verteilung für  $\lambda = 0,5$

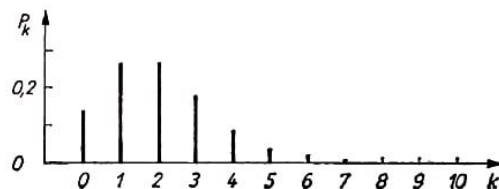


Abb. 28 Wahrscheinlichkeitsfunktion der Poisson-Verteilung für  $\lambda = 2,0$

Wir erkennen aus diesen beiden Wahrscheinlichkeitsverteilungen bereits, dass die Poisson-Verteilung bei kleinen Werten von  $\lambda$  stark unsymmetrisch ist und mit wachsendem  $\lambda$  immer mehr in eine symmetrische Verteilung übergeht.

Für die Poisson-Verteilung existieren ebenfalls Tabellen, aus denen die Wahrscheinlichkeiten  $p_k$  bei unterschiedlicher Größe von  $\lambda$  entnommen werden können.<sup>32</sup>

Die Aussage der Poisson-Verteilung wollen wir uns auch wieder an einem Beispiel deutlich machen.

Die rationelle Gestaltung der Produktionsprozesse erfordert den effektiven Einsatz der vorhandenen Maschinen und Geräte, wobei auch die Störanfälligkeit der Maschinen zu beachten ist. Treten Störungen auf, so müssen sie schnell behoben werden. In diesem Zusammenhang interessiert u. a. die Frage, wieviel Maschinen bzw. Geräte in einem bestimmten Zeitabschnitt, beispielsweise während einer Stunde oder während einer Schicht, Störungen aufweisen werden. Kann darüber eine Aussage getroffen werden, lassen sich Schlussfolgerungen für die erforderlichen Reparaturkapazitäten ableiten.

Um einen Überblick über die ausgefallenen Maschinen zu erhalten, werden die auftretenden Ausfälle über einen längeren Zeitraum ermittelt und analysiert. Zu diesem Zweck seien in einem Betrieb während eines Zeitraumes von insgesamt 220 Stunden die stündlich an den Maschinen aufgetretenen Störungen erfasst worden.

Die Anzahl der pro Stunde aufgetretenen Störungen ist in Tabelle 21, Spalte 2 festgehalten. Wir erkennen, dass in 47 Fällen keine Störung zu verzeichnen war, 78 mal eine Störung pro Stunde auftrat, 57 mal 2 Störungen pro Stunde zu verzeichnen waren usw. Durchschnittlich sind stündlich 1,5 Maschinen ausgefallen, also  $\bar{x} = 1,5$ .

<sup>32</sup>Vgl. z.B. Storm, R.: Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle. VEB Fachbuchverlag, Leipzig 1972, S. 321f.



Anzahl der ausgefallenen Maschinen	Beobachtete Häufigkeiten	Wahrschein- lichkeiten	Theoretische Häufigkeiten
$m$	$h_m$	$p_m$	$np_m$
(1)	(2)	(3)	(4)
0	47	0,223	49,06
1	78	0,335	73,70
2	57	0,251	55,22
3	25	0,126	27,72
4	7	0,047	10,34
5	3	0,014	3,08
6	2	0,003	0,66
7	0	0,001	0,22
8	1	0,000	0,00
$\Sigma$	220 = $n$	1,000	220,00

Tabelle 21 Anzahl ausgefallener Maschinen - beobachtete und theoretische Häufigkeiten (Poisson-Verteilung)

Von Interesse ist ein Vergleich dieser beobachteten Häufigkeiten mit den theoretischen Häufigkeiten. Unterstellen wir, dass die ermittelten Häufigkeiten einer Poisson-Verteilung genügen, so sind die theoretischen Häufigkeiten mit Hilfe von Formel (75) zu bestimmen.

Hierfür fehlt uns aber zunächst die Größe des Parameters  $\lambda$ . Da bei Poisson-Verteilung gemäß Formel (78)  $\mu = \lambda$  gilt, wobei  $\mu$  den Mittelwert der Grundgesamtheit charakterisiert, wird näherungsweise das durch die statistische Erhebung bestimmte arithmetische Mittel  $\bar{x}$  als  $\lambda$  angegeben.

In unserem Beispiel kann somit  $\lambda \approx \bar{x} = 1,5$  gesetzt werden.

Das ist aber nur gestattet, wenn auch geprüft wird, dass nach Formel (79) die Varianz  $\sigma^2$  in der Nähe von  $\bar{x}$  liegt. Im vorliegenden Beispiel ist das der Fall, denn die Varianz der Beobachtungswerte  $s^2 = 1,4$ . Sie ist ein Näherungswert für  $\sigma^2$ .

Nunmehr können die Wahrscheinlichkeiten  $p_k$  der Poisson-Verteilung entweder nach Formel (75) bestimmt oder aus einer Tabelle abgelesen werden. Die für  $\lambda = 1,5$  geltenden Werte sind in Spalte 3 der Tabelle 21 angegeben. Mit ihrer Hilfe wurden durch die Beziehung  $np_k$  die in Spalte 4 vermerkten theoretischen Häufigkeiten berechnet.

Ein Vergleich der beobachteten Häufigkeiten mit den theoretischen Häufigkeiten ergibt, dass beide nur relativ geringfügig voneinander abweichen und die beobachteten Werte näherungsweise einer Poisson-Verteilung entsprechen können.

Eine derartige Feststellung darf aber nicht auf einer Vermutung beruhen, sie muss eindeutig nachgewiesen werden. Die Annahme oder aber auch die Ablehnung einer solchen Hypothese wird in der Statistik durch Prüfverfahren erbracht.

Die im Kapitel 8 vorgenommene Prüfung der beobachteten Häufigkeiten (empirische Verteilung) unseres Beispiels auf Poisson-Verteilung lässt die Schlussfolgerung zu, dass tatsächlich Poisson-Verteilung vorliegt.

Entsprechen die Beobachtungswerte näherungsweise einer Poisson-Verteilung, dann lassen sie sich durch die mathematische Funktion der Poisson-Verteilung beschreiben. Zum Beispiel kann die Frage von Interesse sein, wie groß die Wahrscheinlichkeit dafür ist, dass pro Stunde eine bestimmte Anzahl von Maschinen (0, 1, 2, 3, oder 4 Maschinen) ausfällt.

Die Wahrscheinlichkeiten können in diesem Fall aus Spalte 3 der Tabelle 21 abgelesen werden. Solche Untersuchungen sind u.a. wichtig für den Einsatz von Wartungsingenieuren und Mechanikern. Sie besitzen auch Bedeutung für eventuell bereitzustellende Reservemaschinen.

So kann beispielsweise für den Einsatz von Mechanikern die Wahrscheinlichkeit dafür interessieren, dass pro Stunde nicht mehr als 3 Maschinen ausfallen.

Nach Gleichung (76) ergibt sich dann

$$\begin{aligned}\sum_{k=0}^3 P(X = k) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= \frac{1,5^0}{0!} e^{-1,5} + \frac{1,5^1}{1!} e^{-1,5} + \frac{1,5^2}{2!} e^{-1,5} + \frac{1,5^3}{3!} e^{-1,5} \\ &= 0,223 + 0,335 + 0,251 + 0,126 = 0,935\end{aligned}$$

Die Wahrscheinlichkeit dafür, dass pro Stunde höchstens 3 Maschinen ausfallen, beträgt  $0,935 = 93,5\%$ . Analog dazu ist die Wahrscheinlichkeit dafür, dass mehr als 3 Maschinen ausfallen, gleich  $6,5\%$ .

Zur optimalen Planung des Einsatzes von Mechanikern, dem Bereitstellen von Reservemaschinen usw. bietet die Bedienungstheorie günstige Möglichkeiten.

Viele bedienungstheoretische Modelle bauen darauf auf, dass die in ein Bedienungssystem einlaufenden Forderungen (in unserem Fall sind das die ausgefallenen und auf den Beginn der Reparatur - also auf Bedienung - wartenden Maschinen) bezüglich ihrer Ankunftshäufigkeiten einer Poisson-Verteilung genügen.

Es gibt in der Industrie, im Verkehrswesen, in Dienstleistungsbereichen eine Vielzahl von Prozessen, die hinsichtlich bestimmter Eigenschaften einer Poisson-Verteilung entsprechen.

## 6.4 Inhalt und Zweck der Normalverteilung

Die Normalverteilung ist die bekannteste und zugleich auch die wichtigste Verteilung stetiger Zufallsgrößen. Sie wird häufig auch nach dem deutschen Mathematiker C. F. Gauß als Gaußsche Normalverteilung bezeichnet. Gauß hat für diese Verteilung als erster den mathematischen Beweis erbracht.

Da bei stetigen Verteilungen die Zufallsvariable nicht nur ganzzahlige Werte - wie bei der Binomial- und der Poisson-Verteilung -, sondern jeden Wert innerhalb eines bestimmten Intervalls annehmen kann, ist es nicht möglich, die Wahrscheinlichkeit für einen bestimmten Punkt zu errechnen.

Statt dessen wird die Wahrscheinlichkeit dafür berechnet, dass ein Punkt innerhalb eines bestimmten Intervalls liegt, und die Dichte der Verteilung wird für ein Intervall ermittelt. Das geschieht mit Hilfe der Dichtefunktion. Diese lautet für die Normalverteilung

$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{für } -\infty < x < \infty \quad (80)$$

Die Größen  $\mu$  und  $\sigma^2$  sind die Parameter der Normalverteilung. Bei bekannten Werten von  $\mu$  und  $\sigma^2$  ist die Gestalt der Dichte völlig bestimmt.

Die Verteilungsfunktion der Normalverteilung lautet

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1 \quad (81)$$

Es bedeuten

$\varphi(x)$ : Wahrscheinlichkeitsdichte

$\Phi(x)$ : Verteilungsfunktion  
 $x$ : Wert der Zufallsgröße  $X$   
 $e$ : Konstante (Wert: 2,718)  
 $\sigma^2$ : Varianz der Verteilung  
 $\mu$ : Mittelwert der Verteilung  
 $\pi$ : Konstante (Wert: 3,14).

Die grafische Darstellung der Normalverteilung ergibt immer eine Glockenkurve (vgl. z. B. Abb. 9 oder 29).

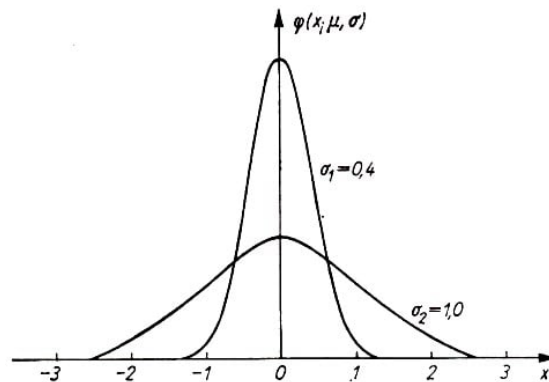


Abb.29 Dichte  $\varphi(x; \mu, \sigma^2)$  der Normalverteilung für  $\mu = 0$  und die Standardabweichungen  $\sigma_1 = 0,4$  bzw.  $\sigma_2 = 1,0$

Diese wird mit steigender Varianz flacher. Die praktische Bedeutung der Normalverteilung liegt vor allem auf dem Gebiet der statistischen Prüf- und Schätzverfahren. Darauf wird in den folgenden Kapiteln noch ausführlich eingegangen.

Wenn z. B. wiederholt Messungen an einem Gegenstand durchgeführt werden (Messungen einer Länge, eines Durchmessers, der Elastizität usw.), ergibt sich nicht bei jeder Messung der gleiche Wert. Die Messwerte weisen kleinere und größere Abweichungen voneinander und vom tatsächlichen Wert auf.

Soweit die Abweichungen nicht durch systematische Fehler, wie falsche Einstellung des Messgerätes, Ungenauigkeiten in der Skale des Messgerätes usw., hervorgerufen werden, ergeben sie sich aus zufälligen Fehlern, die sich auch bei größter Exaktheit der Messungen nicht völlig vermeiden lassen. Sie resultieren vor allem aus geringfügigen Messfehlern.

Die zufälligen Fehler entsprechen in den meisten Fällen zumindest näherungsweise einer Normalverteilung. Das trifft z. B. auch auf das im Kapitel 4 erläuterte Beispiel der Abweichungen vom Nennmaß eines Werkstückes zu.

Bei praktischen Untersuchungen wird in der Regel mit der normierten Normalverteilung gearbeitet.

Die Normierung erfolgt durch die Substitution

$$u = \frac{x - \mu}{\sigma} \quad (82)$$

wobei  $\mu = 0$  und  $\sigma = 1$  gesetzt werden. Dadurch entsteht die normierte Dichtefunktion

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \quad (83)$$

Für diese Funktion existieren Tabellen, aus denen die Werte für  $\varphi(u)$  bei unterschiedlicher Größe von  $u$  entnommen werden können (vgl. Tab. 22 und Abb. 30).

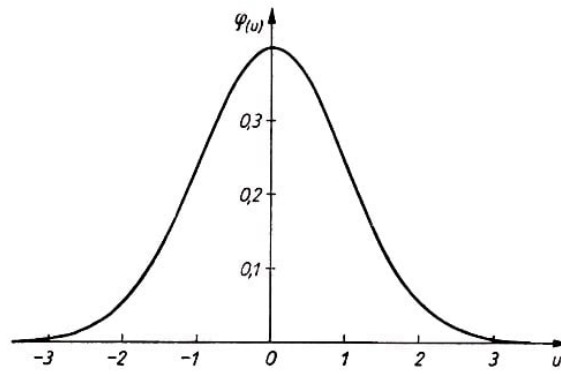


Abb.30 Dichtefunktion  $\varphi(u)$  der normierten Normalverteilung

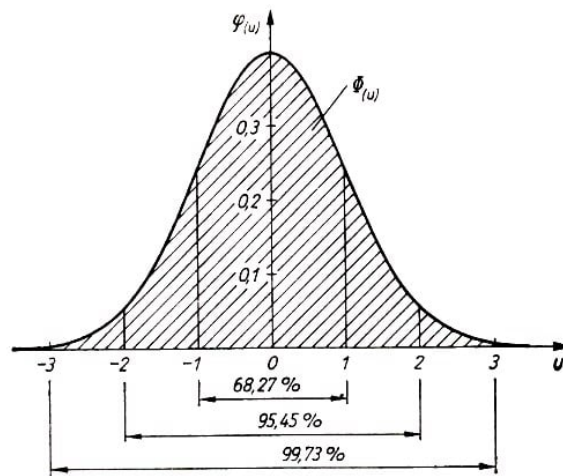


Abb.31 Verteilungsfunktion  $\Phi(u)$  der normierten Normalverteilung

$u$	$\varphi(u)$	$u$	$\varphi(u)$	$u$	$\Phi(u)$	$u$	$\Phi(u)$
0,0	0,3989	2,0	0,0540	0,0	0,0000	2,0	0,9545
0,1	0,3970	2,1	0,0440	0,1	0,0797	2,1	0,9643
0,2	0,3910	2,2	0,0355	0,2	0,1585	2,2	0,9722
0,3	0,3814	2,3	0,0283	0,3	0,2358	2,3	0,9786
0,4	0,3683	2,4	0,0224	0,4	0,3108	2,4	0,9836
0,5	0,3521	2,5	0,0175	0,5	0,3829	2,5	0,9876
0,6	0,3332	2,6	0,0136	0,6	0,4515	2,6	0,9907
0,7	0,3123	2,7	0,0104	0,7	0,5161	2,7	0,9931
0,8	0,2897	2,8	0,0079	0,8	0,5763	2,8	0,9949
0,9	0,2661	2,9	0,0060	0,9	0,6319	2,9	0,9963
1,0	0,2420	3,0	0,0044	1,0	0,6827	3,0	0,9973
1,1	0,2179	3,1	0,0033	1,1	0,7287	3,5	0,9995
1,2	0,1942	3,2	0,0024	1,2	0,7699	4,0	0,9999
1,3	0,1714	3,3	0,0017	1,3	0,8064		
1,4	0,1497	3,4	0,0012	1,4	0,8385		
1,5	0,1295	3,5	0,0009	1,5	0,8664		
1,6	0,1409	4,0	0,0001	1,6	0,8904		
1,7	0,0940			1,7	0,9109		
1,8	0,0790			1,8	0,9281		
1,9	0,0656			1,9	0,9426		

Tabelle 22 Dichtefunktion  $\varphi(u)$  der normierten Normalverteilung

Tabelle 23 Verteilungsfunktion  $\Phi(u)$  der normierten Normalverteilung mit den Integrationsgrenzen  $-u$  bis  $+u$

Die normierte Normalverteilung hat folgende Verteilungsfunktion

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-u}^{+u} e^{-\frac{u^2}{2}} du \quad (84)$$

Auch die Werte für  $\Phi(u)$  liegen tabelliert vor (vgl. Tab. 23 und Abb. 31).

Aus Tabelle 23 ist z. B. zu erkennen, dass

- 68,27% aller Werte im Bereich von  $u \leq |1|$  und 27,63% aller Werte außerhalb des Bereiches von  $u \leq |1|$  liegen;
- 95,45% aller Werte im Bereich von  $u \leq |2|$  und 4,55% aller Werte außerhalb des Bereiches von  $u \leq |2|$  liegen;
- 99,73% aller Werte im Bereich von  $u \leq |3|$  und 0,27% aller Werte außerhalb des Bereiches von  $u \leq |3|$  liegen;
- nahezu 100% aller Werte im Bereich von  $u \leq |4|$  liegen.

Diese Werte sind vor allem für das Festlegen der statistischen Sicherheit bei statistischen Untersuchungen von Bedeutung, worauf wir im folgenden Kapitel noch ausführlich zu sprechen kommen. Häufig interessieren nicht nur die Werte  $\Phi(u)$  des Integrals

$$\frac{1}{\sqrt{2\pi}} \int_{-u}^{+u} e^{-\frac{u^2}{2}} du$$

sondern auch die Werte

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+u} e^{-\frac{u^2}{2}} du \quad (85)$$

Sie werden u. a. für das Bestimmen der Wahrscheinlichkeiten bei der Prüfung empirischer Verteilungen auf Normalverteilung benötigt (vgl. Kap. 8) und liegen, wie Tabelle 24 zeigt, ebenfalls berechnet vor.

Ein Einblick in die praktische Bedeutung der Normalverteilung soll wiederum durch ein Beispiel gegeben werden.

Im Rahmen eines Arbeitsstudiums wurde durch 200 Stichproben die Dauer einer Arbeitsverrichtung gemessen. Dabei wurden die in Tabelle 25, Spalte 3 dargestellten Häufigkeiten festgestellt. Entsprechend den in den Spalten 4 bis 7 enthaltenen Rechenschritten, sind daraufhin die theoretischen Häufigkeiten der Normalverteilung für dieses Beispiel bestimmt worden.

$u$	$\Phi(u)$	$u$	$\Phi(u)$
-4,0	>0,0000	0,0	0,5000
-3,5	0,0003	+0,1	0,5398
-3,0	0,0014	+0,2	0,5793
-2,9	0,0019	+0,3	0,5179
-2,8	0,0026	+0,4	0,6554
-2,7	0,0035	+0,5	0,6915
-2,6	0,0047	+0,6	0,7257
-2,5	0,0062	+0,7	0,7580
-2,4	0,0082	+0,8	0,7881
-2,3	0,0107	+0,9	0,8159
-2,2	0,0139	+1,0	0,8413
-2,1	0,0179	+1,1	0,8643
-2,0	0,0227	+1,2	0,8849
-1,9	0,0287	+1,3	0,9032
-1,8	0,0359	+1,4	0,9192
-1,7	0,0446	+1,5	0,9332
-1,6	0,0548	+1,6	0,9452
-1,5	0,0668	+1,7	0,9554
-1,4	0,0808	+1,8	0,9641
-1,3	0,0968	+1,9	0,9713
-1,2	0,1151	+2,0	0,9772
-1,1	0,1357	+2,1	0,9821
-1,0	0,1587	+2,2	0,9861
-0,9	0,1841	+2,3	0,9893
-0,8	0,2119	+2,4	0,9918
-0,7	0,2420	+2,5	0,9938
-0,6	0,2742	+2,6	0,9953
-0,5	0,3085	+2,7	0,9965
-0,4	0,3446	+2,8	0,9974
-0,3	0,3821	+2,9	0,9981
-0,2	0,4207	+3,0	0,9986
-0,1	0,4602	+3,5	0,9997
0,0	0,5000	+4,0	<1,0000

Tabelle 24 Verteilungsfunktion  $\Phi(u)$  der Normalverteilung mit den Integrationsgrenzen  $-\infty$  bis  $+u$

Zunächst war nach der angegebenen Formel das Argument  $\Phi(a_m)$  zu bestimmen, wobei  $e_{mo}$  die obere Klassengrenze der  $m$ -ten Klasse,  $\mu$  den geschätzten Mittelwert der Grundgesamtheit und  $\sigma$  die geschätzte Standardabweichung der Grundgesamtheit angeben.

Die Parameter  $\mu_0$  und  $\sigma_0$  wurden näherungsweise durch das arithmetische Mittel  $\bar{x}$  und die Standardabweichung  $s$  der Stichprobe bestimmt:

$$\mu_0 \approx \bar{x} = 36,98 \text{ s} \quad , \quad \sigma_0 \approx s = 1,58 \text{ s}$$

Es können dann die Werte  $\Phi(a_m)$  aus der entsprechenden Tafel der Normalverteilung abgelesen werden. In Tabelle 24 sind jedoch die  $u$ -Werte (entsprechen im konkreten Beispiel den  $a_m$ -Werten) nur in Abständen von einem Zehntel und nicht in Hunderts-tel wie bei den Argumenten unseres Beispiels angegeben. Die exakten Werte für die

Klasse $m$	Dauer (s) $e_{mu}e_{mo}$	Beobachtete Häufigkeiten $h_m$	Argumente $a_m = \frac{e_{mo} - e_{mu}}{\sigma_0}$	Verteilungsfunktion $\Phi(a_m)$	Wahrscheinlichkeiten $p_m = \Phi(a_m) - \Phi(a_{m-1})$	Theoretische Häufigkeiten $np_m$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	32...33	1	-2,52	0,0059	0,0059	1,18
2	33...34	2	-1,89	0,0294	0,0235	4,70
3	34...35	17	-1,25	0,1057	0,0763	15,26
4	35...36	34,5	-0,62	0,2673	0,1616	32,32
5	36...37	45,5	+0,01	0,5040	0,2367	47,34
6	37...38	52	+0,65	0,7422	0,2382	47,64
7	38...39	29	+1,28	0,8927	0,1505	30,10
8	39...40	13	+1,91	0,9719	0,0792	15,84
9	40...41	4	+2,54	0,9945	0,0226	4,52
10	41...42	1	+3,18	0,9993	0,0048	0,96
11	42...43	1	+3,81	0,9999	0,0006	0,12
12	43...44	0	+4,44	<1,0000	0,0001	0,02
$\Sigma$		200,0 = $n$			1,0000	200,00

Fällt ein Wert auf eine Klassengrenze, dann wird er beiden Klassen je zur Hälfte zugeordnet.

Tabelle 25 Dauer von Arbeitsverrichtungen - beobachtete und theoretische Häufigkeiten (Normalverteilung)

Argumente wurden daher statistischen Tafelwerken entnommen.<sup>33</sup> Die abgelesenen Tafelwerte enthält Spalte 5.

Nachdem die Argumente  $\Phi(a_m)$  vorliegen, wird in Spalte 6 die Differenz der jeweils übereinanderstehenden Werte  $\Phi(a_m)$  und  $\Phi(a_{m-1})$  gebildet. Sie ergibt die theoretische Wahrscheinlichkeit der  $m$ -ten Klasse, mit deren Hilfe in Spalte 7 die theoretischen Häufigkeiten berechnet werden konnten.

Die im Kapitel 8 für dieses Beispiel durchgeführte Prüfung mit Hilfe des  $\chi^2$ -Tests lässt den Schluss zu, dass die Stichprobe aus einer normalverteilten Grundgesamtheit stammt und die Werte annähernd normalverteilt sind.

Wenn Normalverteilung nachgewiesen wurde, können darauf aufbauend weitere Berechnungen durchgeführt werden. Im eben dargestellten Beispiel kann z. B. die Wahrscheinlichkeit dafür interessieren, dass ein gemessener Zeitwert zwischen  $x_1 = 35$  s und  $x_2 = 40$  s liegt.

Es betragen:  $\mu = 36,98$  s;  $\sigma = 1,58$  s. Hierzu werden die Werte  $x_1$  und  $x_2$  mit Hilfe der Beziehung  $u = \frac{x-\mu}{\sigma}$  normiert. Folglich ergibt sich

$$u_1 = \frac{35 - 36,98}{1,58} = -1,25 \quad , \quad u_2 = \frac{40 - 36,98}{1,58} = +1,91$$

Es ist zu bilden

$$\Phi(1,91) - \Phi(-1,25) = 0,9719 - 0,1057 = 0,8662$$

Somit ist

$$P(35 < x < 40) = 0,8662 = 86,62\%$$

Der gemessene Zeitwert liegt also mit einer Wahrscheinlichkeit von 86,62% zwischen 35 und 40 Sekunden.

---

<sup>33</sup>Vgl. ebenda, S. 326f.



## 7 Wenige Elemente geben Auskunft über die Gesamtheit

Bei statistischen Erhebungen ist eine vollständige Erfassung (Totalerhebung) aller Elemente eines Untersuchungsobjektes oft nicht zweckmäßig oder gar nicht möglich. Das ist z. B. dann der Fall, wenn die untersuchten Elemente bei der Prüfung zerstört werden, wenn die Erfassung der Elemente verhältnismäßig hohe Kosten verursacht oder die Zahl der Elemente sehr groß ist. In diesen Fällen wird eine Teilerhebung - eine Stichprobe - durchgeführt.

Totalerhebungen statistischer Gesamtheiten großen Umfangs erfordern neben einem großen Aufwand an finanziellen Mitteln und Arbeitskräften auch einen relativ langen Bearbeitungszeitraum.

Oft kommt es aber gerade darauf an, die Ergebnisse einer statistischen Erhebung recht schnell zu erhalten. Eine Stichprobe kann in verhältnismäßig kurzer Zeit aufbereitet werden und erfordert in jeder Hinsicht einen geringeren Aufwand als eine Totalerhebung.

Daher sprechen in zahlreichen Fällen die Erfordernisse ganz eindeutig für Teilerhebungen, wobei allerdings ein Fehler, der durch das Beschränken auf einen Teil der Gesamtheit bedingt ist, in Kauf genommen werden muss. Wir wollen uns deshalb in diesem Kapitel mit einigen grundsätzlichen Problemen von Stichprobenerhebungen und mit den Aussagen von Stichproben befassen.

### 7.1 Was ist eine Stichprobe?

Wir wissen jetzt bereits, dass man Teilerhebungen als Stichproben bezeichnet. Der Begriff Stichprobe tritt im täglichen Leben sehr oft auf. Durch eine Stichprobe wird in einem Milchgeschäft geprüft, ob die Milch bereits sauer geworden ist oder nicht. In den Straßenbahnen wird durch Stichproben kontrolliert, ob alle Fahrgäste im Besitz eines gültigen Fahrtausweises sind.

Durch Stichproben werden die Abläufe chemischer Prozesse überwacht oder Bodenproben entnommen.

Diese Beispiele könnten beliebig fortgesetzt werden. Aber nicht immer, wenn das Wort Stichprobe fällt, handelt es sich tatsächlich um Stichproben im Sinne der mathematischen Statistik. Auch in Fällen, wo man fest überzeugt ist, dass eine statistische Stichprobe vorliegt, stellt sich bei genauer Analyse häufig heraus, dass die an eine Stichprobe im Sinne der mathematischen Statistik zu stellenden konkreten Anforderungen nicht erfüllt werden.

Durch die Stichproben sollen in der Regel Angaben über ein bestimmtes Merkmal der Gesamtheit gewonnen werden. Aus Kapitel 2 ist uns bereits bekannt, dass in der Statistik für derartige Gesamtheiten der Begriff Grundgesamtheit verwendet wird, wobei die Grundgesamtheit die Menge aller möglichen Realisationen einer Zufallsvariablen bildet.

Eine Stichprobe muss ein möglichst wahrheitsgetreues Abbild der Grundgesamtheit lie-

fern. Man sagt dafür auch, die Stichprobe muss für die Grundgesamtheit repräsentativ sein. Eine derartige repräsentative Stichprobe ist nur durch die zufällige Auswahl der Elemente aus der Grundgesamtheit und einen ausreichend großen Umfang der Stichprobe zu erreichen.

Stichproben im Sinne der Statistik sind Teilerhebungen, die durch zwei Merkmale charakterisiert werden:

- Die Auswahl der Elemente, die aus der Grundgesamtheit in die Stichprobe gelangen, erfolgt rein zufällig. Das bedeutet auch, dass jedes Element der Grundgesamtheit eine bestimmte berechenbare und von Null verschiedene Wahrscheinlichkeit besitzt, in die Stichprobe zu gelangen.
- Der Fehlerbereich der Stichprobenerhebung, der durch die Beschränkung auf einen Teil der Grundgesamtheit entsteht, ist berechenbar.

Wir erkennen daraus, dass z. B. überall dort keine Stichprobe im Sinne der Statistik vorliegt, wo ein Prozess lediglich durch eine ein- oder zweimalige Kontrolle dahingehend geprüft wird, wie weit er bereits fortgeschritten ist. Statistische Stichproben dienen der Analyse von Massenerscheinungen, wobei eine bestimmte Zahl von Elementen aus der Grundgesamtheit ausgewählt wird.

Die für eine repräsentative Aussage notwendige Zahl zu erfassender Elemente, der erforderliche Stichprobenumfang  $n$ , kann ebenfalls berechnet werden (siehe Abschnitt 7.2.).

Für die zufällige Auswahl der Elemente aus der Grundgesamtheit bestehen verschiedene Möglichkeiten. Zum Beispiel können wir aus einem Warenposten die zu prüfenden Elemente willkürlich herausgreifen. Oft verwendet man dazu auch sogenannte Zufallszahlen.

Die Elemente werden in diesem Fall fortlaufend nummeriert, und die Stichprobe wird entsprechend den Zahlen einer Zufallszahlentafel ausgewählt. Eine Zufallsauswahl ist aber auch gewährleistet, wenn wir das erste Element aus einer Grundgesamtheit zufällig auswählen und dann jedes  $i$ -te Element in die Stichprobe einbeziehen.

Aus einer Kartei kann z. B. jede 10. oder 20. Karte entnommen werden, um daraus bestimmte Daten zu ermitteln. Bei der Analyse von Kundenwünschen kann jeder  $i$ -te Kunde, bei der Qualitätskontrolle jedes  $i$ -te Erzeugnis, bei Verkehrskontrollen jedes  $i$ -te Fahrzeug erfasst werden. Man spricht bei den hier erläuterten Auswahlmöglichkeiten von einer reinen Zufallsauswahl.

Sie lässt sich durch verschiedene Auswahltechniken realisieren. So bezeichnet man die "willkürliche" Auswahl der Elemente bzw. die Auswahl mit Hilfe von Zufallszahlen als uneingeschränkte Zufallsauswahl und die Auswahl jedes  $i$ -ten Elementes in der erläuterten Form als systematische Auswahl.

Von einer reinen Zufallsauswahl wollen wir in den folgenden Betrachtungen ausgehen. Wir müssen allerdings betonen, dass es bei bestimmten praktischen Untersuchungen zweckmäßig sein kann, ein anderes Auswahlverfahren (geschichtete Auswahl, mehrstufige Auswahl) anzuwenden, bei dem die Struktur der Grundgesamtheit in gewisser Hinsicht berücksichtigt wird und darauf dann die Zufallsauswahl aufbaut. In der statis-

tischen Arbeit mit diesen Auswahlverfahren sind einige Besonderheiten zu beachten.

Eine Stichprobenerhebung zielt in der Regel darauf ab, durch die Stichprobe eine Information über die Grundgesamtheit zu erhalten. Die Stichprobe soll mittels ihrer Maßzahlen (z. B. des Mittelwertes und der Standardabweichung) Schlussfolgerungen auf diese Größen in der Grundgesamtheit gestatten.

So ist es z. B. mit Hilfe einer Stichprobe möglich, Angaben über die Qualität hergestellter Erzeugnisse, über das Durchschnittsalter von Werkträgern einer bestimmten Lohngruppe, über die Durchschnittsgeschwindigkeit von Kraftfahrzeugen auf einem bestimmten Streckenabschnitt und über viele andere Größen zu gewinnen.

Dieser Schluss von der Stichprobe auf die Grundgesamtheit ist der sogenannte Repräsentationsschluss, auch als Rückschluss oder indirekter Schluss bezeichnet. Er gilt als der typische Stichprobenschluss.

Neben dem Rückschluss kann aber auch die umgekehrte Schlussweise angewendet und von der Grundgesamtheit auf eine daraus abgeleitete Stichprobe geschlossen werden. Dieser Schluss wird als Inklusionsschluss oder direkter Schluss bezeichnet (vgl. Abb. 32).

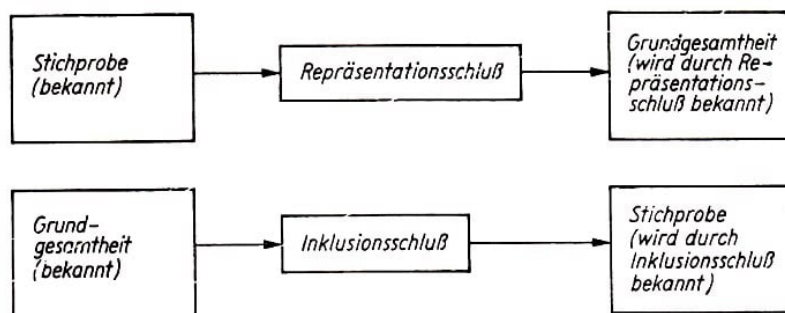


Abb. 32 Repräsentationsschluss und Inklusionsschluss

Zum Beispiel sei die durchschnittliche Ausschussquote bei der Herstellung eines Erzeugnisses bekannt oder vorgegeben. Mit Hilfe des Inklusionsschlusses ist es nunmehr möglich, festzustellen bzw. vorauszusagen, in welchen Grenzen die Ausschussquote für eine bestimmte Anzahl hergestellter Erzeugnisse liegen darf.

Wenn dieser Prozentsatz berechnet wurde, so kann durch eine Stichprobe beurteilt werden, ob beispielsweise die in einer Schicht hergestellten Erzeugnisse diese durchschnittlichen Qualitätsforderungen eingehalten haben oder nicht. Es wird geprüft, ob die Stichprobe mit einer Grundgesamtheit vereinbar ist, in der die vorgegebene durchschnittliche Ausschussquote nicht überschritten wird.

Der Repräsentationsschluss wird durch Schätzverfahren realisiert, der Inklusionsschluss durch die Prüfverfahren. Zwischen beiden bestehen enge Zusammenhänge.

In den folgenden beiden Abschnitten wollen wir uns mit dem Schätzen des Mittelwertes und dem Schätzen von Anteilswerten beschäftigen. Im Kapitel 8 werden dann einige grundsätzliche Fragen der statistischen Prüfverfahren erläutert.

Zunächst sind aber noch einige weitere Probleme zu nennen, die bei Stichprobenerhebungen beachtet werden müssen.

In der Stichprobentheorie ist zwischen dem Fall "mit Zurücklegen" und dem Fall "ohne Zurücklegen" zu unterscheiden (vgl. auch Abschnitt 6.2.).

Bei einer Stichprobe "mit Zurücklegen" wird der Grundgesamtheit jeweils nur ein Element entnommen und geprüft, es wird nach seiner Prüfung der Grundgesamtheit wieder zugeführt, dann wird das nächste Element entnommen usw. Bei Stichproben "ohne Zurücklegen" werden der Grundgesamtheit nacheinander ein oder mehrere Elemente entnommen und ihr nach erfolgter Prüfung nicht wieder zur Auswahl zugeführt.

Für die Praxis hat der Fall "ohne Zurücklegen" die weitaus größere Bedeutung, weil er sich besser realisieren lässt.

Ferner sei noch darauf hingewiesen, dass in der Stichprobentheorie eine Trennung zwischen kleinen Stichproben ( $n \leq 30$ ) und großen Stichproben ( $n > 30$ ) vorgenommen wird.

## 7.2 Wie schätzen wir den Mittelwert?

Die Maßzahlen und Parameter von Stichprobe und Grundgesamtheit stimmen im allgemeinen nicht genau überein. Zwischen den Mittelwerten, den Standardabweichungen oder den Anteilswerten von Stichprobe und Grundgesamtheit bestehen gewisse Abweichungen.

Unter bestimmten Bedingungen, die noch näher zu erläutern sind, wird diese Abweichung einen Maximalwert mit einer bestimmten angebbaren Wahrscheinlichkeit nicht überschreiten.

Diesen Maximalwert bezeichnet man als Stichprobenfehler  $a$ . Nehmen wir an, eine Stichprobe hat den Wert  $M$  geliefert, dann lässt sich ein Intervall  $M \pm a$  abgrenzen, das den gesuchten Parameter  $P$  der Grundgesamtheit mit der Wahrscheinlichkeit einschließt, mit der der Stichprobenfehler den Maximalwert  $a$  besitzt.

Da der Stichprobenfehler ein Intervall angibt, spricht man statt vom Stichprobenfehler häufig auch vom Fehlerbereich einer Stichprobe. Für den Stichprobenfehler schreibt man in allgemeiner Form

$$a = \max |M - P| \quad (86)$$

Der Stichprobenfehler ist für die Beurteilung der Aussagekraft von Teilerhebungen von großer Bedeutung. Deshalb wollen wir uns beim Schätzen des Mittelwertes (Erwartungswertes) einer Grundgesamtheit zunächst dem Stichprobenfehler zuwenden.

Es sei noch betont, dass in diesem Abschnitt unter dem Begriff Mittelwert immer das arithmetische Mittel zu verstehen ist.

Für den Mittelwert kann der Stichprobenfehler wie folgt formuliert werden

$$a_{\bar{x}} = \max |\bar{x} - \mu| \quad (87)$$

Zur Ermittlung dieser Differenz wurde folgende Formel entwickelt:

$$a_{\bar{x}} = \frac{u \cdot \sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (88)$$

Dabei bedeuten

$a_{\bar{x}}$ : absoluter Stichprobenfehler des arithmetischen Mittels

$u$ : Genauigkeitsfaktor, abhängig von der gewählten statistischen Sicherheit  $S$

$\sigma$ : Standardabweichung der Elemente in der Grundgesamtheit

$n$ : Umfang der Stichprobe

$N$ : Umfang der Grundgesamtheit

Der Ausdruck  $\sqrt{\frac{N-n}{N-1}}$  wird als Korrekturfaktor bezeichnet. Wenn die Grundgesamtheit sehr groß ist, kann auf den Korrekturfaktor verzichtet werden.

Oft interessiert für Vergleichszwecke auch der relative Stichprobenfehler. Dieser kann für das arithmetische Mittel nach der Formel

$$a_{\bar{x}_r} = \frac{a_{\bar{x}}}{\bar{x}} \quad (89)$$

bestimmt werden. Der Stichprobenfehler ist somit abhängig

- von der gewählten statistischen Sicherheit
- von der Standardabweichung der Grundgesamtheit
- vom Stichprobenumfang und
- von der Größe der Grundgesamtheit.

Der Stichprobenfehler vergrößert sich mit wachsender statistischer Sicherheit und steigender Standardabweichung, und er verringert sich mit der Erhöhung des Stichprobenumfanges bzw. des Auswahlatzes.

Wenden wir uns zunächst der statistischen Sicherheit  $S$  bzw. dem Genauigkeitsfaktor  $u$  zu. Unter der statistischen Sicherheit ist die Wahrscheinlichkeit dafür zu verstehen, dass der Unterschied zwischen dem arithmetischen Mittel der Grundgesamtheit und dem arithmetischen Mittel der Stichprobe nicht größer als der Stichprobenfehler  $a_{\bar{x}}$  wird. Oft wird statt der statistischen Sicherheit auch die Irrtumswahrscheinlichkeit  $\alpha$  angegeben. Die Irrtumswahrscheinlichkeit stellt die Ergänzung der statistischen Sicherheit zu 1 bzw. zu 100% dar

$$S + \alpha = 1$$

Für Stichprobenmittelwerte gilt, dass sie immer einer Normalverteilung entsprechen, wenn die Grundgesamtheit normalverteilt ist. Ist die Grundgesamtheit nicht normalverteilt, dann ist die Verteilung der Mittelwerte annähernd normalverteilt.

Die Annäherung ist um so besser, je größer der Stichprobenumfang gewählt wird. Bei der Bestimmung des Stichprobenfehlers des arithmetischen Mittels kann der Genauigkeitsfaktor deshalb für große Stichproben ( $n > 30$ ) aus der Verteilungsfunktion der Normalverteilung entnommen werden (vgl. Tab. 23). Für kleine Stichproben ( $n \leq 30$ ) folgt die normierte Variable der  $t$ -Verteilung (vgl. Kap. 8). Dafür gibt es ebenfalls Tabellen, die in statistischen Fachbüchern enthalten sind. In diesem Fall ist  $u$  durch  $t$  zu ersetzen.

Bei statistischen Untersuchungen ist man stets bestrebt, mit der nötigen statistischen Sicherheit zu arbeiten. Bei ihrer Festlegung ist von der Regel auszugehen, dass die

statistische Sicherheit um so höher zu wählen ist, je schwerwiegender die Folgen aus einem aus der Stichprobe abgeleiteten Fehltritt sein können.

Dabei muss man immer von praktischen und nicht von rechnerischen Überlegungen ausgehen. Wir kommen auf dieses Problem auf den folgenden Seiten noch ausführlicher zu sprechen. Im allgemeinen wird mit statistischen Sicherheiten zwischen 90 und 99,9% gearbeitet. Eine 100%ige statistische Sicherheit ist nicht erreichbar, weil in diesem Fall das Vertrauensintervall unendlich groß würde und der Fehler nicht eingegrenzt werden könnte (vgl. hierzu auch Abb. 31).

Der Stichprobenfehler ist weiter von der Standardabweichung der Grundgesamtheit abhängig, wobei sich der größere Stichprobenfehler bei wachsender Standardabweichung daraus ergibt, dass bei großer Standardabweichung in der Grundgesamtheit auch recht unterschiedliche Werte in die Stichprobe gelangen können, die dann einen großen Stichprobenfehler bedingen.

Eine kleine Standardabweichung lässt auf geringe Unterschiede der Elemente der Grundgesamtheit schließen, was auch einen kleinen Stichprobenfehler zur Folge hat.

Als dritte Größe, die für den Stichprobenfehler von Bedeutung ist, war der Stichprobenumfang  $n$  genannt worden. Es ist leicht einzusehen, dass entsprechend dem Gesetz der großen Zahlen der Fehler einer Stichprobe bei einer Vergrößerung des Stichprobenumfanges sinkt, während er bei der Verringerung des Stichprobenumfanges steigt.

Nunmehr noch einige Bemerkungen zum Korrekturfaktor  $K$ . Er ist für endliche Grundgesamtheiten von Bedeutung und durch die Größe

$$K = \sqrt{\frac{N - n}{N - 1}} \quad (90)$$

gegeben. Da  $N - 1 \approx N$ , kann man auch schreiben

$$K = \sqrt{1 - \frac{n}{N}} \quad (91)$$

Die Beziehung lässt erkennen, dass der Stichprobenfehler durch das Verhältnis von Stichprobenumfang zur Größe der Grundgesamtheit beeinflusst wird. Das Verhältnis  $\frac{n}{N}$ , das die Größe des Korrekturfaktors bestimmt, bezeichnet man als *Auswahlsatz*  $A$ . Die Abhängigkeit von Korrekturfaktor  $K$  und Auswahlsatz  $A$  ist in Tabelle 26 dargestellt.

Der Korrekturfaktor nähert sich immer mehr dem Wert 1, wenn der Umfang der Grundgesamtheit sehr groß gegenüber dem Stichprobenumfang ist bzw. der Auswahlsatz abnimmt.

$A$	$K$
1	0
0,9	0,3162
0,8	0,4472
0,7	0,5477
0,6	0,6325
0,5	0,7071
0,4	0,7746
0,3	0,8367
0,2	0,8944
0,1	0,9487
0,05	0,9747
0,01	0,9950
0,001	0,9995
0,0001	0,9999
...	...
0 ( $N = \infty$ )	1

Tabelle 26 Korrekturfaktor  $K$  in Abhängigkeit vom Auswahlsatz  $A$

Aus Tabelle 26 ist ersichtlich, dass der Korrekturfaktor spätestens ab  $A = 0,01$  gleich 1 gesetzt und damit bei der Berechnung vernachlässigt werden kann. Es ist auch zu erkennen, dass bei einer Totalerhebung ( $K = 0$ ) der Stichprobenfehler gleich Null wird. Bei großen Grundgesamtheiten können wir somit den Stichprobenfehler berechnen als

$$a_{\bar{x}} = \frac{u \cdot \sigma}{\sqrt{n}} \quad (92)$$

Die zum Stichprobenfehler gegebenen Erläuterungen sollen nunmehr an einem Beispiel demonstriert werden.

a) Auf einem Parkplatz ist durch eine Stichprobe von  $n = 200$  die Parkdauer der Kraftfahrzeuge bestimmt worden, um daraus Schlussfolgerungen für eine eventuell notwendig werdende Beschränkung der Parkdauer ziehen zu können. Die durchschnittliche Parkdauer hat 2,35 Stunden betragen; die Standardabweichung 1,7 Stunden. Der Stichprobenfehler soll nun mit einer statistischen Sicherheit von 95,45% bestimmt werden. Es sind also gegeben

$N$ : sehr groß;  $n$ : 200 Fahrzeuge,  $u$ : 2,00 ( $S = 3,45\%$ )  $\bar{x}$ : 2,35 Stunden,  $\sigma \approx s$ : 1,70 Stunden.

Der Stichprobenfehler beträgt dann

$$a_{\bar{x}} = \frac{u \cdot \sigma}{\sqrt{n}} = \frac{200 \cdot 1,70}{\sqrt{200}} = 0,24$$

Es gilt somit  $\mu = 2,35 \pm 0,24$  ( $S = 9,45\%$ ). Das Intervall 2,11...2,59 schließt den unbekannten Mittelwert der Grundgesamtheit, in unserem Fall die durchschnittliche Parkdauer, mit der Wahrscheinlichkeit von 95,45% ein.

Setzen wir  $\mu$  als in diesem Intervall befindlich an, dann wird dieser Ansatz durchschnittlich nur in vier oder fünf von hundert Fällen unzutreffend sein.

Der relative Stichprobenfehler beträgt

$$a_{\bar{x}} = \frac{a_{\bar{x}}}{\bar{x}} = \frac{0,24}{2,35} = 0,102 = 10,2\%$$

Der Stichprobenfehler macht somit 10,2% des berechneten Mittelwertes aus.

b) Beträgt die Zahl der auf dem Parkplatz in einem bestimmten Zeitraum (z.B. in einer Woche) geparkten Fahrzeuge 10000 und soll die Größe bei der Berechnung des Stichprobenfehlers mit beachtet werden, ergibt das mit den gleichen Größen wie unter a) bei Berücksichtigung des Korrekturfaktors

$$a_{\bar{x}} = \frac{u \cdot \sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{200 \cdot 170}{\sqrt{200}} \sqrt{\frac{10000-200}{10000-1}} = 0,24 \cdot 0,990 = 0,238 \approx 0,24$$

Die Abweichung gegenüber dem Rechnen mit unendlich großer Grundgesamtheit unter a) ist äußerst gering.

c) Bei einer Grundgesamtheit von 1000 Fahrzeugen verringert sich der Stichprobenfehler wie folgt:

$$a_{\bar{x}} = \frac{u \cdot \sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{200 \cdot 170}{\sqrt{200}} \sqrt{\frac{1000-200}{1000-1}} = 0,24 \cdot 0,895 = 0,21$$

Jetzt wird der Einfluss der Grundgesamtheit auf den Stichprobenfehler deutlich sichtbar. Der Stichprobenfehler ist beim Rechnen mit unendlicher Grundgesamtheit am größten und verringert sich auch bei verhältnismäßig großen Grundgesamtheiten nur unwesentlich.

Erst bei relativ kleinen Grundgesamtheiten verringert sich der Stichprobenfehler unter Beachtung des Korrekturfaktors spürbar.

Es ist also in der Regel sinnvoll, bei relativ großen Grundgesamtheiten den Korrekturfaktor zu vernachlässigen. In diesen Fällen wird der Stichprobenfehler geringfügig überhöht ausgewiesen, was sich in der Praxis nicht nachteilig bemerkbar macht.

d) Stützt sich die Erhebung, bei sonst gleichen Ausgangsgrößen wie unter a) auf einen Stichprobenumfang von 900 Fahrzeugen ( $n = 900$ ), ergibt das den Stichprobenfehler

$$a_{\bar{x}} = \frac{u \cdot \sigma}{\sqrt{n}} = \frac{200 \cdot 170}{\sqrt{900}} = 0,11$$

Der Stichprobenfehler ist gegenüber a) wesentlich geringer. Er beträgt nur noch 0,11 Stunden, die durchschnittliche Parkdauer wird mit einer statistischen Sicherheit von 95,45% vom Intervall  $2,35 \pm 0,11$  Stunden, also 2,24 ... 2,46 Stunden eingeschlossen.

Aus den bisherigen Betrachtungen ergab sich, dass wir mit Hilfe des Stichprobenfehlers einen Bereich angeben können, der den Mittelwert der Grundgesamtheit mit einer bestimmten statistischen Sicherheit einschließt.

Unter a) stellten wir beispielsweise fest, dass die durchschnittliche Parkdauer der Fahrzeuge mit einer statistischen Sicherheit von 95,452,11 ... 2,59 Stunden eingeschlossen



wird. Damit ist bereits eine Schätzung des Mittelwertes erfolgt.

Wenn durch eine Stichprobe das arithmetische Mittel  $\bar{x}$  berechnet wurde, dann ergibt sich für den Mittelwert der Grundgesamtheit das Intervall

$$\mu = \bar{x} \pm a_{\bar{x}} \quad (93)$$

Für diesen Ausdruck kann auch geschrieben werden

$$\bar{x} - a_{\bar{x}} < \mu < \bar{x} + a_{\bar{x}} \quad (94)$$

Es bedeuten

$\mu$ : Mittelwert der Grundgesamtheit

$\bar{x}$ : arithmetisches Mittel der Stichprobe

$a_{\bar{x}}$ : Stichprobenfehler.

Das durch diese Ungleichung eingeschlossene Intervall wird Konfidenz- oder Vertrauensintervall genannt. Es schließt den unbekannten Mittelwert der Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit (statistischen Sicherheit) ein. Der Mittelwert der Grundgesamtheit kann dabei an jeder Stelle des Konfidenzintervalles liegen.

Aus den zur Berechnung des Stichprobenfehlers für große Stichproben angegebenen Beziehungen können für das Konfidenzintervall bei der Schätzung des Mittelwertes folgende Formeln abgeleitet werden:

- für unendliche (sehr große) Grundgesamtheiten

$$\mu = \bar{x} \pm \frac{u \cdot \sigma}{\sqrt{n}} \quad (95)$$

- für endliche (relativ kleine) Grundgesamtheiten

$$\mu = \bar{x} \pm \frac{u \cdot \sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (96)$$

Betrachten wir aus dieser Sicht unser Beispiel, dann können wir die unter a) bis d) angegebenen Intervalle als die Konfidenz- oder Vertrauensintervalle für die durchschnittliche Parkdauer bezeichnen, die wir durch Schätzung ermittelt haben.

Eine solche Form der Schätzung, bei der für den zu ermittelnden Parameter ein Intervall bestimmt wird, bezeichnet man als Intervall- oder Konfidenzschätzung.

Außer der Intervallschätzung gibt es noch die Punktschätzung. Die Punktschätzung dient dazu, für einen Parameter einen ganz bestimmten Wert festzulegen. So mussten wir zur Berechnung des Stichprobenfehlers die unbekannte Standardabweichung  $\sigma$  der Grundgesamtheit schätzen und haben hierfür als Näherungswert die aus der Stichprobe gewonnene Standardabweichung  $s$  angegeben. Dabei wurde kein Intervall zugrunde gelegt, sondern ein fester Wert.

Punktschätzungen erfolgen vor allem dann, wenn mit dem Schätzwert noch weitere Berechnungen durchgeführt werden müssen. Von Nachteil bei den durch Punktschätzung

ermittelten Schätzwerten ist u.a., dass die Punktschätzung keine Anhaltspunkte über die Genauigkeit der Aussage liefert.

Wir wollen uns jetzt wieder dem Verhältnis von Stichprobenfehler und statistischer Sicherheit zuwenden. Gehen wir in unserem Beispiel von Aufgabe a) aus (also  $N =$  sehr groß,  $n = 200$  Fahrzeuge,  $\bar{x} = 2,35$  Stunden,  $\sigma \approx s = 1,70$  Stunden) und verändern die statistische Sicherheit, dann können wir für das Konfidenzintervall schreiben

$$2,35 - u \frac{1,70}{\sqrt{200}} < \mu < 2,35 + u \frac{1,70}{\sqrt{200}}$$

bzw.

$$2,35 - u \cdot 0,12 < \mu < 2,35 + u \cdot 0,12$$

Es ergibt sich im einzelnen

- für  $S = 68,27\%$

$$2,35 - 1,00 \cdot 0,12 < \mu < 2,35 + 1,00 \cdot 0,12 \quad , \quad 2,23 < \mu < 2,47$$

- für  $S = 95,45\%$

$$2,35 - 2,00 \cdot 0,12 < \mu < 2,35 + 2,00 \cdot 0,12 \quad , \quad 2,11 < \mu < 2,59$$

- für  $S = 99,00\%$

$$2,35 - 2,58 \cdot 0,12 < \mu < 2,35 + 2,58 \cdot 0,12 \quad , \quad 2,04 < \mu < 2,66$$

- für  $S = 99,73\%$

$$2,35 - 3,00 \cdot 0,12 < \mu < 2,35 + 3,00 \cdot 0,12 \quad , \quad 1,99 < \mu < 2,71$$

Bei hohen Anforderungen an die statistische Sicherheit ist das entsprechende Konfidenzintervall verhältnismäßig groß, wir haben einen großen Stichprobenfehler zu verzeichnen. Andererseits ist dann aber die Wahrscheinlichkeit gering, dass der durch eine Stichprobe ermittelte Wert außerhalb des Vertrauensintervalls liegt.

Bei einer statistischen Sicherheit von 99,00% ist zu erwarten, dass der durch die Stichprobe bestimmte Mittelwert nur in 1% aller Fälle nicht im berechneten Intervall liegt (Irrtumswahrscheinlichkeit von 1%).

Wird die statistische Sicherheit verringert, verkleinert sich zwar der Stichprobenfehler, gleichzeitig verringert sich aber auch die Wahrscheinlichkeit dafür, dass sich der festgestellte Mittelwert innerhalb des Konfidenzintervalls befindet. Bei einer statistischen Sicherheit von 68,27% wird er also in fast 32% der Fälle nicht im betreffenden Intervall liegen. Die Aussage ist dann wenig brauchbar.

Wir hatten bei der Ermittlung des Stichprobenfehlers schon betont, dass die statistische Sicherheit um so höher zu wählen ist, je schwerwiegender die Folgen bei einem aus der Stichprobe abgeleiteten Fehlurteil sein können, und wir hierbei von praktischen und nicht von rechnerischen Überlegungen ausgehen müssen.

Beispielsweise wird man bei einer überschlägigen Ermittlung der durchschnittlichen Parkzeit auf einem Parkplatz oder der Wartezeit an einem Schalter eine geringere statistische Sicherheit wählen können als bei einer Ermittlung der durchschnittlichen Dauer einzelner Arbeitsverrichtungen als Grundlage für den Aufbau von Zeitnormativen.

Wir wollen bei praktischen Aufgaben natürlich auch den Stichprobenfehler nicht allzu groß werden lassen, weil dann ein wenig brauchbares Konfidenzintervall für die zu ermittelnde Größe entsteht. Wie wir gesehen haben, dürfen wir den Stichprobenfehler in der Regel jedoch nicht "zu Lasten" der statistischen Sicherheit verringern.

Man nimmt deshalb auf den Stichprobenfehler durch einen entsprechend großen Stichprobenumfang Einfluss. Wir wissen, dass sich der Stichprobenfehler bei steigendem Umfang der Stichprobe verkleinert. Damit können wir den Stichprobenfehler bei einer Schätzung so klein wie vorgesehen halten. Auch hierbei geht man von praktischen Überlegungen aus.

Wenn eine statistische Erhebung in Form einer Stichprobe durchgeführt werden soll, ist deshalb zunächst vor allem die Frage von Interesse, wieviel Elemente denn überhaupt zu erfassen sind, um ein repräsentatives Ergebnis zu erhalten. Es muss der erforderliche Stichprobenumfang berechnet werden.

Durch Umstellung der Formel (92) erhalten wir für große Grundgesamtheiten zur Berechnung des Stichprobenumfangs die Beziehung

$$n = \frac{u^2 \cdot \sigma^2}{a_{\bar{x}}^2} \quad (97)$$

Die statistische Sicherheit und der Stichprobenfehler werden vorgegeben. Die Varianz der Grundgesamtheit ist unbekannt. Sie wird aus früheren Erhebungen übernommen oder durch eine Probeerhebung näherungsweise bestimmt.

Für die Berechnung des Stichprobenumfangs ein Beispiel:

Es soll das Gewicht der männlichen Personen einer bestimmten Altersgruppe mit einer statistischen Sicherheit von 95% bestimmt werden. Der Stichprobenfehler soll 0,5 kg nicht überschreiten. Die aus älteren Untersuchungen übernommene Standardabweichung beträgt 6,5 kg.

Gegeben sind

$N$ : sehr groß;  $u$ : 1,96 ( $S = 95,00\%$ );  $\sigma$ : 6,5 kg;  $a_{\bar{x}}$ : 0,5 kg

Dann ist

$$n = \frac{u^2 \cdot \sigma^2}{a_{\bar{x}}^2} = \frac{1,96^2 \cdot 6,5^2}{0,5^2} = 649,0 \approx 650$$

Nach den Prinzipien der Zufallsauswahl ist somit das Gewicht von 650 Personen der betreffenden Altersgruppe zu ermitteln, um die geforderten Parameter einzuhalten.

Bei der Schätzung von Parametern wird in der Praxis zweckmäßig so vorgegangen, dass man zuerst eine Stichprobe mit bestimmten Eigenschaften hinsichtlich der statistischen Sicherheit und des Stichprobenfehlers plant, darauf aufbauend den erforderlichen Stichprobenumfang berechnet und erst dann die Stichprobe durchführt.

### 7.3 Wir schätzen Anteilswerte

Zahlreichen praktischen Untersuchungen liegt eine alternative Fragestellung zugrunde, z. B. der Frage nach dem Anteil der weiblichen und der männlichen Beschäftigten in

einer bestimmten Berufsgruppe, der Frage nach dem Besitz eines Fernsehgerätes (ja, nein), der Frage nach dem Wachstum von Setzlingen in der Biologie (gut, schlecht), der Frage nach dem Anteil der mit Qualitätsmängeln behafteten Produkte eines Warenpostens usw.

Auch hier interessiert, welche Aussagekraft die Stichprobenergebnisse für die betreffende Grundgesamtheit besitzen. Deshalb wollen wir uns jetzt nach dem Schätzen des Mittelwertes quantitativer Merkmale mit dem Schätzen des Anteilswertes alternativer, also qualitativer Merkmale beschäftigen.

Beim Schätzen des Anteilswertes  $p$  einer Grundgesamtheit wird das Konfidenzintervall durch die Beziehung

$$p = p' \pm a_p \quad (98)$$

bestimmt. Es bedeuten

$p$ : Anteilswert mit der Eigenschaft  $A$  in der Grundgesamtheit

$p'$ : Anteilswert mit der Eigenschaft  $A$  in der Stichprobe

$a_p$ : Stichprobenfehler.

Der Stichprobenfehler  $a_p$  kann nach folgender Formel berechnet werden:

$$a_p = \frac{u\sqrt{pq}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (99)$$

Hierauf treffen die Gesetzmäßigkeiten der Binomialverteilung bzw. der hypergeometrischen Verteilung zu (vgl. Abschn. 6.2.). Es bedeuten

$a_p$ : Stichprobenfehler des Anteilswertes  $p$

$u$ : Genauigkeitsfaktor, abhängig von der gewählten statistischen Sicherheit  $S$

$p$ : Anteilswert mit der Eigenschaft  $A$

$q$ : Anteilswert mit der Eigenschaft  $\bar{A}$

$n$ : Umfang der Stichprobe

$N$ : Umfang der Grundgesamtheit.

Für die jeweiligen Größen der Formel haben sinngemäß die gleichen Regeln Gültigkeit, die bereits bei der Berechnung des Stichprobenfehlers für das arithmetische Mittel erläutert worden sind. Bei sehr großer Grundgesamtheit kann auch hier den Korrekturfaktor  $\sqrt{\frac{N-n}{N-1}}$  verzichtet werden, und wir erhalten

$$a_p = \frac{u\sqrt{pq}}{\sqrt{n}} \quad (100)$$

An Stelle der Standardabweichung  $\sigma$  beim Berechnen des Stichprobenfehlers des arithmetischen Mittels ist jetzt der Ausdruck  $\sqrt{pq}$  getreten, der auf die Standardabweichung der Binomialverteilung zurückgeht, die für das Ermitteln von Anteilswerten Bedeutung hat.

Dabei ergibt sich der größte Stichprobenfehler bei  $p = q = 0,5$ . Jedes andere Verhältnis zwischen  $p$  und  $q$  liefert für den Stichprobenfehler kleinere Werte. Man setzt für  $p$  und  $q$  in der Regel die durch die Stichprobe ermittelten Anteile  $p'$  und  $q'$  ein.

Damit erhalten wir für große Stichproben durch ähnliche Überlegungen wie im Abschnitt 7.2. folgende Formeln zur Ermittlung des Konfidenzintervalls:

- für unendliche (sehr große) Grundgesamtheiten

$$p = p' \pm \frac{u\sqrt{p'q'}}{\sqrt{n}} \quad (101)$$

- für endliche (relativ kleine) Grundgesamtheiten

$$p = p' \pm \frac{u\sqrt{p'q'}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (102)$$

Wir wollen folgende Aufgabe lösen:

In einer Ambulanz wurde durch eine Stichprobe ( $n = 160$ ) der Anteil der Patienten ermittelt, die länger als 30 Minuten auf den Beginn der Behandlung warten müssen. Dieser Anteil hat 30% betragen. Wie groß ist das Konfidenzintervall, wenn die statistische Sicherheit 95% betragen soll?

Gegeben sind somit

$N$ : sehr groß,  $n$ : 160,  $u$ : 1,96 ( $S = 3,00\%$ )

$p'$ : 0,30,  $q'$ : 0,70.

Wir berechnen nach Formel (101)

$$p = 0,30 \pm \frac{1,96 \cdot \sqrt{0,3 \cdot 0,7}}{\sqrt{160}} = 0,30 \pm 0,071$$

Als Konfidenzintervall können wir also angeben

$$p = 0,30 \pm 0,071 = 0,229 \dots 0,371$$

Der Anteil der Patienten, die länger als 30 Minuten auf den Beginn der Behandlung warten müssen, wird mit einer statistischen Sicherheit von 95% vom Intervall 0,229 ... 0,371 (oder: 22,9% ... 37,1%) eingeschlossen.

Zur Berechnung des erforderlichen Stichprobenumfangs beim Ermitteln von Anteilswerten großer Grundgesamtheiten ist Formel (100) nach  $n$  umzustellen

$$n = \frac{u^2 \cdot pq}{a_p^2} \quad (103)$$

Die statistische Sicherheit und der zulässige Stichprobenfehler werden vorgegeben. Die Anteilsgröße  $p$  (und dadurch gleichzeitig  $q$ ) in der Grundgesamtheit sind abzuschätzen.

Durch eine Stichprobe soll beispielsweise in einer Großmarkthalle bei der Anlieferung von Tafelobst festgestellt werden, wie groß der Anteil ( $p$ ) der Früchte ist, die Qualitätsmängel (Druckstellen, Fäulnis) aufweisen.

Wir arbeiten mit folgenden Ausgangsdaten:

$N$ : sehr groß,  $u$ : 1,96 ( $S = 9,00\%$ ),  $p$ : 0,1;  $q$ : 0,9;  $a_p$ : 0,02 (2,00%)

Die Rechnung ergibt

$$n = \frac{u^2 \cdot pq}{a_p^2} = \frac{1,96^2 \cdot 0,1 \cdot 0,9}{0,02^2} = 864 \approx 870$$

Es sind also ca. 870 Früchte in die Qualitätskontrolle einzubeziehen.

Mit diesem Beispiel wollen wir die Betrachtungen über das Schätzen statistischer Parameter beenden. Die hier behandelten Beispiele sollten einen kleinen Einblick in diese Problematik vermitteln, die insgesamt wesentlich vielseitiger und umfassender ist.

## 8 Hypothesen werden geprüft

Nachdem wir uns im Kapitel 7 mit Stichproben und ihrem Aussagevermögen für Grundgesamtheiten durch statistische Schätzverfahren beschäftigt haben, wollen wir jetzt einige Probleme des Prüfens von Hypothesen mit Hilfe von Stichprobenergebnissen kennenlernen, wobei auch hier wiederum nur ein kleiner Einblick in eine vielseitige und komplizierte Problematik gegeben werden kann.

### 8.1 Statistische Hypothesen und statistische Prüfverfahren

Als Hypothese wird in der wissenschaftlichen Arbeit eine noch unbewiesene Annahme bzw. Vermutung bezeichnet. Wir haben derartige Fälle bereits im Kapitel 6 bei der Erörterung der theoretischen Verteilungen kennengelernt.

Dort hatten wir zunächst nur angenommen, dass eine empirische Verteilung z. B. der Poisson-Verteilung oder der Normalverteilung entspricht. Es ist in diesem Zusammenhang auch bereits darauf aufmerksam gemacht worden, dass die Beweisführung dafür im Kapitel 8 erbracht wird.

Hypothesen spielen bei statistischen Untersuchungen eine große Rolle. Auf ihnen beruht die große Gruppe der Prüfverfahren. Statistische Prüfverfahren - auch statistische Tests genannt - werden dazu verwendet, mit Hilfe von Stichproben bestimmte Annahmen (Hypothesen) über die Grundgesamtheit, aus der die Stichprobe entnommen wurde, zu prüfen.

Sie dienen auch dazu, Kennwerte (z. B. Mittelwerte, Streuungen, Korrelationskoeffizienten) aus zwei oder mehreren Stichproben dahingehend zu prüfen, ob sie mit der Annahme vereinbar sind, aus ein und derselben Grundgesamtheit zu entstammen.

Die Prüfverfahren ermöglichen schließlich auch auf Grund von Stichproben, eine nicht bekannte Grundgesamtheit mit anderen bekannten bzw. hypothetisch angenommenen Grundgesamtheiten zu vergleichen. Die Grundlage hierfür ist der Inklusionsschluss, der Schluss von der Grundgesamtheit auf die Stichprobe, der bereits im Abschnitt 7.1. erläutert wurde.

Bei der statistischen Prüfung wird zunächst eine Hypothese formuliert. Sie wird allgemein Nullhypothese  $H_0$  genannt. Im Ergebnis der Auswertung von Stichproben wird diese Hypothese verworfen oder angenommen. Jeder Nullhypothese  $H_0$  können wir eine Alternativhypothese  $H_1$  gegenüberstellen. Sie ist die Negation der Nullhypothese.

Wir wollen zunächst ein Beispiel betrachten. Eine Stichprobe bei Werkstücken über die Abweichung des Durchmessers vom Nennmaß (40 mm) ergab als durchschnittliche Abweichung  $\bar{x} = 56,03$   $\mu\text{m}$ .

Insgesamt wurden 200 Messungen vorgenommen. Durch einen statistischen Test soll nun festgestellt werden, ob die aus der Stichprobe ermittelte durchschnittliche Abweichung darauf schließen lässt, dass ein vorgegebener höchstzulässiger Sollwert von 55  $\mu\text{m}$  eingehalten wurde. Die Irrtumswahrscheinlichkeit soll 0,27% betragen.

Dieser Aufgabe liegt folgende mathematisch-statistische Problemstellung zugrunde: Es

wird ein messbares Merkmal untersucht, so dass wir eine stetige Zufallsvariable  $X$  (Abweichung vom Sollwert) vorliegen haben, wobei die Grundgesamtheit die Tages- oder Monatsproduktion des betreffenden Werkstücks in einem Betrieb sein kann.

Die Verteilungsfunktion  $F(x)$  der Zufallsvariablen  $X$  in der Grundgesamtheit wird als normalverteilt mit den unbekannten Parametern  $\mu$  und  $\sigma^2$  angenommen.

Wir fragen nach der Einhaltung des Sollwertes von  $55\ \mu\text{m}$  und formulieren dies als Nullhypothese. Dafür schreiben wir bei zweiseitiger Fragestellung (der Begriff wird auf den folgenden Seiten erläutert)

$$H_0: \mu = 55\mu\text{m} \text{ oder allgemein } H_0: \mu = \mu_0 \quad (104)$$

$$\text{Die Alternativhypothese lautet dann } H_1: \mu = \mu_1, \text{ wobei } \mu_1 \neq \mu_0. \quad (105)$$

Derartige Hypothesen, bei denen ein Parameter der Grundgesamtheit geprüft wird, bezeichnen wir als Parameterhypothesen.

Sie beziehen sich also auf die einzelnen Parameter der Grundgesamtheit, wenn die Gestalt der Verteilungsfunktion bis auf einzelne Parameter gegeben ist. Zum Beispiel ist die eben dargestellte Hypothese über den Mittelwert  $\mu$  einer Grundgesamtheit  $H_0: \mu = \mu_0$  eine Parameterhypothese unter der Annahme, der Mittelwert der Grundgesamtheit  $\mu$  kann dem gegebenen Wert  $\mu_0$  entsprechen.

Außer den Parameterhypothesen gibt es die nichtparametrischen Hypothesen, bei denen die betreffende Verteilungsfunktion nicht gegeben ist. So sind z. B. Annahmen über das Verteilungsgesetz der Grundgesamtheit in der Form

$$H_0: F(x) = F_0(x) \quad (106)$$

nichtparametrische Hypothesen. Dabei enthält  $F(x)$  die auf ihre Verteilung zu prüfenden Werte, und  $F_0(x)$  ist eine angenommene Wahrscheinlichkeitsverteilung (z. B. Normalverteilung oder Poisson-Verteilung).

Die Nullhypothese geht also von einer Annahme über die unbekannten Parameter bzw. über das unbekannte Verteilungsgesetz der Grundgesamtheit aus. Derartige Annahmen können auf unterschiedliche Art und Weise gewonnen werden, etwa durch das Verwenden von Plan- oder Normzahlen (Sollwerte), wie das in unserem Beispiel der Fall war, sie können aber auch auf bereits vorliegenden Analysen vergangener Zeiträume aufbauen, wobei zu prüfen ist, ob die darin enthaltenen Werte noch Gültigkeit haben oder nicht, sie können schließlich auch auf bestimmten theoretischen Werten (Vermutungen) beruhen, die durch wahrscheinlichkeitstheoretische Ansätze formuliert werden.

Diese Hypothese wird nun anhand von Stichprobenergebnissen geprüft, d. h., die Werte einer Stichprobe oder durch Stichproben ermittelte Verteilungen werden z. B. mit den Parametern bzw. mit dem Verteilungsgesetz der Grundgesamtheit verglichen.

Es wird geprüft, ob die zwischen den Werten der Grundgesamtheit und der Stichprobe auftretenden Differenzen als zufällig angesehen werden können oder nicht. Sind die beobachteten Unterschiede so groß, dass der Test die Nullhypothese verwirft, so heißen sie signifikant.



Die zwischen Grundgesamtheit und Stichprobe auftretenden Unterschiede können dann nicht mehr auf den der Stichprobe anhaftenden Stichprobenfehler zurückgeführt werden. Wenn ein Test keine Signifikanz für die geprüften Unterschiede ergibt, dann besagt das lediglich, dass man keine statistisch gesicherten Wirkungen nachweisen kann. Ohne zusätzliche Hinweise dürfen wir deshalb aus einer nicht signifikanten Differenz keinesfalls bereits schlussfolgern, dass tatsächlich nur eine Zufallswirkung vorliegt. Es ist zu beachten, dass durch die statistischen Prüfverfahren nur Unterschiede nachgewiesen werden können, aber noch keine qualitative Wertung vorgenommen wird.

Im Kapitel 7 war festgestellt worden, dass die Ergebnisse von Stichproben vom tatsächlichen Wert der Grundgesamtheit zufallsbedingt abweichen. Den maximal zulässigen Wert dieser Abweichung gibt der Stichprobenfehler an. Mit seiner Hilfe kann z. B. für den Mittelwert ein Intervall angegeben werden, in dem die Stichprobenergebnisse aus einer Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit liegen werden.

Liegen die Stichprobenergebnisse innerhalb dieses Intervalls, kann die Abweichung als eine auf die Zufallsauswahl zurückzuführende zufällige Abweichung gelten. Liegt aber ein Stichprobenergebnis außerhalb des Vertrauensintervalls, so wird die Abweichung nicht mehr als zufällig angesehen; Stichprobe und Grundgesamtheit sind in diesem Fall nicht miteinander vereinbar.

Das Vertrauensintervall (Konfidenzintervall) für das Prüfen von Mittelwerten ist in Abbildung 33 dargestellt.

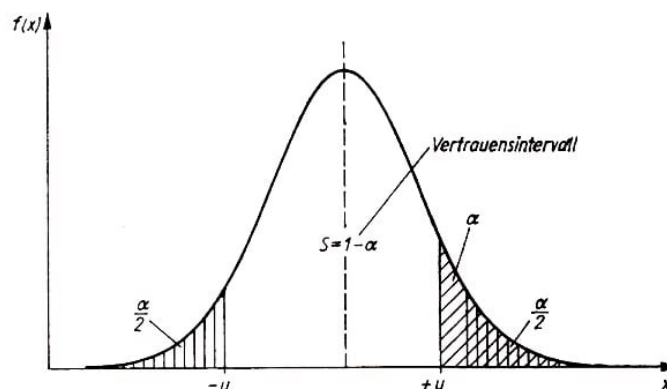


Abb. 33 Vertrauensintervall bei Normalverteilung

Wir hatten bereits bei der Berechnung des Stichprobenfehlers (Abschnitt 7.2.) erläutert, dass Mittelwerte bei genügend großer Stichprobe auch dann, wenn die Grundgesamtheit nicht normalverteilt ist, näherungsweise normalverteilt sind. Deshalb kann für die Ermittlung der Größe des Vertrauensintervalls beim Prüfen von Mittelwerten großer Stichproben die Normalverteilung zugrunde gelegt werden.

Zur Prüfung wird aus den Maßzahlen der Stichprobe und den angenommenen Parametern der Grundgesamtheit eine Testgröße berechnet. Diese Testgröße wird mit einem Tafelwert verglichen. Dabei ist eine statistische Sicherheit  $S$  bzw. Irrtumswahrscheinlichkeit  $\alpha$  festzulegen, um das Vertrauensintervall bestimmen und abgrenzen zu können. Während beim Stichprobenfehler hierfür in der Regel die statistische Sicherheit angegeben wird, ist es bei den Prüfverfahren üblich, die zulässige Irrtumswahrscheinlichkeit

zu nennen, die man auch als Signifikanzniveau bezeichnet.

Die Tafelwerte ergeben sich z. B. beim Prüfen des Mittelwertes durch große Stichproben aus der Verteilungsfunktion der normierten Normalverteilung.

Es wird nach einseitiger und zweiseitiger Fragestellung getrennt, je nachdem ob das Vertrauensintervall nur auf einer Seite oder auf beiden Seiten begrenzt ist. Die einseitige Begrenzung tritt z. B. auf, wenn die Untersuchung nur darauf hinausläuft, ob ein Mittelwert nicht zu groß ist, während kleine Mittelwerte unberücksichtigt bleiben.

Das kann z. B. vorkommen, wenn bei Werkstücken eine bestimmte Sollgrenze nicht überschritten werden darf, während eine Unterschreitung - natürlich in gewissen Grenzen - zulässig ist. Einige in der statistischen Arbeit häufig verwendete Größen für ein- und zweiseitige Fragestellung sind in Tabelle 27 vermerkt (vgl. hierzu auch die Tab. 23 und 24).

Symbol	Tafelwerte						
$\alpha$	0,004	0,0027	0,010	0,0455	0,050	0,100	0,3173
$S$	0,999	0,9973	0,990	0,9545	0,950	0,900	0,6827
$u$	3,29	3,00	2,58	2,00	1,96	1,65	1,00
$U'$	3,09	2,78	2,33	1,69	1,65	1,28	0,48

Tabelle 27 Tafelwerte für das Prüfen von Parameterhypothesen bei ein- und zweiseitiger Fragestellung und Normalverteilung

zur Tabelle 27: Es bedeuten  $\alpha$ : Irrtumswahrscheinlichkeit,  $S$ : statistische Sicherheit,  $u$ : Tafelwert bei zweiseitiger Fragestellung,  $u'$ : Tafelwert bei einseitiger Fragestellung

Beim Test werden folgende Fälle unterschieden:

Testgröße < Tafelwert: Annahme der Nullhypothese, die Differenz ist zufallsbedingt;

Testgröße > Tafelwert: Ablehnung der Nullhypothese, die Differenz ist signifikant.

Handelt es sich um nichtparametrische Prüfverfahren, können jedoch auch andere Beziehungen zwischen Testgröße und berechnetem Wert auftreten.

Die bisher erläuterten Prüfverfahren bezeichnet man auch als klassische Prüfverfahren. Hier stammen die Untersuchungsdaten aus einer Stichprobe von festgelegtem Umfang. Der Umfang der Stichprobe wird dabei bereits vor der Prüfung bestimmt.

Daneben gibt es noch die Gruppe der sequentiellen Prüfverfahren (Folgetestverfahren), die hier lediglich erwähnt sein sollen. Bei diesen Prüfverfahren wird der Stichprobenumfang nicht vor der Prüfung festgelegt, sondern es wird schrittweise vorgegangen. Nach jedem neu einbezogenen Element wird geprüft, ob schon eine Aussage über die Annahme oder Ablehnung der Hypothese möglich ist.

Der Stichprobenumfang hängt bei den sequentiellen Verfahren vom jeweiligen Beobachtungsergebnis ab und stellt in gewissem Sinn eine Zufallsvariable dar.

Nach diesen zum Teil recht theoretischen Ausführungen, die aber zum Verständnis der Prüfverfahren notwendig waren, sollen in folgendem der konkrete Ablauf sowie Ziel und Zweck der klassischen Prüfverfahren erläutert werden.

## 8.2 Das Prüfen von Parameterhypothesen

Wir hatten festgestellt, dass Parameterhypothesen Hypothesen sind, bei denen einzelne Parameter einer Grundgesamtheit geprüft werden, wenn das Verteilungsgesetz  $F(x)$  der Zufallsvariablen  $X$  bekannt ist.

Dabei hatten wir an einem Beispiel bereits die Nullhypothese für das Prüfen des Mittelwertes normalverteilter Grundgesamtheiten formuliert. Diesem Beispiel wollen wir uns jetzt wieder zuwenden.

Die durchschnittliche Abweichung vom Nennmaß (40 mm) war bei Werkstücken dahingehend zu prüfen, ob ein vorgegebener höchstzulässiger Sollwert von 55 um eingehalten wird oder nicht. Die Ausgangsdaten waren

$\mu_0$ : 55,00  $\mu\text{m}$ ;  $\sigma$ : 6,00  $\mu\text{m}$ ;  $\bar{x}$ : 56,03  $\mu\text{m}$   
 $s$ : 6,30  $\mu\text{m}$ ;  $n$ : 200 Stück;  $u$ : 3,00 ( $\alpha = 0,27\%$ ).

Für unser Beispiel hatten wir die Nullhypothese entsprechend Formel (104) bereits als

$$H_0 : \mu = \mu_0 = 55\mu\text{m}$$

formuliert. Für das Berechnen der Testgröße  $\hat{u}$  gilt beim Prüfen von Mittelwerten die Formel

$$\hat{u} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n} \quad (107)$$

Dabei bedeuten

$\hat{u}$ : Testgröße

$\bar{x}$ : Mittelwert der Stichprobe

$\mu_0$ : Mittelwert der Grundgesamtheit

$\sigma$ : Standardabweichung der Grundgesamtheit

$n$ : Stichprobenumfang.

Die angegebene Formel hat streng genommen nur bei normalverteilter Grundgesamtheit und bekannter Varianz  $\sigma^2$  Gültigkeit. Beide Voraussetzungen sind in unserem Beispiel erfüllt.

Die Formel kann jedoch bei großen Stichproben auch näherungsweise zur Prüfung des Mittelwertes verwendet werden, wenn diese Voraussetzungen nicht gegeben sind. Bei unbekannter Varianz  $\sigma^2$  und kleinen bzw. mittleren Stichproben ist der Vergleichswert der  $t$ -Verteilung zu entnehmen.

Die Testgröße kann für unser Beispiel berechnet werden als

$$\hat{u} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n} = \frac{56,03 - 55,00}{6,00} \sqrt{200} = 2,43$$

Bei zweiseitiger Fragestellung wird für

$$|\hat{u}| < u \quad (108)$$

die Hypothese angenommen, für

$$|\hat{u}| \geq u \quad (109)$$

wird sie abgelehnt.

In unserem Beispiel soll die Irrtumswahrscheinlichkeit 0,27% betragen. Dafür ergibt sich bei zweiseitiger Fragestellung laut Tabelle 27 eine Vergleichsgröße (Tafelwert)  $u = 3,00$ . Somit ist

$$|\hat{u}| < u, \quad \text{denn} \quad |2,43| < 3,00$$

Die Nullhypothese wird angenommen. Die festgestellte Abweichung des arithmetischen Mittels vom Sollwert kann als zufallsbedingt entstanden angesehen werden. Das bedeutet allerdings nicht, dass sei noch einmal ausdrücklich betont, dass damit die Richtigkeit der Hypothese bewiesen wäre, sondern lediglich, dass wir keinen Grund haben anzunehmen, dass die Hypothese falsch ist.

Interessiert bei der gegebenen Aufgabenstellung z. B. nur der Fall, dass ein bestimmter Durchschnittswert nicht überschritten wird, während das Unterschreiten für die Weiterverarbeitung keine Bedeutung hat, so liegt eine einseitige Fragestellung vor.

Jetzt lautet die Alternativhypothese  $H_1 : \mu > \mu_0$ , und die Nullhypothese müssten wir demzufolge formulieren als  $H_0 : \mu \leq \mu_0$  (bzw. im umgekehrten Fall:  $H_0 : \mu \geq \mu_0$ ,  $H_1 : \mu < \mu_0$ ). Der Nullhypothese widersprechen somit nur "zu große" Stichprobenmittelwerte.

Die Nullhypothese lautet in unserem Beispiel dann

$$H_0 : \mu \leq \mu_0 = 55,00 \mu\text{m} \quad \text{und} \quad H_1 : \mu > 55,00 \mu\text{m}$$

Mit den gegebenen Ausgangsdaten ergibt die Testgröße wiederum den Wert

$$\hat{u} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n} = \frac{56,03 - 55,00}{6,00} \sqrt{200} = 2,43$$

Der Tafelwert für die einseitige Fragestellung beträgt laut Tabelle 27 bei einer Irrtumswahrscheinlichkeit von 0,27%  $u' = 2,78$ .

Für unser Beispiel gilt somit  $|\hat{u}| < u'$ , weil  $2,43 < 2,78$ .

Die Nullhypothese kann angenommen werden. Allerdings ist der Grad der Annahme schwächer als bei der zweiseitigen Fragestellung. Man sagt dafür auch, die einseitige Fragestellung sei trennschärfer als die zweiseitige Fragestellung.

Für das Prüfen von Mittelwerten soll ein weiteres Beispiel die vielfältigen Anwendungsmöglichkeiten aufzeigen. Ein Produkt soll in Packungen zu 500 g in den Handel kommen. Eine Stichprobe von  $n = 20$  Packungen aus der Tagesproduktion ergibt ein durchschnittliches Gewicht von 505,3 g. Die Standardabweichung ergab  $s = 3,4$  g.

Mit einer Irrtumswahrscheinlichkeit von 1% ist zu prüfen, ob die durch die Stichprobe gefundenen Werte mit der Annahme übereinstimmen, dass der Durchschnittswert der insgesamt gefertigten Packungen 500 g beträgt.

Hier haben wir den Fall zu verzeichnen, dass die Standardabweichung der Grundgesamtheit nicht aus älteren Untersuchungen übernommen werden kann. Daher wird die Standardabweichung der Stichprobe zur Berechnung der Testgröße herangezogen. Weil eine kleine Stichprobe vorliegt und die Standardabweichung der Grundgesamtheit nicht

bekannt ist, müssen wir die Testgröße  $\hat{t}$  berechnen:

$$\hat{t} = \frac{\bar{x} - \mu_0}{s} \sqrt{n} \quad (110)$$

Die Hypothesen lauten

$$H_0 : \mu = \mu_0 = 500 \text{ g} \quad \text{und} \quad H_1 : \mu \neq 500 \text{ g}$$

Es sind gegeben  $\mu_0 : 500 \text{ g}$ ;  $n: 20$ ;  $\bar{x}: 505,3 \text{ g}$ ;  $\sigma \approx s: 3,48$ .

Nach (110) ergibt sich

$$\hat{t} = \frac{\bar{x} - \mu_0}{s} \sqrt{n} = \frac{505,3 - 500}{3,4} \sqrt{20} = 6,97$$

Bei  $n = 20$  und einer Irrtumswahrscheinlichkeit von 1% beträgt die Vergleichsgröße  $t$  für zweiseitige Fragestellung 2,86.

Tabellen der t-Verteilung sind in der Mehrzahl der statistischen Fachbücher enthalten.<sup>34</sup>

Somit ist  $\hat{t} > t$ , denn  $6,97 > 2,86$ .

Die Nullhypothese muss abgelehnt werden. Wir nehmen die Alternativhypothese an, da die festgestellte Abweichung signifikant ist. Das durchschnittliche Gewicht von 500 g wird nicht eingehalten. Es ist eine Veränderung an der Abfüllanlage vorzunehmen.

Wir haben an unseren Beispielen gesehen, dass die Entscheidung über die Annahme bzw. die Ablehnung einer statistischen Hypothese auf Grund einer Stichprobe getroffen werden muss.

Bei der Entscheidung über die Gültigkeit von Hypothesen lassen sich unterschiedliche Maßstäbe anlegen. Das hängt von der vorliegenden praktischen Aufgabenstellung ab. Durch das Festlegen der Irrtumswahrscheinlichkeit bzw. des Signifikanzniveaus  $\alpha$  wird dieser Maßstab bestimmt. Sind wir z. B. bereit, das Risiko zu übernehmen, durchschnittlich in fünf von hundert Fällen ein Fehlurteil zu treffen, dann entscheiden wir uns für die Irrtumswahrscheinlichkeit  $\alpha = 0,05 = 5\%$ .

Hier genügen bereits relativ kleine Differenzen, um die Nullhypothese zurückzuweisen. Stellen wir höhere Ansprüche an die Zuverlässigkeit unserer Entscheidung, so müssen wir z. B. mit einer Irrtumswahrscheinlichkeit von  $\alpha = 0,01 = 1\%$  arbeiten.

Um eine Hypothese dann zurückzuweisen, bedarf es bereits größerer Kennwertunterschiede.

Arbeitet man mit einer sehr kleinen Irrtumswahrscheinlichkeit (z. B.  $\alpha = 0,001 = 0,1\%$ ), bei der der Untersuchende eine Fehlentscheidung mit möglichst großer Sicherheit vermeiden will, dann kann die Nullhypothese nur bei sehr großen Differenzen zwischen den Maßzahlen abgelehnt werden.

Beim Prüfen von Hypothesen können zwei Arten von Fehlern auftreten:

- eine richtige Hypothese wird abgelehnt (Fehler erster Art)
- eine falsche Hypothese wird angenommen (Fehler zweiter Art).

---

<sup>34</sup>Vgl. z. B. ebenda, S. 330

Wählen wir z. B. eine Irrtumswahrscheinlichkeit von  $\alpha = 0,05$ , können wir evtl. eine Hypothese zurückweisen (signifikante Unterschiede annehmen), obwohl sie richtig war und die wir bei einer Irrtumswahrscheinlichkeit von  $\alpha = 0,001$  angenommen hätten. Will man einen Fehler erster Art vermeiden, ist man deshalb bestrebt, die Irrtumswahrscheinlichkeit möglichst klein zu halten. Jedoch ist dann das Risiko größer, einen Fehler zweiter Art zu begehen, man vermutet keine wesentlichen Unterschiede, die aber in Wirklichkeit in der Grundgesamtheit vorhanden sind. Genauere Aussagen über den Fehler zweiter Art erhält man durch das Gegenüberstellen von Nullhypothese und Alternativhypothese.

In jedem speziellen Fall müssen wir entscheiden, ob lieber ein Fehler erster oder lieber ein Fehler zweiter Art in Kauf genommen werden soll. Ist es folgenswer, eine Nullhypothese irrtümlicherweise abzulehnen, werden wir dieses Risiko möglichst klein halten, also ein hohes Signifikanzniveau wählen.

Ist es dagegen folgenswer, eine Nullhypothese fälschlicherweise anzunehmen, ist mit einer geringeren Verlässlichkeit zu prüfen.<sup>35</sup>

Als weiteres Beispiel eines parametrischen Prüfverfahrens wollen wir jetzt das Prüfen von Anteilswerten kennenlernen. Die Nullhypothese lautet in diesem Fall

$$H_0 : p = p_0 \quad (111)$$

Die Testgröße kann wie folgt berechnet werden:

$$\hat{u} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} \quad (112)$$

Es bedeuten

$\hat{u}$ : Testgröße

$x$ : absolute Häufigkeit, mit der ein Ergebnis in der Stichprobe eintrat

$n$ : Stichprobenumfang

$p_0$ : Wahrscheinlichkeit des Ergebnisses in der Grundgesamtheit.

Die Rechnung basiert auf der Binomialverteilung. Sinngemäß gelten aber die gleichen Regeln, die bereits bei der Prüfung des Mittelwertes erläutert worden sind.

Wir wollen folgendes Problem untersuchen:

Aus einer früheren Ermittlung im innerstädtischen Verkehr einer Großstadt sei der Anteil der Monatskarteninhaber an den Fahrgästen insgesamt bekannt. Er betrage  $p_0 = 0,2$ . Durch eine Stichprobe soll geprüft werden, ob gegenwärtig dieser Anteilswert noch Gültigkeit hat.

Die Stichprobe liefert  $n = 900$  und  $x = 230$ . Wir formulieren mit  $p_0 = 0,2$  bei zweiseitiger Fragestellung  $H_0 : p = 0,2$  ( $H_1 : p \neq 0,2$ ).

---

<sup>35</sup>Vgl. u.a. Clauß, G., Ebner, H.: Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Volk und Wissen - Volkseigener Verlag, Berlin 1970, S. 1708.

Das Berechnen der Testgröße ergibt

$$\hat{u} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{230 - 900 \cdot 0,2}{\sqrt{900 \cdot 0,2(1 - 0,2)}} = 4,17$$

Bei einer Irrtumswahrscheinlichkeit von 5% ( $\alpha = 0,05$ ) und zweiseitiger Fragestellung erhält man als Vergleichsgröße  $u = 1,96$ .

Da  $|\hat{u}| > u$ , denn  $4,17 > 1,96$ , wird die Nullhypothese abgelehnt. Der Anteil der Monatskarteninhaber kann nicht mehr mit 20% angegeben werden.

Von dieser Erkenntnis ausgehend, müssen weitere Untersuchungen durchgeführt werden, um einen Überblick über den tatsächlichen Anteil der Monatskarteninhaber zu erhalten. So könnte mit den angegebenen Größen nach den im Abschnitt 7.3. erläuterten Regeln eine Schätzung über den Anteil der Monatskarteninhaber vorgenommen werden.

Statistische Prüfverfahren dienen auch dazu, die Ergebnisse zweier Stichproben zu testen. Hierfür ein Beispiel: Untersuchungen über die Vitalkapazität (Luftfassungsvermögen der Lunge) von Oberschülern lieferten folgende Werte:<sup>36</sup>

- Für 113 vierzehnjährige Oberschüler, die lediglich am Schulsport teilnahmen, ergaben sich für die Vitalkapazität der Mittelwert  $\bar{x}_1 = 3480$  ml und die Standardabweichung  $s_1 = 712$  ml.

- Für 172 gleichaltrige Oberschüler, die einer Sportgemeinschaft angehörten, ergaben sich für die Vitalkapazität die Werte  $\bar{x}_2 = 3650$  ml und  $s_2 = 766$  ml.

Es wurde mit einer Irrtumswahrscheinlichkeit von  $\alpha = 0,05$  geprüft, ob die Annahme gerechtfertigt ist, dass die Vitalkapazität (für Oberschüler dieser Altersstufe) durch zusätzliche sportliche Betätigung erhöht wird. Da die Betätigung in der Sportgemeinschaft sicher nicht zu einer Verringerung der Vitalkapazität führen wird, ist der Test als einseitige Fragestellung formuliert worden, und zwar als

$$H_0 : \mu_1 = \mu_2 \quad (H_1 : \mu_1 < \mu_2) \quad (113)$$

Es liegen zwei große, voneinander unabhängige Stichproben vor, und die Zufallsvariable wird in beiden Grundgesamtheiten mit derselben Standardabweichung  $\sigma$  als normalverteilt angenommen. Die Testgröße kann nach folgender Formel berechnet werden:

$$\hat{u} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \quad (114)$$

Der Parameter  $\sigma$  wird näherungsweise durch  $s$  ersetzt, das nach Formel (115) berechnet werden kann.

$$\sigma \approx s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (115)$$

<sup>36</sup>Vgl. ebenda, S. 186f.

Wir berechnen zunächst

$$s = \sqrt{\frac{112 \cdot 712^2 + 171 \cdot 766^2}{113 + 172 - 2}} = 744,7$$

und bestimmen dann nach Formel (114) die Testgröße

$$|\hat{u}| = \frac{|3480 - 3650|}{744,7} \sqrt{\frac{113 \cdot 172}{113 + 172}} = 1,89$$

Der Tafelwert beträgt bei einseitiger Fragestellung laut Tabelle 27 bei einer Irrtumswahrscheinlichkeit von  $\alpha = 0,05$ :  $u' = 1,65$ .

Für das Beispiel gilt deshalb  $|\hat{u}| > u'$ , denn  $1,89 > 1,65$ .

Die Nullhypothese wird also zurückgewiesen und die Alternativhypothese angenommen. Die Vitalkapazität ist bei vierzehnjährigen Oberschülern, die Mitglied einer Sportgemeinschaft sind, im Mittel größer als bei Schülern, die nur am Schulsport teilnehmen.

### 8.3 Wie prüfen wir Verteilungen?

Nach dem Prüfen von Parameterhypothesen soll jetzt an zwei Beispielen das Prüfen nichtparametrischer Hypothesen erläutert werden.

Statistische Erhebungen liefern zunächst empirische Häufigkeitsverteilungen (vgl. Kap. 4). Im allgemeinen wird man bestrebt sein, diese Häufigkeitsverteilungen dann durch mathematische Funktionen zu beschreiben. Dabei wird untersucht, ob den empirischen Verteilungen mathematische Gesetzmäßigkeiten zugrunde liegen, die zu der Annahme berechtigen, die betreffende Verteilung entspreche zumindest näherungsweise einer theoretischen Verteilung.

Die prinzipiellen Bemerkungen über Inhalt und Zweck der theoretischen Verteilungen sind bereits im Kapitel 6 enthalten. Jetzt soll erläutert werden, wie geprüft werden kann, ob eine empirische Verteilung einer theoretischen Verteilung genügt.

Das Prüfverfahren hat also die Aufgabe, den Grad der Übereinstimmung zwischen einer empirischen und einer theoretischen Verteilung festzustellen.

Liegen die Abweichungen zwischen beiden Verteilungen in einem bestimmten Zufälligkeitsbereich, kann die empirische Verteilung durch die Funktion der theoretischen Verteilung beschrieben werden.

Durch die Nullhypothese wird beim Prüfen von Verteilungen behauptet, dass die Abweichungen zwischen der theoretischen und der empirischen Häufigkeitsverteilung zufällig sind und zwischen beiden Verteilungen keine wesentlichen (signifikanten) Unterschiede bestehen.

Die beim Prüfen von Verteilungen am häufigsten angewendete Methode ist der  $\chi^2$ -Test (Chi-Quadrat-Test). Hierbei wird zur Prüfung die Testgröße

$$\hat{\chi}^2 = \sum_{m=1}^k \frac{(h_m - np_m)^2}{np_m} \quad (116)$$



berechnet, und es bedeuten

$\hat{\chi}^2$ : Testgröße

$m$ :  $m$ -te Klasse

$k$ : Zahl der Klassen insgesamt

$h_m$ : beobachtete Häufigkeit der  $m$ -ten Klasse

$n$ : Umfang der Stichprobe

$p_m$ : Wahrscheinlichkeit der  $m$ -ten Klasse.

Die Testgröße wird der Vergleichsgröße  $\chi^2$  gegenübergestellt. Dabei gilt:

$\hat{\chi}^2 < \chi^2$ : Die Hypothese wird angenommen, die Abweichungen sind zufällig.

$\hat{\chi}^2 \geq \chi^2$ : Die Hypothese wird verworfen, die Abweichungen sind signifikant.

Der  $\chi^2$ -Test beruht auf den Eigenschaften der  $\chi^2$ -Verteilung.

Die  $\chi^2$ -Verteilung ist eine sogenannte Prüfverteilung. Zur Prüfung werden die Differenzen zwischen der empirisch gefundenen Verteilung und der zur Prüfung hypothetisch gesetzten theoretischen Verteilung verwendet. Sie werden normiert, indem ihre Quadrate durch die zu erwartenden Werte dividiert werden. Die Summe dieser normierten Werte ergibt den  $\chi^2$ -Wert.<sup>37</sup>

Wenn die Testgröße näherungsweise einer  $\chi^2$ -Verteilung entsprechen soll, dürfen die theoretischen Häufigkeiten  $np_m$  - das sind die sich aus den theoretischen Verteilungen ergebenden Häufigkeiten - nicht zu klein sein. Es gilt deshalb die Regel

$$np_m \geq 5 \quad (117)$$

Sie besagt, dass kein Wert der theoretischen Häufigkeiten kleiner als 5 sein darf. Ist für einige Klassen diese Forderung nicht erfüllt, müssen benachbarte Klassen zusammengefasst werden, bis die Bedingung eingehalten wird.

Wir müssen dann auch die Größen  $h_m$  der entsprechenden Klassen addieren. Durch eine derartige Zusammenfassung verringert sich die Zahl der Klassen von  $k$  auf  $k^*$ .

Die Vergleichsgröße  $\chi^2$  wird für eine vorzugebende statistische Sicherheit bzw. Irrtumswahrscheinlichkeit und eine bestimmte Anzahl der Freiheitsgrade

$$FG = k^* - 1 - r \quad (118)$$

aus der Tafel der  $\chi^2$ -Verteilung entnommen (siehe Tab. 28).

---

<sup>37</sup>Vgl. auch Weber, E.: Grundriss der biologischen Statistik. VEB Gustav Fischer Verlag, Jena 1967, S. 484ff.

Freiheitsgrade $FG$	Irrtumswahrscheinlichkeit		Freiheitsgrade $FG$	Irrtumswahrscheinlichkeit	
	$\alpha = 0,05$	$\alpha = 0,1$		$\alpha = 0,05$	$\alpha = 0,1$
1	3,84	6,63	2	5,99	9,21
3	7,81	11,3	4	9,49	13,3
5	11,1	15,1	6	12,6	16,8
7	14,1	18,5	8	15,5	20,1
9	16,9	21,7	10	18,3	23,2
11	19,7	24,7	12	21,0	26,2
13	22,4	27,7	14	23,7	29,1
15	25,0	30,6	16	26,3	32,0
17	27,6	33,4	18	28,9	34,8
19	30,1	36,2	20	31,4	37,6
21	32,7	38,9	22	33,9	40,3
23	35,2	41,6	24	36,4	43,0
25	37,7	44,3	30	43,8	50,9
40	55,8	63,7	50	67,5	76,2

Dabei bedeuten

$FG$ : Anzahl der Freiheitsgrade (frei verfügbare Beobachtungswerte)

$k^*$ : Anzahl der zusammengefassten Klassen, für die die Bedingung  $np_m \geq 5$  erfüllt sein muss

$r$ : Anzahl der aus der Stichprobe geschätzten Parameter der Wahrscheinlichkeitsverteilung.

Im Abschnitt 6.3. war an einem Beispiel die Poisson-Verteilung erläutert worden. Dabei wurde an  $n = 220$  Tagen die Anzahl der täglich ausgefallenen Maschinen registriert (vgl. Tab. 21).

Pro Tag fielen durchschnittlich 1,5 Maschinen aus. Damit war  $\lambda = 1,5$ . Die durch die Erfassung ermittelten Werte sind in den Spalten 1 und 2 der Tabelle 29 noch einmal dargestellt. Die Spalten 3 und 4 enthalten die Wahrscheinlichkeiten  $p_m$  und die theoretischen Häufigkeiten  $np_m$ . Es ist zu prüfen, ob die beobachteten Häufigkeiten tatsächlich einer Poisson-Verteilung entsprechen.

Die Nullhypothese lautet beim Prüfen auf Poisson-Verteilung

$$H_0 : F(x) = \sum_{k < x} \frac{\lambda_0^k}{k!} e^{-\lambda_0} \quad (119)$$

Die Berechnungen sind den Spalten 4 bis 7 der Tabelle 29 zu entnehmen. In Spalte 4 sind die in Spalte 3 enthaltenen Wahrscheinlichkeiten der Poisson-Verteilung für  $\lambda = 1,5$  mit dem Stichprobenumfang  $n = 220$  multipliziert worden. In Spalte 5 wurden die Absolutwerte der Differenzen zwischen Spalte 4 und Spalte 2 gebildet und in Spalte 6 ihr Quadrat, um in Spalte 7 nach Formel (116) die normierten Werte zur Ermittlung der Testgröße  $\hat{\chi}^2$  bilden zu können.

Um die Bedingung  $np_m \geq 5$  zu erfüllen, mussten die letzten 5 Klassen zu einer Klasse zusammengefasst werden.

Für die Testgröße ergab sich der Wert  $\hat{\chi}^2 = 0,77$ .

Anzahl der ausgefallenen Maschinen	Beobachtete Häufigkeiten	Wahrscheinlichkeiten	Theoretische Häufigkeiten			
$m$	$h_m$	$p_m$	$np_m$	$ h_m - np_m $	$(h_m - np_m)^2$	$\frac{(h_m - np_m)^2}{np_m}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0	47	0,223	49,06	2,06	4,24	0,08
1	78	0,335	73,70	4,30	18,49	0,25
2	57	0,251	55,22	1,78	3,17	0,05
3	25	0,126	27,72	2,72	7,40	0,27
4	7	0,047	10,34			
5	3	0,014	3,08			
6	2	0,003	0,66			
7	0	0,001	0,22			
8	1	0,000	0,00			
$\sum_{i=4}^8$	13		14,30	1,30	1,69	0,12
$\sum$	$220 = n$	1,000	220,00			$0,77 = \hat{\chi}^2$

Tabelle 29 Berechnung von  $\hat{\chi}^2$  für das Beispiel Prüfen auf Poisson-Verteilung

Die Anzahl der Freiheitsgrade bestimmen wir mit  $k^* = 5$  und  $r = 1$  (ein Schätzwert:  $\lambda_0 = \bar{x} = 1,5$ ). Somit ist die Anzahl der Freiheitsgrade

$$FG = k^* - 1 - r = 5 - 1 - 1 = 3$$

Die Vergleichsgröße  $\chi^2$  hat für 3 Freiheitsgrade bei einer Irrtumswahrscheinlichkeit von 5% laut Tabelle 28 den Wert 7,81.

Es ist also  $\hat{\chi}^2 < \chi^2$ , denn  $0,77 < 7,81$ .

Die Nullhypothese wird angenommen. Die vorliegende empirische Verteilung widerspricht nicht der Annahme, dass die Werte aus einer poissonverteilten Grundgesamtheit mit dem Parameter  $\lambda = \mu = 1,5$  stammen.

Im Abschnitt 6.4. war bei der Erörterung der Normalverteilung als Beispiel ein Problem des Arbeitsstudiums aufgeführt worden.

Durch 200 Stichproben wurde die Dauer einer Arbeitsverrichtung gemessen. Die dabei ermittelten Werte und die Berechnung der theoretischen Häufigkeiten enthielt Tabelle 25.

Für die Größen  $\mu_0$  und  $\sigma_0$  ergab sich näherungsweise

$$\mu_0 = \bar{x} = 36,98 \text{ s} \quad , \quad \sigma_0 = s = 1,58 \text{ s}$$

Wir wollen jetzt prüfen, ob die beobachteten Häufigkeiten tatsächlich einer Normalverteilung entsprechen. Zu diesem Zweck sind in Tabelle 30, Spalten 1 bis 4, die hierfür erforderlichen Werte aus Tabelle 25 noch einmal aufgeführt.

Die Berechnung der Testgröße  $\hat{\chi}^2$  wird in den Spalten 5 bis 7 der Tabelle 30 vorgenommen.

Die Nullhypothese lautet beim Prüfen auf Normalverteilung

$$H_0 : F(x) = \Phi_0(x : \mu_0; \sigma_0^2) \quad (120)$$

$\Phi_0$  ist die Verteilungsfunktion der Normalverteilung mit den Parametern  $\mu_0$  und  $\sigma_0^2$ . Um die Bedingung  $np_m \geq 5$  zu erfüllen, mussten die ersten 2 und die letzten 4 Klassen zu jeweils einer Klasse zusammengefasst werden.

Für die Testgröße ergab sich  $\hat{\chi}^2 = 2,81$ .

Die Anzahl der Freiheitsgrade ist zu bestimmen mit  $k^* = 8$  und  $r = 2$  (Schätzwerte:  $\mu_0 = \bar{x} = 36,98 \text{ s}$ ,  $\sigma_0 = s = 1,58 \text{ s}$ ). Daher ist die Zahl der Freiheitsgrade

$$FG = k^* - 1 - r = 8 - 1 - 2 = 5$$

Der Tafelwert  $\chi^2$  hat für 5 Freiheitsgrade und eine Irrtumswahrscheinlichkeit von 5% den Wert 11,1.

$$\hat{\chi}^2 < \chi^2, \quad \text{denn} \quad 2,81 < 11,1$$

Die Nullhypothese wird angenommen. Die vorliegende empirische Verteilung kann zumindest näherungsweise als normalverteilt angesehen werden.

Neben dem  $\chi^2$ -Test gibt es noch eine Reihe anderer parameterfreier Prüfverfahren,

Klasse	Dauer (s)	Beobachtete Häufigkeiten	Theoretische Häufigkeiten				
$m$	$e_{mu}e_{mo}$	$h_m$	$np_m$	$ h_m - np_m $	$(h_m - np_m)^2$	$\frac{(h_m - np_m)^2}{np_m}$	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	
1	32...33	1	1,18				
2	33...34	2	4,70				
$\sum_{i=1}^2$		3	5,88	2,88	8,29	1,41	
3	34...35	17	15,26	1,74	3,03	0,20	
4	35...36	34,5	32,32	2,18	4,75	0,15	
5	36...37	45,5	47,34	1,84	3,39	0,07	
6	37...38	52	47,64	4,36	19,01	0,40	
7	38...39	29	30,10	1,10	1,21	0,04	
8	39...40	13	15,84	2,84	8,07	0,51	
9	40...41	4	4,52				
10	41...42	1	0,96				
11	42...43	1	0,12				
12	43...44	0	0,02				
$\sum_{i=9}^{12}$		6	5,62	0,38	0,14	0,03	
$\Sigma$		200,0 = $n$	200,00			$2,18 = \chi^2$	

Tabelle 30 Berechnung von  $\chi^2$  für das Beispiel Prüfen auf Normalverteilung (Normalverteilung)

die für zahlreiche statistische Aufgabenstellungen entwickelt worden sind, insbesondere wenn die Verteilung der Grundgesamtheit unbekannt ist.

Auf diese soll hier jedoch nicht näher eingegangen werden, da sie schon recht spezielle Kenntnisse der mathematischen Statistik voraussetzen.

## 9 Zeitreihen

Alle Erscheinungen in Natur, Gesellschaft und Technik sind durch sachliche, örtliche und zeitliche Merkmale bestimmt. Die bisher behandelten statistischen Untersuchungsmethoden waren geeignet, eine Analyse der Beobachtungswerte zu ausgewählten Zeitpunkten vorzunehmen. Unbeachtet blieb dabei die der Mehrzahl der Erscheinungen innewohnende Dynamik ihrer Entwicklung.

Der Zeitreihenforschung ist die Aufgabe gestellt, die Art und Weise der zeitlichen Variabilität der einzelnen Erscheinungen zu untersuchen. Diese Analyse der zeitlichen Entwicklung ist eine wesentliche Grundlage für die wissenschaftliche Planung der sozialistischen Volkswirtschaft und ihrer einzelnen Zweige.

Ein bewusster Einfluss auf die planmäßige proportionale Entwicklung der Volkswirtschaft im nationalen Maßstab und unter dem Aspekt der Zusammenarbeit der RGW-Länder mit dem Ziel einer höchstmöglichen Befriedigung der Bedürfnisse aller Werktätigen kann nur ausgeübt werden, wenn auch die zurückliegenden Entwicklungstendenzen und deren Ursachen bekannt sind.

Gleichermaßen sind diese Untersuchungen auch unabdingbare Voraussetzungen der Leitung und Planung in den Zweigen und Betrieben der sozialistischen Volkswirtschaft. So muss die Planung der Selbstkosten im Betrieb auf einer gründlichen Analyse der bisherigen Entwicklung der Selbstkosten aufbauen.

Die Planung des Aufkommens an tierischen Produkten muss von der bisherigen Entwicklung der Viehbestände und des Aufkommens an tierischen Produkten ausgehen.

Die Zeitreihenforschung hat gleichfalls Bedeutung für die Technik, etwa bei der Analyse von Entwicklungstendenzen technisch-technologischer Neuerungen. In der Biologie interessieren in starkem Maße Wachstumsprozesse, die an Lebewesen und Pflanzen erkannt werden können. Aber auch auf anderen Gebieten ergeben sich vielfache Anwendungsmöglichkeiten für die Zeitreihenforschung, so dass die Analyse von Zeitreihen heute ein fester Bestandteil des statistischen Instrumentariums geworden ist.

Wenn die Beobachtungswerte verschiedener Zeitpunkte oder Zeitintervalle entsprechend der zeitlichen Reihenfolge geordnet werden, sprechen wir von einer Zeitreihe.

Eine Zeitpunktreihe ergibt sich, wenn der Bestand an ausgewählten Stichtagen erfasst wird (Volkszählung, Gebäudezählung).

Es handelt sich dabei meist um Erscheinungen, die entweder nur geringen Veränderungen über längere Zeitabschnitte unterworfen sind oder die einer permanenten Erfassung nicht oder nur schwer zugänglich sind.

Beobachten wir dagegen die in einem laufenden Prozess vor sich gehende ständige Veränderung, so erhalten wir eine Zeitintervallreihe, in der die Größe der Erscheinung summarisch für das einzelne Zeitintervall erfasst ist.

Die in der volkswirtschaftlichen Planabrechnung typischen Zeitintervalle sind der Monat und das Jahr, während sich auf betrieblicher Ebene zum Teil die Notwendigkeit ergibt, das zu beobachtende Zeitintervall bis auf den Tag oder bestimmte Stunden des Tages zu reduzieren. Ein typisches Beispiel ist hier die tägliche Kassenabrechnung.

Für technisch-technologische, aber auch naturwissenschaftliche Untersuchungen kann sich eine Verkürzung der Intervalle bis zur Minute oder gar Sekunde erforderlich machen. Im folgenden sollen die einzelnen Beobachtungswerte in ihrer zeitlichen Reihenfolge mit

$$y_1, y_2, \dots, y_k$$

bezeichnet werden, während die Zeit durch die Variable  $t$  angegeben wird. Eine Zeitreihe setzt sich aus den drei Komponenten:

Trend ( $T$ )  
periodische Schwankung ( $S$ ) und  
unregelmäßige Schwankung ( $E$ )

zusammen, wobei nicht in jeder Zeitreihe alle Komponenten auftreten. Wird eine Zeitreihe in diese Komponenten zerlegt, so bezeichnet man diesen Untersuchungsprozess als Zeitreihenanalyse.

Der Trend, der die wesentlichen Merkmale einer zeitlichen Veränderung statistischer Massen ausdrückt, charakterisiert den Grundverlauf einer Zeitreihe.

Die zeitliche Variabilität kommt aber nicht nur im Trend, sondern auch in regelmäßig wiederkehrenden sogenannten periodischen Schwankungen während verschiedener Tagesstunden, Wochentage, Dekaden, Monate oder Quartale zum Ausdruck.

Derartigen Schwankungen unterschiedlicher Intensität sind z. B. die Produktion in den verschiedenen Zweigen der Volkswirtschaft, das Verkehrsaufkommen im Güter- und Personenverkehr, der Krankenstand oder die Urlaubsabwicklung unterworfen, ohne dass hier näher auf die verschiedenen Ursachen dieser Schwankungen eingegangen werden soll.

Die unregelmäßigen Schwankungen sind letztlich Ausdruck besonderer Umstände, die eine einmalige oder mehrmalige unregelmäßige extreme Abweichung der Erscheinung vom Normalfall zur Folge haben. Diese unregelmäßigen Schwankungen können mit Hilfe statistischer Methoden nicht näher beschrieben werden, die qualitative Ursache-Wirkung-Forschung muss hierbei im Vordergrund stehen.

Aufgabe der Zeitreihenanalyse ist es nun, einen quantitativen Ausdruck insbesondere für den Trend und die periodischen Schwankungen zu finden und auf Grund der Ergebnisse die Ursachenforschung voranzutreiben.

## 9.1 Die Grundrichtung der Entwicklung wird bestimmt

Der Grundverlauf einer zeitlichen statistischen Reihe kann durch verschiedene Verfahren, je nach dem Ziel der Untersuchung, bestimmt werden. Eine erste Einschätzung dieses Grundverlaufs wird durch die absolute Gesamtveränderung oder die durchschnittliche absolute Veränderung vorgenommen.

Der Ansatz für die absolute Gesamtveränderung, die die Differenz zwischen Beobachtungswert des letzten und des ersten Beobachtungszeitabschnittes angibt, lautet

$$G = y_k - y_1 \quad (121)$$



Je nach dem Vorzeichen von  $G$  wird durch Formel (121) die absolute Gesamtzunahme oder -abnahme zwischen dem ersten und dem letzten Wert einer Zeitreihe zum Ausdruck gebracht.

Interessiert jedoch nicht die absolute Gesamtveränderung, sondern die durchschnittliche Zu- oder Abnahme der untersuchten Erscheinung je Beobachtungsabschnitt, so ist die durchschnittliche absolute Veränderung zu berechnen als

$$d = \frac{y_k - y_1}{k - 1} \quad (122)$$

wobei  $k$  die Anzahl der Zeitpunkte oder -intervalle angibt.

Auch in diesem Fall bringt das negative Vorzeichen eine Abnahme und das positive Vorzeichen eine Zunahme der Erscheinung, jetzt aber im Durchschnitt je Zeitabschnitt der Reihe, zum Ausdruck.

Beide Ansätze liefern allerdings nur dann brauchbare Ergebnisse, wenn der Zeitreihe eine relativ gleichgerichtete Entwicklung zugrunde liegt. Die Beziehungen (121) und (122) sind ungeeignet, wenn sich an eine steigende Entwicklung eine fallende anschließt und umgekehrt.

Insbesondere für vergleichende Betrachtungen der Veränderung mehrerer Erscheinungen während des gesamten Untersuchungszeitraumes ist die absolute Gesamtveränderung wenig geeignet. In diesen Fällen werden wir vielmehr die relative Gesamtveränderung

$$I_{1/k} = \frac{y_k}{y_1} \cdot 100(\%) \quad (123)$$

oder die durchschnittliche relative Veränderung bestimmen.

Ausgangspunkt für die Ermittlung der durchschnittlichen relativen Veränderung, auch durchschnittliches Wachstumstempo genannt, ist die Gesamtveränderung einer Erscheinung innerhalb des betrachteten Zeitabschnittes.

Gehen wir davon aus, dass zum Zeitpunkt  $t_1$  die Größe  $y_1$  gemessen wurde, so verändert sich die Größe  $y_1$  mit fortschreitendem  $t$  um den Faktor  $u$ , der zunächst zwischen benachbarten Zeitabschnitten unterschiedliche Größen annimmt. Allgemein können wir das darstellen als

$$\begin{aligned} y_1 &= y_1 \\ y_2 &= y_1 \cdot u_1 \\ y_3 &= y_1 \cdot u_1 \cdot u_2 \\ y_4 &= y_1 \cdot u_1 \cdot u_2 \cdot u_3 \quad \text{usw.} \end{aligned}$$

Werden die Größen  $u_i$  zu einer gemeinsamen durchschnittlichen Größe  $u$  umgeformt, so erhalten wir z. B.

$$y_4 = y_1 \cdot u \cdot u \cdot u = y_1 \cdot u^3$$

Formen wir diese letzte Beziehung um, so erhalten wir

$$u^3 = \frac{y_4}{y_1}, \quad u = \sqrt[3]{\frac{y_4}{y_1}}$$

bzw. allgemein

$$u = \sqrt[k-1]{\frac{y_k}{y_1}} \cdot 100(\%) \quad (124)$$

Für die Beziehungen (123) und (124) gilt allgemein

$$u > 100 \text{ Zunahme, } u = 100 \text{ Stagnation, } u < 100 \text{ Abnahme}$$

Am Beispiel der Entwicklung der Elektroenergieerzeugung in der DDR von 1955 bis 1971<sup>38</sup> (Tab. 31) soll die Berechnung der Maßzahlen der absoluten und relativen Veränderung demonstriert werden.

Jahr	Elektroenergieerzeugung in 10 <sup>3</sup> GWh
1955	28,7
1960	40,3
1965	53,6
1969	65,5
1970	67,7
1971	69,4

Tabelle 31 Entwicklung der Elektroenergieerzeugung in der DDR von 1955 bis 1971

Für die Entwicklung der Elektroenergieerzeugung wird nach Formel (121) die absolute Gesamtveränderung bestimmt als

$$G = (69,4 - 28,7) \cdot 10^3 \text{ GWh} = +40,7 \cdot 10^3 \text{ GWh}$$

Die durchschnittliche absolute Veränderung nach Formel (122) ergibt

$$d = \frac{69,4 - 28,7}{16} \cdot 10^3 \text{ GWh} = +2,54 \cdot 10^3 \text{ GWh}$$

Die absolute Gesamtzunahme der Elektroenergieerzeugung betrug danach zwischen 1955 und 1971 insgesamt  $40,7 \cdot 10^3$  GWh, was einer durchschnittlichen Zunahme pro Jahr von  $2,54 \cdot 10^3$  GWh entspricht.

Die relative Gesamtveränderung, berechnet nach Formel (123), beträgt

$$I_{1955/1971} = \frac{69,4}{28,7} \cdot 100 = 241,5\%$$

und die durchschnittliche relative Veränderung

$$u = \sqrt[17-1]{\frac{69,4}{28,7}} \cdot 100 = \sqrt[16]{2,42} \cdot 100 = 0,10568 \cdot 100 = 10,6\%$$

Demzufolge stieg die Elektroenergieerzeugung in der DDR von 1955 bis 1971 auf 241,8% bzw. im Durchschnitt jährlich um 10,6%. Die Beobachtungswerte und die absolute Gesamtveränderung sind in Abbildung 34 dargestellt.

<sup>38</sup>Statistisches Jahrbuch der DDR 1972. Staatsverlag der DDR, Berlin 1972, S. 128

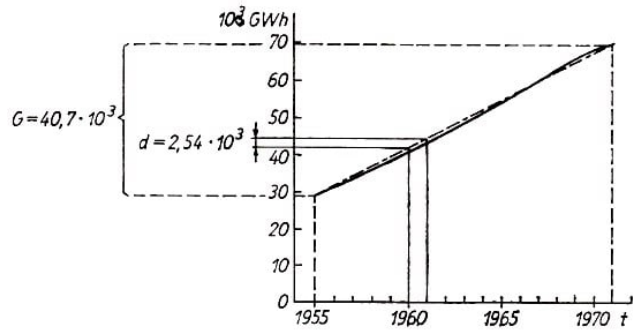


Abb. 34 Entwicklung der Elektroenergieerzeugung in der DDR von 1955 bis 1971

Ein weiteres einfaches Verfahren zur näherungsweisen Bestimmung der Grundrichtung der Entwicklung ist die schrittweise Ausgleiche der Zeitreihe auf der Grundlage einer Durchschnittsbildung aus einer geraden oder ungeraden Anzahl von Beobachtungswerten.

Je nach der Anzahl der zur Durchschnittsbildung herangezogenen Werte verkürzt sich aber die Reihe der gleitenden Durchschnitte gegenüber der Reihe der Beobachtungswerte. Vom Bearbeiter muss deshalb zunächst auf Grund der vorliegenden statistischen Zeitreihe entschieden werden, in welchem Umfang eine Ausgleiche erfolgen kann und wieviel Beobachtungswerte demzufolge für die Berechnung der Durchschnitte verwendet werden dürfen.

Allgemein wird empfohlen, die Zahl der für die Durchschnittsbildung heranzuziehenden empirischen Werte von der festgestellten Periodenlänge abhängig zu machen. Im Ergebnis erhält man gegenüber der mehr oder weniger stark schwankenden Kurve der Beobachtungswerte eine ausgeglichene Kurve, die bereits deutlich die Grundtendenz der Entwicklung sichtbar macht.

Die Berechnung der gleitenden Durchschnitte aus einer ungeraden Anzahl von Beobachtungswerten, hier aus drei Werten, geschieht wie folgt:

$$\begin{aligned}\bar{y}_2 &= \frac{y_1 + y_2 + y_3}{3} \\ \bar{y}_3 &= \frac{y_2 + y_3 + y_4}{3} \\ \bar{y}_{k-1} &= \frac{y_{k-2} + y_{k-1} + y_k}{3}\end{aligned}\quad (125)$$

Analog zu Formel (125) werden die gleitenden Durchschnitte aus fünf, sieben, neun usw. Werten bestimmt.

Ist jedoch der gleitende Durchschnitt aus einer geraden Anzahl von Beobachtungswerten zu bilden, so muss im Interesse einer klaren Zuordnung der einzelnen Durchschnitte  $\bar{y}_i$  zu den entsprechenden Zeitwerten  $t_i$  nach folgendem Bildungsprinzip, hier für 4 Werte, vorgegangen werden:

$$\begin{aligned}\bar{y}_3 &= \frac{0,5y_1 + y_2 + y_3 + y_4 + 0,5y_5}{4}, \quad \bar{y}_4 = \frac{0,5y_2 + y_3 + y_4 + y_5 + 0,5y_6}{4} \\ \bar{y}_5 &= \frac{0,5y_{k-4} + y_{k-3} + y_{k-2} + y_{k-1} + 0,5y_k}{4}\end{aligned}\quad (126)$$

Am Beispiel des wöchentlichen Handelsumsatzes eines Kaufhauses soll die Berechnung der gleitenden Durchschnitte in Tabelle 32 gezeigt werden.

Woche	Handels- Umsatz (in Mill. M)	Gleitender 3-Wochen- Durchschnitt	7-Wochen- Durchschnitt
1	4,8	-	-
2	4,3	4,7	-
3	5,1	4,6	4,7
4	4,4	4,8	4,8
5	4,9	4,9	4,9
6	5,3	5,1	4,9
7	5,0	5,0	5,0
8	4,6	4,9	5,0
9	5,1	5,0	-
10	5,4	-	-

Tabelle 32 Entwicklung des Handelsumsatzes eines Kaufhauses in 10 Wochen

Entsprechend den Formeln (125) und (126) werden die gleitenden Durchschnitte für 3 und 4 Werte bestimmt. Die ersten gleitenden 3-Wochen-Durchschnitte wären zu berechnen als

$$\bar{y}_2 = \frac{4,8 + 4,3 + 5,1}{3} = 4,7 \quad \text{und} \quad \bar{y}_3 = \frac{4,3 + 5,1 + 4,4}{3} = 4,6$$

Sind hingegen die 4-Wochen-Durchschnitte für die Untersuchung erforderlich, so werden die ersten bestimmt als

$$\bar{y}_3 = \frac{2,4 + 4,3 + 5,1 + 4,4 + 2,45}{4} = 4,7 \quad \text{und}$$

$$\bar{y}_4 = \frac{2,15 + 5,1 + 4,4 + 4,9 + 2,65}{4} = 4,8$$

Aus Tabelle 32 ist ersichtlich, dass sich mit zunehmender Zahl der in die Berechnung einbezogenen Werte eine immer stärkere Ausgleicheung ergibt, jedoch gleichzeitig eine immer stärkere Verkürzung der Reihe der Durchschnitte gegenüber der der Beobachtungswerte eintritt.

Auch in Abbildung 35, in der die Beobachtungswerte und die zwei Reihen der gleitenden Durchschnitte dargestellt sind, ist diese Ausgleicheung deutlich erkennbar.

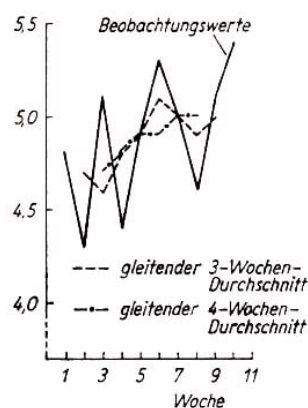


Abb. 35 Beobachtungswerte und gleitende Durchschnitte des Woche Handelsumsatzes eines Kaufhauses

Mit Hilfe der bisher behandelten Verfahren ist es möglich, die Entwicklung der Grundrichtung grob zu bestimmen, wobei die erzielten Aussagen häufig als ausreichend für erste Entscheidungen anzusehen sind. Ist es jedoch notwendig, einen knappen und übersichtlichen Ausdruck für den Grundverlauf einer statistischen Zeitreihe zu bestimmen, so muss eine mathematische Funktion gefunden werden, zu der die Beobachtungswerte auszugleichen sind.

Für diese mathematische Ausgleichung können ganze rationale Funktionen  $n$ -ten Grades, aber auch weitere Funktionsansätze, wie Potenz- oder Exponentialfunktion, verwendet werden. Da zudem einzelne Erscheinungen im Verlauf ihrer Entwicklung einer Sättigungsgrenze zustreben, kann dieser Umstand z. B. durch die dafür geeigneten Hyperbel- oder logistischen Funktionen berücksichtigt werden.

Welche der Funktionen für den konkreten Sachverhalt zu verwenden ist, hängt von den der Erscheinung innewohnenden Gesetzmäßigkeiten ab. Da diese Gesetzmäßigkeiten jedoch ihren konkreten Ausdruck in den Beobachtungswerten finden, wird die geeignete Funktion zweckmäßigerweise anhand der grafischen Darstellung der Beobachtungswerte ausgewählt.

Es muss an dieser Stelle jedoch noch auf die Parallelität zur Regressionsrechnung (vgl. Abschnitt 5.3.) hingewiesen werden, denn formal wird die Trendfunktion  $Y = f(t)$  als Abhängigkeit der Erscheinung  $y$  von der Zeit  $t$  bestimmt.

Die Zeit  $t$  ist jedoch nur Ausdruck des Wirkens einer Reihe von Faktoren auf die Erscheinung  $y$ , so dass die Trendfunktion lediglich als Ausdruck der Proportionalität zwischen der Variabilität der Zeit  $t$  und der des Tatbestandes  $y$  zu verstehen ist.

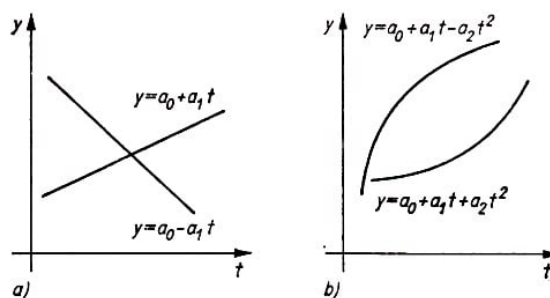


Abb. 36 Typische Verlaufformen ganzer rationaler Funktionen  $n$ -ten Grades  
a) lineare Funktionen, b) quadratische Funktionen

Die ganze rationale Funktion  $n$ -ten Grades lautet allgemein

$$Y = a_0 + a_1 t + a_2 t^2 + \dots + a_n t^n \quad (127)$$

Aus ihr können die verschiedenen Trendformen abgeleitet werden. Typische Verlaufsformen ganzer rationaler Funktionen  $n$ -ten Grades sind in Abbildung 36 wiedergegeben.

Um eine möglichst gute Annäherung der Trendfunktion an die Beobachtungswerte zu

erreichen, wird der Berechnung der Konstanten  $a_0$  bis an die quadratische Minimumbedingung

$$\sum_{i=0}^n (y_i - Y_i)^2 = \text{Minimum} \quad (128)$$

zugrunde gelegt. Aus der Ableitung dieser Minimumbedingung, die die Forderung beinhaltet, dass die Summe der Quadrate der Abweichungen der empirischen Werte von den Funktionswerten ein Minimum sein soll, ergeben sich die Normalgleichungen in ihrer allgemeinen Form:

$$\begin{aligned} \sum_{i=1}^n y_i &= a_0 k + a_1 \sum_{i=1}^n t_i + \dots + a_n \sum_{i=1}^n t_i^n \\ \sum_{i=1}^n y_i t_i &= a_0 \sum_{i=1}^n t_i + a_1 \sum_{i=1}^n t_i^2 + \dots + \sum_{i=1}^n t_i^{n+1} \\ \sum_{i=1}^n y_i t_i^n &= a_0 \sum_{i=1}^n t_i^n + a_1 \sum_{i=1}^n t_i^{n+1} + \dots + a_n \sum_{i=1}^n t_i^{2n} \end{aligned} \quad (129)$$

Je nach der zu berechnenden Trendfunktion und den in ihr enthaltenen Konstanten  $a_i$  müssen wir die entsprechenden Normalgleichungen bilden. Für die lineare Trendfunktion (Trend 1. Grades)

$$Y = f(t) = a_0 + a_1 t \quad (130)$$

lauten die zwei Normalgleichungen nach Formel (129)

$$\begin{aligned} \sum_{i=1}^n y_i &= a_0 k + a_1 \sum_{i=1}^n t_i \\ \sum_{i=1}^n y_i t_i &= a_0 \sum_{i=1}^n t_i + a_1 \sum_{i=1}^n t_i^2 \end{aligned} \quad (131)$$

Die Konstante  $a_0$  bringt den Betrag der Funktion zum Zeitpunkt  $t_0$  zum Ausdruck und ist in der grafischen Darstellung der Schnittpunkt der Funktion mit der Ordinate. Die Konstante  $a_1$  gibt die Veränderung der Größe  $Y$  in der gewählten Zeiteinheit an. Für die Parabelfunktion (Trend 2. Grades)

$$Y = f(t) = a_0 + a_1 t + a_2 t^2 \quad (132)$$

können wir die Konstanten  $a_0$ ,  $a_1$  und  $a_2$  durch die Normalgleichungen

$$\begin{aligned} \sum_{i=1}^n y_i &= a_0 k + a_1 \sum_{i=1}^n t_i + a_2 \sum_{i=1}^n t_i^2 \\ \sum_{i=1}^n y_i t_i &= a_0 \sum_{i=1}^n t_i + a_1 \sum_{i=1}^n t_i^2 + a_2 \sum_{i=1}^n t_i^3 \\ \sum_{i=1}^n y_i t_i^2 &= a_0 \sum_{i=1}^n t_i^2 + a_1 \sum_{i=1}^n t_i^3 + a_2 \sum_{i=1}^n t_i^4 \end{aligned} \quad (133)$$

bestimmen.

Besondere Aufmerksamkeit müssen wir auch der Wahl geeigneter  $t$ -Werte schenken. Zunächst können die einzelnen Zeitabschnitte in ihrer Reihenfolge mit Ordnungsnummern von  $1 \dots k$  bezeichnet und diese Ordnungsnummern in den weiteren Berechnungen verwendet werden. Dabei muss jedoch beachtet werden, dass sich der Abstand zwischen den einzelnen Zeitabschnitten auch in den für die einzelnen Zeitabstände gewählten Ordnungsnummern widerspiegelt.

Sind die Beobachtungswerte aus fortlaufenden Zeitabschnitten gegeben und ist die Reihe der Beobachtungswerte nicht unterbrochen, kann die Zeitachsenmitte in den Ursprung verschoben werden. Da nun negative und positive Ordnungsnummern auftreten, die als  $t_i$ -Werte verwendet werden, gilt für deren Potenzsumme mit ungeraden Exponenten

$$\sum_{i=-n}^{+n} t_i = \sum_{i=-n}^{+n} t_i^3 = \sum_{i=-n}^{+n} t_i^5 = \dots = 0$$

wodurch eine wesentliche Vereinfachung der Normalgleichungen erzielt wird. An Beispielen soll in Tabelle 33 die Wahl entsprechender  $t_i$ -Werte demonstriert werden.

nichtfortlaufende Zeitabschnitte			Ordnungsnummern für fortlaufende Zeitabschnitte			
			bei gerader Anzahl von Zeitabschnitten		bei ungerader Anzahl von Zeitabschnitten	
Zeitabschnitt	$t_i$	$t_i$	Zeitabschnitt	$t_i$	Zeitabschnitt	$t_i$
1955	1	-8	Jan.	-7	So.	3
1960	6	-3	Febr.	-5	Mo.	2
1965	11	+2	März	-3	Di.	-1
1967	13	+4	Apr.	-1	Mi.	0
1968	14	+5	Mai	+1	Do.	+1
1969	15	+6	Juni	+3	Fr.	+2
1970	16	+7	Juli	+5	Sa.	+3
1971	17	+8	Aug.	+7		
Summe $t_i$	93	+17	Summe $t_i$	0	Summe $t_i$	0

Tabelle 33 Beispiele für die Wahl von  $t_i$ -Werten

Die Normalgleichungen lauten dann, wenn die Summe  $t_i$  gleich Null ist, beim Trend 1. Grades

$$\sum_i y_i = a_0 k \quad , \quad \sum_i y_i t_i = a_1 \sum_i t_i^2 \quad (134)$$

und beim Trend 2. Grades

$$\begin{aligned} \sum_i y_i &= a_0 k + a_2 \sum_i t_i^2 \\ \sum_i y_i t_i &= a_1 \sum_i t_i^2 \\ \sum_i y_i t_i^2 &= a_0 \sum_i t_i^2 + a_2 \sum_i t_i^4 \end{aligned} \quad (135)$$

Am Beispiel der Selbstkostenentwicklung eines Erzeugnisses über einen längeren Zeitraum (vgl. Tab. 34) sollen die Trendfunktionen 1. und 2. Grades berechnet werden. Entsprechend den Gleichungen (134) und (135) müssen die Produkte und Summen bestimmt werden, wobei auch hier wieder die tabellarische Berechnung entsprechend Tabelle 34 zu bevorzugen ist.

Jahr	$t_i$	Selbstkosten in M/ Erzeugnis ( $y_i$ )	$t_i^2$	$t_i^4$	$y_i t_i$	$y_i t_i^2$	$Y_I$	$Y_{II}$
1966	-7	8,8	49	2401	-61,6	431,2	8,2	8,75
1967	-5	7,9	25	625	-39,5	197,5	7,8	7,92
1968	-3	7,3	9	81	-21,9	65,7	7,4	7,13
1969	-1	6,7	1	1	-6,7	6,7	7,0	6,66
1970	+1	6,1	1	1	+6,7	6,1	6,6	6,22
1971	+3	5,9	9	81	+17,7	53,1	6,2	5,81
1972	+5	5,9	25	625	+29,5	147,5	5,8	5,72
1973	+7	5,6	9	2401	+39,2	274,4	5,4	5,66
	0	54,2	168	6216	-37,2	1182,2	-	-

Tabelle 34 Berechnung der Trendfunktionen 1. und 2. Grades für die Selbstkostenentwicklung eines Erzeugnisses

Wir erhalten die Konstanten der Trendfunktion 1. Grades mit

$$a_0 = \frac{54,2}{8} = 6,8 \quad , \quad a_1 = \frac{-37,2}{168} = -0,2$$

aus denen die Trendfunktion

$$Y_I = 6,8 - 0,2t$$

folgt. Als Trendfunktion 2. Grades gewinnen wir analog durch Einsetzen der Summen aus Tabelle 34 in das Gleichungssystem (135)

$$Y_{II} = 6,43 - 0,22t + 0,016t^2$$

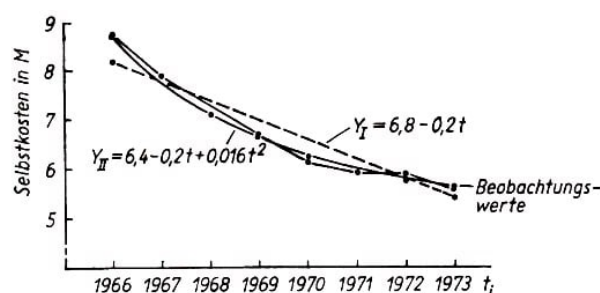


Abb. 37 Beobachtungs- und Trendwerte der Selbstkostenentwicklung eines Erzeugnisses

Um die Funktionswerte zu bestimmen, müssen die einzelnen  $t_i$ -Werte in diese Trendfunktionen eingesetzt werden. Es ergeben sich die in den beiden letzten Spalten von



Tabelle 34 enthaltenen Werte.

Zusammen mit den Beobachtungswerten sind die beiden Trendfunktionen in Abbildung 37 dargestellt. Es muss jetzt allerdings entschieden werden, welche der Funktionen am besten die Grundrichtung der Selbstkostenentwicklung für das Erzeugnis widerspiegelt. Das kann einmal auf Grund des optischen Eindrucks vom Grad der Anpassung der Trendkurven an die Reihe der Beobachtungswerte erfolgen. Eine solche Vorgehensweise schließt aber Fehlentscheidungen nicht aus.

Es empfiehlt sich deshalb, die Güte der Anpassung mit Hilfe der bereits früher eingeführten Standardabweichung (vgl. Abschnitt 4.4.) zu beurteilen, wonach die Trendfunktion mit der geringsten Abweichung zwischen empirischen und Funktionswerten als für den Sachverhalt zutreffend ausgewählt wird.

Die Standardabweichung lautet für diesen Fall allgemein

$$s = \sqrt{\frac{1}{k} \sum_i (y_i - Y_i)^2} \quad (136)$$

Für unser Beispiel sind dazu die in Tabelle 35 enthaltenen Werte zu berechnen.

$y_i$	$Y_{Ii}$	$y_i - Y_{Ii}$	$(y_i - Y_{Ii})^2$	$Y_{IIi}$	$y_i - Y_{IIi}$	$(y_i - Y_{IIi})^2$
8,8	8,2	+0,6	+0,36	8,75	+0,05	+0,0025
7,9	7,8	+0,1	+0,01	7,92	-0,02	+0,0004
7,3	7,4	-0,1	+0,01	7,13	+0,17	+0,0289
6,7	7,0	-0,3	+0,09	6,66	+0,04	+0,0016
6,1	6,6	-0,5	+0,25	6,22	-0,12	+0,0144
5,9	6,2	-0,3	+0,09	5,81	+0,09	+0,0081
5,9	5,8	+0,1	+0,01	5,72	+0,18	+0,0324
5,6	5,4	+0,2	+0,04	5,66	-0,06	+0,0036
Summe	-	-	+0,86	-	-	+0,0919

Tabelle 35 Berechnung der Abweichungsquadrate für die Trendfunktionen 1. und 2. Grades

Für die berechneten Trendfunktionen ergeben sich nach Formel (136) folgende Standardabweichungen:

$$s_I = \sqrt{\frac{1}{8} \cdot 0,86} = 0,3279 \quad , \quad s_{II} = \sqrt{\frac{1}{8} \cdot 0,0919} = 0,01070$$

Auf Grund der Tatsache, dass der Trend 2. Grades die geringere Abweichung von den Beobachtungswerten aufweist, ist er zur Beurteilung der Selbstkostenentwicklung für das Erzeugnis gegenüber dem Trend 1. Grades vorzuziehen.

Neben den ganzen rationalen Funktionen  $n$ -ten Grades sind jedoch auch andere Funktionsansätze für die Bestimmung des Trends denkbar. Hierzu gehören insbesondere die Potenz-, Exponential- und logarithmischen Funktionen sowie Funktionen mit Sättigungsgrenze.

Es würde hier zu weit führen, wenn diese Funktionen hinsichtlich der konkreten Bestimmung der Konstanten alle behandelt würden. Vielmehr sollen nur die zuerst genannten Funktionen in ihrer allgemeinen Form dargestellt und in ihren typischen Verlaufsformen interpretiert werden.

Die allgemeine Form der Potenzfunktionen lautet

$$Y = a_0 \cdot t^{a_1} \quad (137)$$

die, um mit Hilfe der früher genannten Normalgleichungen die

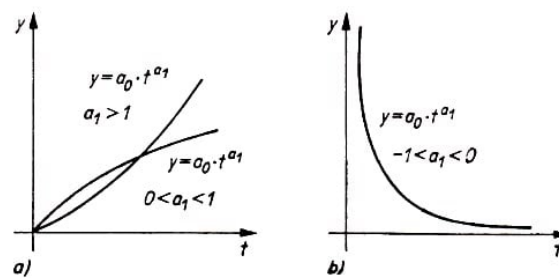


Abb. 38 Potenzfunktionen

Konstanten  $a_0$  und  $a_1$  bestimmen zu können, umgeformt werden muss in

$$\lg Y = \lg a_0 + a_1 \lg t$$

Je nach der Größe von  $a_1$  ergeben sich unterschiedliche Kurven, die in Abbildung 38 wiedergegeben sind.

Der Exponentialfunktion liegt die Beziehung

$$Y = a_0 \cdot a_1^t \quad (138)$$

zugrunde, die zur Bestimmung der Konstanten gleichfalls umgeformt werden muss in

$$Y = \lg a_0 + t \lg a_1$$

Um eine allgemeine Vorstellung vom Verlauf der Exponentialfunktion zu vermitteln, sind in Abbildung 39 typische Verlaufsformen dieser Funktion dargestellt.

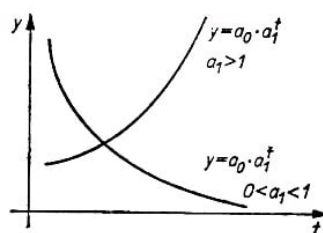


Abb. 39 Exponentialfunktionen

Die Exponentialfunktion wird vorwiegend zur Beschreibung progressiver Entwicklungstendenzen eingesetzt, wobei entweder ein progressives Ansteigen oder Abfallen der Kurve festzustellen ist.

Im Gegensatz zur Exponentialfunktion kann die logarithmische Funktion zur Beschreibung degressiv steigender oder fallender Entwicklungsprozesse eingesetzt werden. Die allgemeine Form lautet

$$Y = a_0 + a_1 \lg t \quad (139)$$

in der die Konstanten  $a_0$  und  $a_1$  wieder analog den Beziehungen (131) berechnet werden können. In Abbildung 40 sind Kurven von logarithmischen Funktionen wiedergegeben.

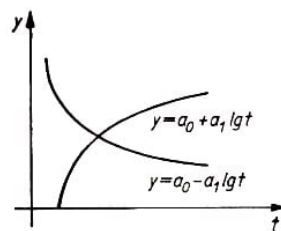


Abb. 40. Logarithmische Funktionen

Es sei jedoch in diesem Zusammenhang betont, dass alle Trendfunktionen, werden sie zur Interpretation bestimmter Entwicklungsvorgänge verwendet, nur mathematische Hilfsmittel sind, um die Grundrichtung der Entwicklung näher bestimmen zu können. Sie können nie losgelöst von der untersuchten Erscheinung und deren Ursachen betrachtet werden.

## 9.2 Berechnung der Periodizität einer Erscheinung

Wenden wir uns einer zweiten Komponente zu, den periodischen Schwankungen. Sowohl bei naturwissenschaftlichen als auch technisch-technologischen und ökonomischen Erscheinungen können derartige regelmäßig wiederkehrende Schwankungen festgestellt werden.

Sie haben ihre Ursache in Faktoren, die mit periodisch sich wiederholender unterschiedlicher Intensität auf die Erscheinung einwirken.

Bei der Berechnung periodischer Schwankungen ist allerdings zu beachten, dass diese verzerrt wiedergegeben werden, wenn der Entwicklung ein Trend zugrunde liegt. Wird die Reihe der Beobachtungswerte nicht vom Trend bereinigt, werden in diesen Fällen, je nach der Richtung des Trends, die durchschnittlichen Schwankungsgrößen der ersten Perioden zu hoch bzw. zu niedrig und die der letzten Perioden zu niedrig bzw. zu hoch angegeben.

Aus diesem Grund werden für stationäre Zeitreihen, deren Entwicklungsniveau keinen oder nur unwesentlichen Veränderungen unterworfen ist, und für Zeitreihen mit Trend einfluss unterschiedliche Verfahren angewandt.

Wenden wir uns zunächst der Bestimmung periodischer Schwankungen bei stationären Zeitreihen zu. Als Beispiele hierfür wären insbesondere meteorologische Erscheinungen,

wie Temperaturschwankungen und Niederschlagsmengen, aber auch viele technisch-technologische Abläufe zu nennen.

Die periodischen Schwankungen können bei derartigen Zeitreihen relativ einfach mit Hilfe des Phasendurchschnittsverfahrens gewonnen werden. Die Einzelwerte werden mit  $y_{ij}$  bezeichnet, wobei der Zeiger  $i$  für den Untersuchungszeitraum steht und von  $1 \dots n$  läuft, während der Zeiger  $j$  den Zeitabschnitt innerhalb des Untersuchungszeitraumes angibt und von  $1 \dots k$  laufen soll.

Beispiele für den Inhalt der Zeiger  $i$  und  $j$  sind Jahr und Monat, Monat und Dekade, Dekade und Tag sowie Tag und Stunde. Beim Phasendurchschnittsverfahren sind nacheinander die Phasenzahlen, die durchschnittliche Phasenzahl, die Phasenveränderungszahlen und/oder die Phasenindexzahlen zu bestimmen.

Die Phasenzahlen ergeben sich allgemein nach der Formel

$$\bar{y}_j = \frac{1}{k} \sum_{i=1}^n y_{ij} \quad (140)$$

und geben den Durchschnitt für mehrere gleiche Zeitabschnitte verschiedener Untersuchungszeiträume an. Aus diesen Phasenzahlen wird die durchschnittliche Phasenzahl gebildet, indem die Phasenzahlen summiert und durch die Anzahl der Zeiträume geteilt werden. Allgemein kann danach die durchschnittliche Phasenzahl geschrieben werden als

$$\bar{y} = \frac{1}{k} \sum_{j=1}^n \bar{y}_j \quad (141)$$

Die Phasenveränderungszahlen ergeben sich aus der Differenz zwischen den einzelnen Phasenzahlen und der durchschnittlichen Phasenzahl. Sie bringen die mittlere absolute Abweichung der einzelnen Periodenwerte vom Durchschnitt zum Ausdruck. Allgemein kann die Phasenveränderungszahl geschrieben werden

$$Z_j = \bar{y}_j - \bar{y} \quad (142)$$

↓ Tag (j)	Woche (i) →	1	2	3	4	$\sum_{i=1}^4 y_{ij}$
So.	2,75	2,64	2,68	2,73	10,80	
Mo.	2,30	2,20	2,26	2,28	9,04	
Di.	1,85	1,76	1,82	1,87	7,30	
Mi.	2,40	2,33	2,44	2,50	9,67	
Do.	1,97	1,88	2,02	2,05	7,92	
Fr.	1,58	1,47	1,58	1,61	6,24	
Sa.	1,14	1,05	1,16	1,08	3,43	

Tabelle 36 Täglicher Bestand an festen Brennstoffen (in 100 t) in einem Industriebetrieb

Sollen hingegen die relativen Abweichungen der Periodenwerte  $\bar{y}_i$  vom Durchschnitt aller Periodenwerte  $\bar{y}$  zum Ausdruck gebracht werden, sind die Phasenindexzahlen als

$$I_j = \frac{\bar{y}_j}{\bar{y}} \cdot 100(\%) \quad (143)$$

zu bilden.

Die Anwendung des Phasendurchschnittsverfahrens soll am Beispiel der täglichen Schwankungen des Bestandes an festen Brennstoffen in einem Industriebetrieb demonstriert werden. Die täglichen Bestandszahlen über 4 Wochen sind in Tabelle 36 zusammengestellt. Die grafische Darstellung der Beobachtungswerte in Abbildung 41 zeigt deutlich die sich wöchentlich wiederholenden Schwankungen des Bestandes.

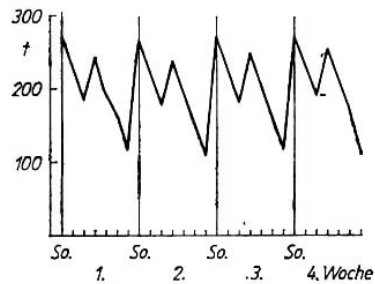


Abb. 41 Schwankungen des Bestandes an festen Brennstoffen

In der letzten Spalte der Tabelle 36 sind außerdem die Summen der  $y_{ij}$  über alle Zeiträume  $i$  summiert, die für die Bestimmung der Phasenzahlen erforderlich sind. Die erste Phasenzahl wird gebildet als

$$\bar{y}_1 = \frac{2,75 + 2,64 + 2,68 + 2,73}{4} = 2,7$$

Die Phasenzahlen werden summiert und durch die Anzahl der Zeitabschnitte ( $j = 7$ ) geteilt. Wir erhalten als durchschnittliche Phasenzahl  $\bar{y} = 1,9$ .

Tag ( $j$ )	$\bar{y}_j$	$Z_j$	$I_j$ (%)
So.	2,70	+0,76	139,2
Mo.	2,26	+0,32	116,5
Di.	1,83	-0,11	94,3
Mi.	2,42	+0,48	124,7
Do.	1,98	+0,04	102,0
Fr.	1,56	-0,38	80,4
Sa.	0,86	-1,08	44,3
Summe	13,61		

Tabelle 37 Phasenveränderungs- und Phasenindexzahlen des Bestandes an festen Brennstoffen in einem Industriebetrieb

In Tabelle 37 sind nun die Phasenveränderungszahlen und die Phasenindexzahlen berechnet worden.

Deutlich sind bereits die über dem Wochendurchschnitt liegenden Bestände am Sonntag und Mittwoch zu erkennen, während die Bestände am Freitag und Samstag unter den Durchschnitt absinken. Noch eindrucksvoller kommen diese Schwankungen zum Ausdruck, wenn die Phasenindexzahlen, wie in Abbildung 42, grafisch dargestellt werden.

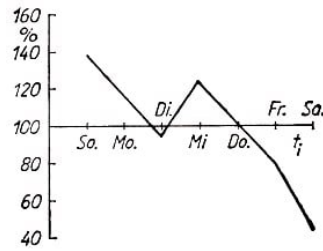


Abb. 42 Phasenindexzahlen des Bestandes an festen Brennstoffen

Die in der Reihe der Beobachtungswerte ursprünglich vorhandenen unterschiedlich großen Schwankungsstärken, hervorgerufen durch einen täglich schwankenden Verbrauch an festen Brennstoffen im untersuchten Betrieb, sind durch die Durchschnittsbildungen ausgeglichen worden.

Das Ergebnis besteht in einer Reihe absoluter oder relativer Schwankungsgrößen, die sich auf den gesamten Untersuchungszeitraum beziehen.

Sie können sowohl der Einschätzung der bisherigen Vorratswirtschaft als auch der Planung der zukünftigen Bestände mit dem Ziel einer bestmöglichen Anpassung des Bestandes an den durchschnittlichen Verbrauch dienen.

Die Mehrzahl der statistischen Zeitreihen ist jedoch nicht wie im vorhergehenden Beispiel frei vom Trendeinfluss. Soweit dieser Einfluss von der Erscheinung und dem Untersuchungsziel her als wesentlich zu werten ist, muss deshalb die Reihe der Beobachtungswerte zunächst vom Einfluss des Trends bereinigt werden, ehe die Berechnung der periodischen Schwankungen erfolgen kann.

Dabei ist zunächst zu überlegen, auf welche Art die beiden Komponenten - Trend ( $T$ ) und periodische Schwankung ( $S$ ) - in der untersuchten Zeitreihe miteinander verbunden sind. Vielfach tritt hier eine additive Verknüpfung in Form von

$$T + S$$

auf, doch liegt bei allen nichtlinearen Trendfunktionen eine Verbindung zwischen den beiden Komponenten durch Multiplikation oder Division vor.

Es soll hier nur der Fall der additiven Verknüpfung näher betrachtet werden, da für viele Entwicklungsprozesse von vornherein eine lineare Trendfunktion zur Beschreibung der Grundrichtung der Entwicklung angesetzt werden kann. Aber selbst bei einer nichtlinearen Entwicklungstendenz wird häufig, da die lineare Funktion am besten zu interpretieren ist, eine Unterteilung der Zeitreihe und getrennte Berechnung linearer Funktionen für die einzelnen Teile vorgenommen.

Berechnet werden können bei der additiven Verknüpfung der beiden Komponenten

- die durchschnittliche Größe der absoluten Schwankung ( $\bar{c}_j$ ) und
- die durchschnittliche Größe der relativen Schwankung ( $\bar{P}_j$ ).

Beide Größen berücksichtigen den Einfluss eines linearen Trends, indem für die Bestimmung der Schwankungsgrößen die Differenzen bzw. Quotienten zwischen den Beobachtungswerten und den Trendwerten verwendet werden.

Um die durchschnittliche Größe der absoluten Schwankung, die vom Trendeinfluss befreit ist, bestimmen zu können, müssen wir zunächst für alle empirischen Werte  $y_{ij}$  die in diesem Fall angenommene lineare Trendfunktion und dann für alle  $t_{ij}$  die Funktionswerte  $y_{ij}$  berechnen.

Da uns nun eine Schwankungsgröße interessiert, die frei vom Trendeinfluss ist, müssen wir für die einzelnen Zeitabschnitte über alle gegebenen Zeiträume die empirischen und Funktionswerte summieren und die Differenz zwischen ihnen bilden. Die jetzt summarisch für alle untersuchten Zeiträume vorliegenden absoluten Schwankungsgrößen werden durch die Anzahl der Zeiträume  $n$  geteilt und ergeben für alle Untersuchungszeitabschnitte die jeweilige durchschnittliche Größe der absoluten Schwankung.

Es sei an dieser Stelle nur noch auf die im Prinzip identische Formel (122) verwiesen. Die Formel für die durchschnittliche Größe der absoluten Schwankung lautet

$$\bar{c}_j = \frac{\sum_i y_{ij} - \sum_i Y_{ij}}{m} \quad (144)$$

mit  $y$  - Beobachtungswerte

$Y$  - Funktionswerte

$i$  - Zeiger für Untersuchungszeitraum (1 bis  $n$ )

$j$  - Zeiger für Untersuchungszeitabschnitt (1 bis  $k$ ).

Werden die gleichen Bezeichnungen wie in Formel (144) verwendet, ergibt sich die Beziehung für die durchschnittliche Größe der relativen Schwankung als

$$\bar{P}_j = \frac{1}{n} \sum_i \frac{y_{ij}}{Y_{ij}} \cdot 100(\%) \quad (145)$$

Auch hier ist unschwer die Erweiterung der bereits diskutierten Formel (123) zu sehen, wobei die Ausgangsüberlegungen für die Berechnung der durchschnittlichen Größe der relativen Schwankung die gleichen wie für Formel (144) sind.

Am Beispiel der Entwicklung der Transportleistung eines Verkehrsbetriebes in drei Jahren soll die Berechnung der durchschnittlichen Größe der absoluten und relativen Schwankung gezeigt werden. Die Entwicklung der Transportleistung, ausgedrückt in Tonnenkilometern, ist in Tabelle 38 enthalten.

Quartal ( $j$ )	Jahr ( $i$ ) $y_{ij}$			$\sum_{i=1}^3 y_{ij}$
	1	2	3	
I	12,8	14,3	16,9	44,0
II	9,7	10,2	13,1	33,0
III	11,0	12,4	13,8	37,2
IV	14,8	16,8	19,3	50,9

Tabelle 38 Entwicklung der Transportleistung eines Verkehrsbetriebes (in  $10^4$  tkm)

Quartal ( $j$ )	Jahr ( $i$ ) $Y_{ij}$			$\sum_{i=1}^3 Y_{ij}$
	1	2	3	
I	14,01	13,04	15,01	39,03
II	14,51	13,51	15,51	40,53
III	12,01	14,04	16,04	42,03
IV	12,51	14,51	16,51	43,53

Tabelle 39 Funktionswerte des Trends 1. Grades für die Entwicklung der Transportleistung (in  $10^4$  tkm)

Wir haben im ersten Bearbeitungsschritt die Trendfunktion nach den Beziehungen (134) zu bestimmen. Auf Grund der Ausgangswerte in Tabelle 38 berechnen wir die Konstanten zu

$$a_0 = 13,76 \quad , \quad a_1 = 0,25$$

und erhalten als Trendfunktion

$$Y = 13,76 + 0,25t$$

Um die einzelnen Funktionswerte zu gewinnen, müssen die gewählten  $t_{ij}$ -Werte (hier -11, -9, ... bis +9, +11) in die Trendfunktion eingesetzt werden. Es ergeben sich die in Tabelle 39 enthaltenen Werte für  $Y_{ij}$ .

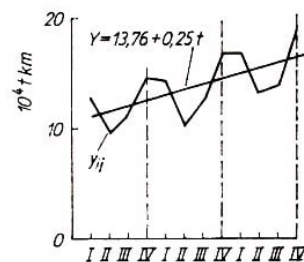


Abb. 43 Beobachtungswerte ( $y_{ij}$ ) und Funktionswerte ( $Y_{ij}$ ) der Entwicklung der Transportleistung

In Abbildung 43 sind die empirischen und die Funktionswerte dargestellt.

Quartal	$\sum_{i=1}^3 y_{ij} - \sum_{i=1}^3 Y_{ij}$	$\bar{c}_j$
I	+4,97	+1,66
II	-7,53	-2,51
III	-4,83	-1,61
IV	+7,37	+2,46

Tabelle 40 Berechnung der  $\bar{c}_j$  für die Transportleistung eines Verkehrsbetriebes

Die durchschnittliche Größe der absoluten Schwankung kann nun anhand der Angaben der jeweils letzten Spalte der Tabellen 38 und 39 berechnet werden. Die Ergebnisse sind in Tabelle 40 enthalten.

Aus den Werten der  $\bar{c}_i$  für die einzelnen Quartale folgt, dass nach Ausschalten des Trendeinflusses auf die Zeitreihe die Transportleistung im I. Quartal um  $1,66 \cdot 10^4$  tkm



und im IV. Quartal um  $2,46 \cdot 10^4$  tkm über dem Jahresdurchschnitt liegt. Hingegen liegen die Leistungen im II. und III. Quartal um  $2,51 \cdot 10^4$  bzw.  $1,61 \cdot 10^4$  tkm unter dem Jahresdurchschnitt.

Quartal ( $j$ )	Jahr ( $i$ ) $y_{ij} : Y_{ij}$			$\sum_{i=1}^3 \frac{y_{ij}}{Y_{ij}}$	$\bar{P}_j = \frac{1}{n} \sum_{i=1}^3 \frac{y_{ij}}{Y_{ij}}$
	1	2	3		
I	1,153	1,099	1,126	3,378	112,6
II	0,843	0,755	0,845	2,443	81,4
III	0,916	0,885	0,862	2,663	88,8
IV	1,183	1,158	1,169	3,510	117,0

Tabelle 41 Berechnung der  $\bar{P}_j$  für die Transportleistung eines Verkehrsbetriebes

Für die Berechnung der durchschnittlichen Größe der relativen Schwankung sind zunächst für alle Zeitabschnitte der drei Jahre die Quotienten aus Beobachtungs- und Funktionswerten zu bilden. Die Ergebnisse werden anschließend für gleiche Zeitabschnitte ( $j$ ) summiert und durch die Anzahl der Zeiträume ( $i$ ) geteilt. Die Berechnungen für unser Beispiel erfolgen in Tabelle 41.

Danach liegen die Transportleistungen im I. Quartal um 12,6% und im IV. Quartal um 17,0% über dem Jahresdurchschnitt, während im II. und III. Quartal deutlich die Leistung unter dem genannten Durchschnitt liegt. Die grafische Darstellung der Schwankungen ist in Abbildung 44 enthalten.

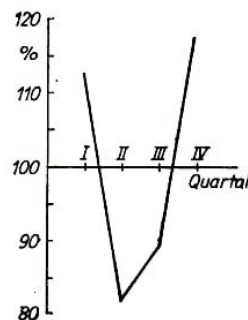


Abb. 44 Periodische Schwankungen der Transportleistung eines Verkehrsbetriebes

### 9.3 Die Veränderung von Zeitabschnitt zu Zeitabschnitt

Wir wollen uns nun im Zusammenhang mit der Untersuchung zeitlicher statistischer Reihen noch den Möglichkeiten zur Messung der relativen Veränderung einer Erscheinung von Zeitabschnitt zu Zeitabschnitt zuwenden. Diese relative Veränderung wird als Index bezeichnet, wobei zwischen einfachen und zusammengesetzten Indexzahlen oder Indizes zu unterscheiden ist.

Sollen die einer einzelnen statistischen Zeitreihe innewohnenden relativen Veränderungen bestimmt werden, kann je nach der Aufgabenstellung ein einfacher Index mit fester oder variabler Basis gebildet werden. Dem Leser sind derartige Zahlen als Informations- und Agitationsmittel bekannt, die gestatten, Reihen mit unterschiedlichem Anfangsniveau und voneinander abweichenden Veränderungen in der Entwicklung der Reihen auf

vergleichbare relative Größen umzurechnen.

Die Entwicklung einzelner Erscheinungen kann zudem durch derartige Indexzahlen deutlicher zum Ausdruck gebracht werden, als es durch die Reihe der Beobachtungswerte möglich ist. Diese Vorteile der Indexzahlen werden in vielfältiger Weise für die Analyse von Entwicklungsvorgängen genutzt.

Wir finden Indexzahlen in den Volkswirtschaftsplänen und den Abrechnungen dieser Pläne, da sie in knapper Form die geplanten und tatsächlich eingetretenen Veränderungen zum Ausdruck bringen. Indexzahlen sind im allgemeinen informativer als die absoluten Werte.

Würden wir in absoluten Werten z. B. die Veränderung der Arbeitsproduktivität mit der des Durchschnittslohnes vergleichen, könnten wir zwar eine Aussage über die absolute Höhe der im betrachteten Zeitabschnitt vor sich gegangenen Erhöhung beider Kennziffern treffen, ein Vergleich jedoch, der zeigt, inwieweit die Arbeitsproduktivität schneller gewachsen ist als der Durchschnittslohn, wird erst durch entsprechende Indexzahlen möglich.

In vielen Fällen interessiert die Entwicklung in Form der relativen Veränderung gegenüber einem typischen Zeitabschnitt, der sogenannten Basis. Dieser Basiszeitabschnitt kann sowohl der erste ( $t_1$ ) oder letzte Zeitabschnitt ( $t_k$ ) einer Reihe als auch ein typischer Zeitabschnitt innerhalb der Reihe ( $t_{1+j}$ ) sein. In diesem Fall ist der Index mit fester Basis zu bestimmen als

$$I_{m/i} = \frac{y_i}{y_m} \cdot 100(\%) \quad (146)$$

mit  $y_i$  - Beobachtungswerte und  $y_m$  - Basiswert der Reihe.

Sind jedoch für die Untersuchung die relativen Veränderungen zwischen aufeinanderfolgenden Zeitabschnitten wesentlich, ist der Index mit variabler Basis zu bilden. Die allgemeine Form lautet

$$I_{i-1/i} = \frac{y_i}{y_{i-1}} \cdot 100(\%) \quad (147)$$

Jahr	Aufkommen an Kartoffeln/ha Nutzfl. (kg)	Index mit	
		fester Basis	variabler Basis
1961	460,9	100,0	100,0
1962	589,1	127,8	127,8
1963	611,4	133,7	103,8
1964	611,0	132,6	99,9
1965	600,7	130,3	98,3
1966	599,1	130,0	99,7
1967	622,4	135,0	103,9
1968	607,3	131,8	97,6
1969	607,6	131,8	100,0
1970	676,7	146,8	111,4
1971	652,9	141,7	96,5

Tabelle 42 Die Entwicklung und relative Veränderung des Aufkommens an Kartoffeln je Hektar landwirtschaftlicher Nutzfläche in der DDR von 1961 bis 1971

Am Beispiel der Entwicklung des Aufkommens an Kartoffeln je Hektar landwirtschaftlicher Nutzfläche sollen die Indizes mit fester und variabler Basis erläutert werden.<sup>39</sup> In Tabelle 42 sind die entsprechenden Werte festgehalten. Während die Indizes mit fester Basis eine annähernd kontinuierliche relative Zunahme während des gesamten Untersuchungszeitraumes von 100% im Jahr 1961 auf 141,7% im Jahr 1972 zum Ausdruck bringen, zeigt die Reihe der Indexzahlen mit variabler Basis, dass die relative Zunahme von Jahr zu Jahr erheblichen Schwankungen unterworfen ist.

Es kommt jedoch auch vor, dass die Beobachtungswerte selbst fehlen und nur eine der beiden Indexreihen bekannt ist. Dann besteht die Möglichkeit, durch Umrechnung die Reihe der jeweils anderen Indexzahlen zu gewinnen. Durch Division der Indizes mit fester Basis können die Indexzahlen mit variabler Basis gefunden werden, da

$$\frac{y_i}{y_m} : \frac{y_{i-1}}{y_m} = \frac{y_i}{y_{i-1}} \quad (148)$$

Sind hingegen die Indizes mit fester Basis gesucht, führt die Multiplikation der Indizes mit variabler Basis zu diesem Ziel

$$\frac{y_1}{y_m} \cdot \frac{y_2}{y_m} = \frac{y_2}{y_m}, \quad \frac{y_1}{y_m} \cdot \frac{y_2}{y_1} \cdot \frac{y_3}{y_2} = \frac{y_3}{y_m} \quad (149)$$

In den bisherigen Betrachtungen veränderte sich der Umfang der Erscheinungen in Abhängigkeit von der Zeit, während die Struktur der Erscheinung konstant blieb. Sind jedoch Umfang und Struktur der Erscheinung im Entwicklungsprozess Veränderungen unterworfen, müssen beide Faktoren zur Beurteilung der relativen Veränderung herangezogen werden. Die zusammengesetzten Indizes tragen dieser Aufgabenstellung Rechnung.

Die gebräuchlichsten Formen zusammengesetzter Indexzahlen sind der Aggregatindex, der Volumenindex und der Strukturindex, doch sind aus der Indextheorie noch eine Vielzahl weiterer zusammengesetzter Indexzahlen bekannt. Als bekannteste Beispiele für die Anwendung zusammengesetzter Indexzahlen seien hier der Index der Arbeitsproduktivitätsentwicklung, der Index der Selbstkostenentwicklung und der Index des Lebensstandards erwähnt.

Der Aggregatindex zeigt den Einfluss eines Faktors auf die relative Veränderung der untersuchten Erscheinung, während ein zweiter Faktor, der im allgemeinen der Gewichtung des ersten Faktors dient, als konstant für den Basis- und Berichtszeitraum angesehen wird.

Wenn die beiden Faktoren mit  $x$  und  $y$  bezeichnet und die relative Veränderung der Größe  $x$  gemessen werden soll, bestehen zwei Möglichkeiten. Entweder kann die Gewichtung durch den Wert des Basiszeitraumes ( $y_m$ )

$$I_p = \frac{\sum x_i \cdot y_m}{\sum x_m \cdot y_m} \cdot 100(\%) \quad (150)$$

<sup>39</sup>Ebenda, S. 243

oder durch den Wert des Berichtszeitraumes ( $y_i$ )

$$I_p = \frac{\sum x_i \cdot y_i}{\sum x_m \cdot y_i} \cdot 100(\%) \quad (151)$$

mit  $i$  - Zeiger für Berichtszeitraum und  $m$  - Zeiger für Basiszeitraum vorgenommen werden. In analoger Weise ist die Veränderung der Größe  $y$  bei Konstanz der Größe  $x$  zu bestimmen.

Durch derartige Indizes kann z. B. die Preisentwicklung für verschiedene Waren und die Selbstkostenentwicklung für unterschiedliche Produkte beurteilt werden. Am Beispiel der Selbstkostenentwicklung sei die Berechnung demonstriert. In einem Betrieb sind vier verschiedene Produkte mit unterschiedlichen Selbstkosten und in unterschiedlichen Mengen hergestellt worden; die entsprechenden Angaben sind in Tabelle 43 enthalten.

Produkt	Basiszeitraum		Berichtszeitraum	
	$x_m$ (in M)	$y_m$ (in Stück)	$x_i$ (in M)	$y_i$ (in Stück)
A	0,78	2000	0,72	2000
B	2,16	1400	2,03	1500
C	1,35	700	1,12	900
D	0,36	12000	0,33	15000

Tabelle 43 Selbstkosten ( $x$ ) und Mengen ( $y$ ) der im Betrieb A hergestellten Produkte

Die Berechnung des Selbstkostenindex auf der Grundlage der Mengen des Basiszeitraumes nach Formel (150) ergibt

$$I_p = \frac{0,72 \cdot 2000 + 2,03 \cdot 1400 + 1,12 \cdot 700 + 0,33 \cdot 12000}{0,78 \cdot 2000 + 2,16 \cdot 1400 + 1,35 \cdot 700 + 0,36 \cdot 12000} = \frac{9026}{9849} \cdot 100 = 91,64\%$$

Danach verminderten sich die Selbstkosten für die vier Erzeugnisse, unter der Annahme, dass auch im Berichtszeitraum die gleichen Mengen wie im Basiszeitraum hergestellt wurden, um 823 M bzw. 8,36%. Der absolute Betrag für die gesunkenen Selbstkosten folgt aus der Differenz zwischen dem Betrag des Zählers und Nenners, in unserem Beispiel zwischen 9026 und 9849.

Werden hingegen die Mengen des Berichtszeitraumes der Berechnung zugrunde gelegt, ergibt sich nach Formel (151)

$$I_p = \frac{0,72 \cdot 2000 + 2,03 \cdot 1500 + 1,12 \cdot 900 + 0,35 \cdot 15000}{0,78 \cdot 2000 + 2,16 \cdot 1500 + 1,35 \cdot 900 + 0,36 \cdot 15000} = \frac{10443}{11415} \cdot 100 = 91,48\%$$

Die Selbstkosten verminderten sich demnach im Berichtsjahr um 972 M bzw. 8,52% gegenüber dem Basisjahr unter der Annahme, dass im Basiszeitraum die gleichen Mengen hergestellt wurden wie im Berichtszeitraum.

Während bisher nur der Einfluss eines Faktors auf die relative Veränderung einer Erscheinung von Interesse war, ist häufig auch die Kenntnis des Einflusses aller Faktoren auf diese Veränderung erforderlich. In unserem Beispiel lautet die Fragestellung:

In welchem Maße veränderten sich die Gesamtselbstkosten des Betriebes A vom Basiszeitraum zum Berichtszeitraum?

Zur Berechnung wird der Volumenindex verwendet, dessen allgemeine Formel lautet

$$I_v = \frac{\sum_i x_i \cdot y_i}{\sum_m x_m \cdot y_m} \cdot 100(\%) \quad (152)$$

Für das gewählte Beispiel in Tabelle 43 ergibt sich der Volumenindex

$$I_v = \frac{0,72 \cdot 2000 + 2,03 \cdot 1500 + 1,12 \cdot 900 + 0,33 \cdot 15000}{0,78 \cdot 2000 + 2,16 \cdot 1400 + 1,35 \cdot 700 + 0,36 \cdot 12000} = \frac{10443}{9849} \cdot 100 = 106,03\%$$

Die Gesamtselbstkosten erhöhten sich demnach um

$$10443 - 9849 = 594 \text{ M bzw. } 6,03\%$$

Zur Beurteilung der relativen Veränderung von Durchschnittswerten, wie dem Durchschnittslohn, den durchschnittlichen Kosten oder dem durchschnittlichen Umsatz, dient der Durchschnittsindex. Allgemein kann er geschrieben werden als

$$I_s = \frac{\frac{\sum_i x_i \cdot y_i}{\sum_i y_i}}{\frac{\sum_m x_m \cdot y_m}{\sum_m y_m}} \cdot 100(\%) \quad (153)$$

wobei der Betrag des Zählers den Durchschnitt im Berichtszeitraum und der des Nenners den im Basiszeitraum angibt.

Ein Vergleich zwischen Zähler und Nenner gestattet deshalb auch bei diesem Index eine Aussage über die absolute Veränderung der jeweils betrachteten Durchschnitte. Am Beispiel der Entwicklung des Durchschnittslohnes in einem Betrieb soll die Berechnung der relativen Veränderung gezeigt werden.

Lohngruppe	Basiszeitraum		Berichtszeitraum	
	$x_m$ Br.lohn (in M)	$y_m$ Anz. d. Besch.	$x_i$ Br.lohn (in M)	$y_i$ Anz. d. Besch.
L	560	38	580	32
M	640	57	650	60
N	690	88	690	86

Tabelle 44 Anzahl der Beschäftigten in den einzelnen Lohngruppen eines Betriebes

Setzen wir die entsprechenden Werte der Tabelle 44 in Formel (153) ein, folgt

$$I_s = \frac{\frac{580 \cdot 32 + 650 \cdot 60 + 690 \cdot 86}{32 + 60 + 86}}{\frac{560 \cdot 38 + 640 \cdot 57 + 690 \cdot 88}{38 + 57 + 88}} = \frac{656,74}{647,43} \cdot 100 = 101,4\%$$

Es ist für den ausgewählten Betrieb festzustellen, dass der Durchschnittslohn für die untersuchten 3 Lohngruppen vom Basis- zum Berichtszeitraum um 9,31 M bzw. um

1,4% stieg.

Auch alle zusammengesetzten Indexzahlen können so gebildet werden, dass ihnen entweder eine feste oder variable Basis zugrunde liegt. Dies ist aber nur möglich, wenn die relative Veränderung über mehrere Zeitabschnitte betrachtet wird.

## 9.4 Korrelationsuntersuchungen bei Zeitreihen

Häufig interessiert bei ökonomischen, aber auch bei naturwissenschaftlichen und technisch-technologischen Untersuchungen die Abhängigkeit zwischen zeitlichen statistischen Reihen. Hierbei ist allerdings zu beachten, dass durch die Dynamik der Entwicklung, die den zu untersuchenden Zeitreihen zugrunde liegt, eine Abhängigkeit vorgetäuscht werden kann.

Werden deshalb Zeitreihen miteinander korreliert, ohne die der jeweiligen Zeitreihe zugrunde liegende Entwicklungstendenz auszuschalten, kann der Betrag des Korrelationskoeffizienten lediglich eine Bestätigung dafür sein, dass die untersuchten Zeitreihen sich gleich- oder gegenläufig entwickeln.

Eine derartige Aussage ist aber bereits auf Grund der Beobachtungswerte und ihrer grafischen Darstellung möglich. Voraussetzung der im Kapitel 5 behandelten Korrelations- und Regressionsanalyse war, dass die einzelnen Beobachtungswerte der untersuchten statistischen Reihen voneinander unabhängig sein müssen. Diese Bedingung wird von Zeitreihen auf Grund der logischen Verbundenheit der aufeinanderfolgenden Beobachtungswerte nicht erfüllt, d. h., sie sind autokorrelativ verbunden.

Daher muss, um der genannten Voraussetzung zu genügen, der Grad der Autokorrelation bei Zeitreihen festgestellt und durch entsprechende Verfahren vermindert bzw. ganz beseitigt werden.

Es sind jedoch auch einfachere Verfahren bekannt, die davon ausgehen, dass bereits die Ausschaltung der Grundrichtung der Entwicklung zu brauchbaren Wertepaaren für die Korrelationsrechnung führt. Zunächst können für die Abhängigkeitsuntersuchung an Stelle der Beobachtungswerte die Werte der absoluten oder relativen Veränderung zwischen aufeinanderfolgenden Zeitreihenwerten verwendet werden.

Wird die unabhängige Erscheinung mit  $x$  und die abhängige Erscheinung mit  $y$  bezeichnet, so ergeben sich die Reihen der Werte der absoluten Veränderung als

$$\Delta x_i = x_i - x_{i-1} \quad , \quad \Delta y_i = y_i - y_{i-1} \quad (154)$$

Sollen hingegen die relativen Veränderungen beider Reihen Gegenstand der Abhängigkeitsuntersuchung sein, müssen die relativen Veränderungen

$$p_{x_i} = \frac{x_i}{x_{i-1}} \quad , \quad p_{y_i} = \frac{y_i}{y_{i-1}} \quad (155)$$

bestimmt werden.

Alle diese Größen besitzen eine spezifische Aussage bei der Analyse zeitlicher Entwicklungsvorgänge. Welche der angegebenen Kennziffern im einzelnen der Korrelationsuntersuchung zugrunde gelegt wird, hängt im wesentlichen vom Ziel der Aussage und vom

Charakter der untersuchten Erscheinung ab.

So werden bei der Untersuchung ökonomischer Vorgänge vorwiegend nach den Formeln (155) bestimmte Kennziffern verwendet.

Für die lineare Regressionsfunktion

$$p_y = f(p_x) = a + bp_x \quad (156)$$

sind, von den relativen Veränderungen ausgehend, die Konstanten  $a$  und  $b$  zu bestimmen als

$$b = \frac{\sum_{i=1}^k p_x p_y - \bar{p}_y \sum_{i=1}^k p_x}{\sum_{i=1}^k p_x^2 - \bar{p}_x \sum_{i=1}^k p_x} \quad (157)$$

und

$$a = \bar{p}_x - b\bar{p}_y \quad (158)$$

Die Strenge des Zusammenhanges, ausgedrückt durch den Korrelationskoeffizienten, wird durch die Varianzen und die Kovarianz berechnet als

$$r = \frac{\sum_{i=1}^k p_x p_y - \bar{p}_x \sum_{i=1}^k p_y}{\sqrt{\left(\sum_{i=1}^k p_x^2 - \bar{p}_x \sum_{i=1}^k p_x\right) \left(\sum_{i=1}^k p_y^2 - \bar{p}_y \sum_{i=1}^k p_y\right)}} \quad (159)$$

Die drei zuletztgenannten Beziehungen sind uns bereits in ihrer Ursprungsform im Kapitel 5 begegnet, wobei dort die unabhängige Größe mit  $x$  und die von ihr abhängige Größe mit  $y$  bezeichnet wurden.

Hier nun stehen als Ursache die relativen Veränderungen der Größe  $x$  in Form der  $p_{x_i}$  und als Wirkung die relative Veränderung der Größe  $y$  in Form der  $p_{y_i}$ .

Als Beispiel soll die Abhängigkeit der relativen Veränderung des Warenumsatzes eines Handelsbetriebes von der relativen Veränderung des Einkommens pro Einwohner von 1960 bis 1969 bestimmt werden. In Tabelle 45 sind die zur Berechnung der Regressionsfunktion und des Korrelationskoeffizienten erforderlichen Werte enthalten.

Jahr	$p_x$	$p_y$	$p_x^2$	$p_y^2$	$p_{xy}$
1961	1,0530	1,0202	1,1088	1,0408	1,0743
1962	1,0371	1,0132	1,0756	1,0266	1,0508
1963	1,0188	1,0627	1,0380	1,1293	1,0827
1964	1,0404	1,0496	1,0824	1,1017	1,0920
1965	1,0255	1,0369	1,0517	1,0752	1,0633
1966	1,0550	1,0239	1,1130	1,0484	1,0802
1967	1,0577	1,0104	1,1187	1,0209	1,0687
1968	1,0592	1,0268	1,1219	1,0543	1,0876
1969	1,0528	1,0369	1,1084	1,0752	1,0916
Summe	9,3995	9,2806	9,8185	9,5724	9,6912

Tabelle 45 Korrelationsberechnung

$$\bar{p}_x = 1,0444, \bar{p}_y = 1,0312$$

Die Berechnung des Regressionskoeffizienten  $b$  ergibt nach Formel (157)

$$b = \frac{9,6912 - 1,0312 \cdot 9,3995}{9,8185 - 1,0444 \cdot 9,3995} = -0,94$$

und die des Koeffizienten  $a$  nach Formel (158)

$$a = 1,0312 - (-0,94)(1,0444) = 2,013$$

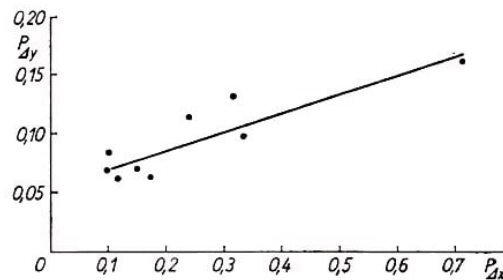


Abb. 45 Lineare Regressionsfunktion für die Abhängigkeit der relativen Veränderung des Warenumsatzes von der relativen Veränderung des Einkommens

Die lineare Regressionsfunktion (vgl. Abb. 45) für die Abhängigkeit der relativen Veränderung des Warenumsatzes des untersuchten Handelsbetriebes von der relativen Veränderung des Einkommens pro Einwohner für den Zeitraum von 1960 bis 1969 lautet dann

$$p_y = 2,013 - 0,94p_x$$

Der lineare Korrelationskoeffizient wird nach Beziehung (159) bestimmt.

$$r = \frac{9,612 - 1,0444 \cdot 9,2806}{\sqrt{(9,815 - 1,0444 \cdot 9,3995)(9,5724 - 1,0312 \cdot 9,2806)}} = -0,78$$

Ersichtlich ist zunächst, dass zwischen den relativen Veränderungen der beiden untersuchten Erscheinungen ein negativer Zusammenhang von  $r = -0,78$  besteht.

Die lineare Regressionsfunktion bringt zum Ausdruck, dass die relative Zunahme des Einkommens  $p_x$  eine relative Abnahme des Warenumsatzes  $p_y$  zur Folge hat, oder anders ausgedrückt, der Warenumsatz nimmt nicht in gleichem Maße wie das Einkommen zu.

Außer den Werten der absoluten und relativen Veränderung können auch die absoluten Veränderungen zwischen den aufeinanderfolgenden Werten der jeweiligen Zeitreihen ( $x_{i-1}$  und  $x_i$  sowie  $y_{i-1}$  und  $y_i$ ) bezogen auf den zutreffenden Ausgangswert ( $x_{i-1}$  bzw.  $y_{i-1}$ ) für die Regressions- und Korrelationsanalyse bei zeitlichen statistischen Reihen verwendet werden. Die Beziehungen lauten

$$p_{\Delta x_i} = \frac{x_i - x_{i-1}}{x_{i-1}} \quad \text{und} \quad p_{\Delta y_i} = \frac{y_i - y_{i-1}}{y_{i-1}} \quad (160,161)$$

Die Konstanten der Regressionsfunktion werden wie schon im vorangegangenen Fall nach den Formeln (157) und (158) und der Korrelationskoeffizient wird nach Formel (159) bestimmt.



Die bisher behandelten Verfahren gingen davon aus, dass durch Kennziffern der Entwicklung die vorhandene Autokorrelation in den Beobachtungswerten in gewissem Umfang eliminiert werden kann. In keinem Fall wurde dabei aber von der mathematisch bestimmten Ausgleichsfunktion abstrahiert.

Eine dritte Methode zur Korrelation zeitlicher statistischer Reihen berücksichtigt diesen Umstand, indem zunächst für jede der zu korrelierenden Reihen der entsprechende Trend bestimmt wird. Die sich aus den Abweichungen zwischen den Beobachtungs- und Trendwerten ergebenden Differenzen bilden dann die Grundlage der Korrelationsrechnung.

## 10 Die Darstellung der Ergebnisse

Zur Erfassung, Darstellung und Auswertung statistischen Zahlenmaterials verwendet man Tabellen und Grafiken. Auf diese Weise ist es relativ leicht und schnell möglich, einen Überblick über die untersuchten Massenerscheinungen zu gewinnen. Wenn es sich um die Darstellung von Ergebnissen mit klarer gesellschaftlicher Bezugnahme handelt, ist vor allem darauf zu achten, dass die politische Aussage in der Tabelle bzw. Grafik deutlich sichtbar wird.

In den voranstehenden Kapiteln haben wir schon eine große Anzahl Tabellen und Grafiken kennengelernt. Jetzt wollen wir uns kurz mit deren grundsätzlichem Aufbau beschäftigen und einige wesentliche Dinge zusammenfassen.

Statistische Tabellen stellen Zahlenübersichten dar, in denen die Kennziffern so geordnet sind, dass sie den Ablauf eines Prozesses, das Niveau einer Erscheinung oder Zusammenhänge zwischen den Erscheinungen auf kleinstem Raum sichtbar machen. Eine schematische Darstellung des Aufbaus einer Tabelle enthält Abbildung 46.

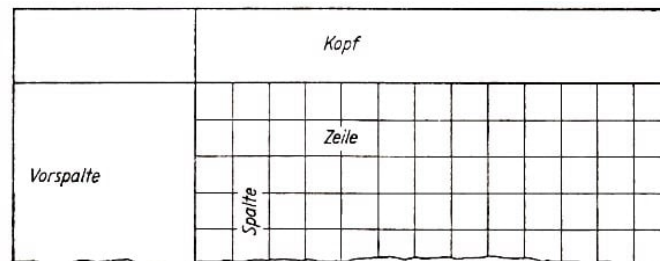


Abb. 46 Schema einer Tabelle

Jede Tabelle besteht aus

- der Bezeichnung
- dem Zahlenteil und
- dem Textteil.

Die Bezeichnung soll in kurzer Form den Inhalt der Tabelle zum Ausdruck bringen. Sie kann unter Umständen entfallen, wenn sie in den Text eingebaut ist und im Text entsprechende Hinweise auf die Tabelle gegeben werden. Der Zahlenteil dient zur Darstellung der Kennziffern. Im Textteil werden die Begriffsbestimmungen zum Zahlenteil aufgenommen.

Der Textteil gliedert sich wiederum in Vorspalte und Tabellenkopf. Die Vorspalte bezieht sich dabei auf den Zahleninhalt der Zeilen. Der Tabellenkopf enthält die Bezeichnungen der Spalten einschließlich der Vorspalte und kann nach mehreren Gruppierungsmerkmalen untergliedert werden.

Dabei hält man sich im allgemeinen an die Regel, dass waagerecht das dominierende Merkmal vermerkt wird und senkrecht die untergeordneten Merkmale eingetragen werden.

Oft ist es sinnvoll, die einzelnen Zeilen und Spalten durch laufende Nummern zu kennzeichnen. Im Zahlenteil einer Tabelle sollen die Kennziffern möglichst übersichtlich

dargestellt werden. Er wird nach Zeilen und Spalten untergliedert.

Nach ihrem Verwendungszweck werden Arbeits- und Ergebnistabellen unterschieden. Die Arbeitstabellen dienen dem Sammeln, Konzentrieren und Bearbeiten der Angaben über das Untersuchungsobjekt. Sie nehmen die Kennziffern aus dem Auswertungsprogramm auf.

Typische Arbeitstabellen sind z. B. die Tabellen 1 und 2, die Ausgangsdaten für die Untersuchung empirischer Häufigkeitsverteilungen enthalten. In den Ergebnistabellen wird das gewonnene Zahlenmaterial aus den Arbeitstabellen zur Analyse zusammengefasst. Die Mehrzahl der in diesem Buch enthaltenen Tabellen sind Ergebnistabellen.

Beim Erarbeiten der Tabellen ist immer darauf zu achten, dass die Tabellen leicht lesbar und übersichtlich gestaltet werden. Betrachten Sie aus dieser Sicht bitte die Tabellen im vorliegenden Buch.

Grafische Darstellungen werden in der Regel aus Tabellen abgeleitet. Sie sind im Vergleich zu den Tabellen anschaulicher, wenn es darum geht, einen allgemeinen Überblick über den Verlauf von Prozessen, über die Struktur von Erscheinungen usw. zu gewinnen.

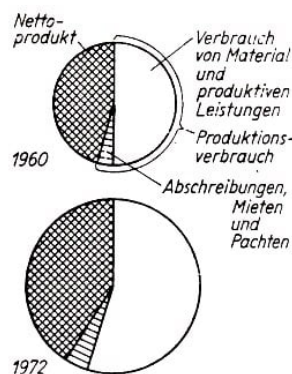


Abb. 47 Bruttonprodukt der DDR nach Wertbestandteilen in den Jahren 1960 1972 und 1972

Die grafischen Darstellungen weisen jedoch einen geringeren Genauigkeitsgrad als die Tabellen auf und können keine so große Zahl von Kombinationen aufnehmen. Tabellen und grafische Darstellungen sollen sich nicht gegenseitig ersetzen, sie müssen sich ergänzen. Auch für grafische Darstellungen sind in den bisherigen Abschnitten bereits zahlreiche Beispiele enthalten.

Wenn statistische Merkmalsvariationen oder Merkmalsgrößen miteinander verglichen bzw. einander gegenübergestellt werden sollen, zeichnet man diese oft als Flächen. Hierzu werden häufig geometrische Gebilde genutzt, vor allem Kreise und Rechtecke, wie sie die Abbildungen 47 und 48 zeigen.<sup>40</sup>

<sup>40</sup>Statistisches Jahrbuch der DDR 1973. Staatsverlag der DDR, Berlin 1973, S. 33 und 49

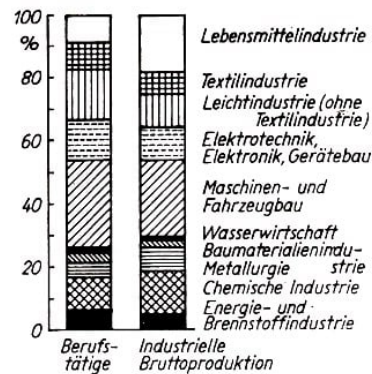


Abb. 48 Berufstätige der Industrie und industrielle Bruttoproduktion in der DDR nach Bereichen im Jahre 1971

Zur besseren optischen Anschaulichkeit werden die abgebildeten Flächen in den meisten Fällen schraffiert oder farbig gestaltet.

Das ist besonders dann zweckmäßig, wenn mehrere Flächen unmittelbar nebeneinander stehen.

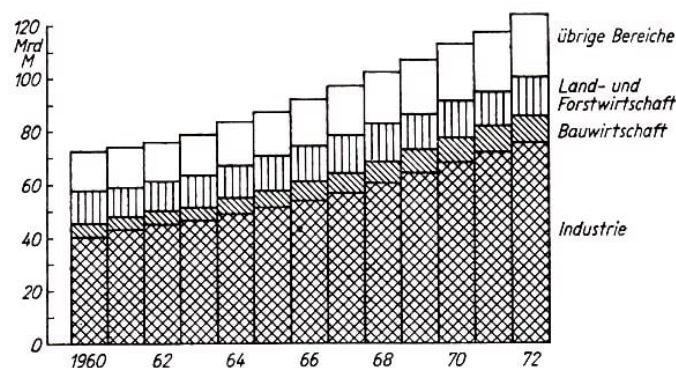
In vielen Fällen wird dabei zugleich versucht, mit der Farbe bzw. der Schraffur eine Wertigkeit für die betreffenden Merkmalsvariationen zum Ausdruck zu bringen. Werden z.B. die Bevölkerungsdichten verschiedener Territorien in einer grafischen Darstellung einander gegenübergestellt, so werden die Gebiete mit einer großen Bevölkerungsdichte dunkler gezeichnet als die mit der geringeren Bevölkerungsdichte.

Dabei ist stets darauf zu achten, dass die Bedeutung der verwendeten Schraffuren, Farben und Symbole erläutert wird.

Statistische Ergebnisse werden grafisch vor allem in Form von Diagrammen dargestellt. Wir kennen zwei Arten von Diagrammen:

- Darstellung durch geometrische Gebilde und Symbole (Punkte, Striche, Kreise, Rechtecke)
- Darstellung durch Kurven.

Die Form des verwendeten Diagramms ist von den darzustellenden Kennziffern abhängig. Die Abbildung 49 zeigt z. B. die Entwicklung des Nettoproduktes der DDR vom Jahr 1960 bis zum Jahr 1972 als Balkendiagramm.<sup>41</sup> Die Höhe des Nettoproduktes wird durch die Länge des jeweiligen Balkens ausgedrückt.



<sup>41</sup>Ebenda, S. 33

Abb. 49 Nettoprodukt der DDR nach Wirtschaftsbereichen von 1960 bis 1972

Das Gegenüberstellen der senkrecht bzw. waagerecht angeordneten Balken ermöglicht einen Vergleich der Größen. Häufig ist man bestrebt, in einem Diagramm nicht nur das Niveau einer Erscheinung zum Ausdruck zu bringen, sondern gleichermaßen die Struktur der Untersuchungsobjekte und die bestehenden Niveauunterschiede. In Abbildung 49 ist deshalb das Nettoprodukt der DDR nach Wirtschaftsbereichen untergliedert. Ein Balkendiagramm liegt auch in Abbildung 48 vor.

Die in Abbildung 47 vorgenommene Aufgliederung des Bruttoproduktes nach Wertbestandteilen ist ebenfalls ein Beispiel für ein Diagramm mittels geometrischer Gebilde. Der Kreis wird als Darstellungsform bevorzugt, wenn nicht das absolute Niveau, sondern die Struktur des Untersuchungsobjektes von Interesse ist.

Der Kreis entspricht dann 100%, während die Größe der Kreisabschnitte den prozentualen Anteil der jeweiligen Gruppe angibt.

Bei der Kurvendarstellung gibt es zwei grundlegende Arten von Kurven:

- Häufigkeitskurven und
- Entwicklungskurven.

Häufigkeitskurven stellen die Verteilung von Einheiten auf Klassen oder Gruppen dar. Dabei werden in der Regel auf der Abszisse die Klassen und auf der Ordinate die Häufigkeiten der jeweiligen Klassen aufgetragen (vgl. Abb. 50).

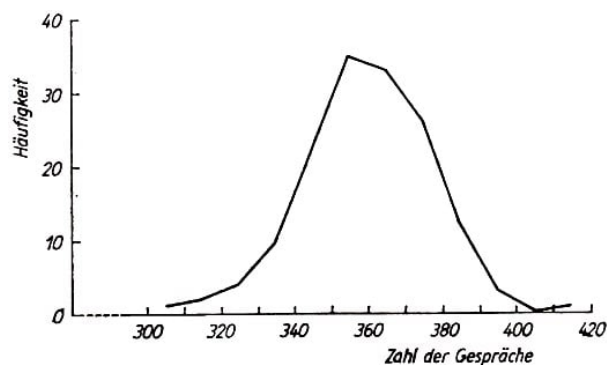


Abb. 50 Zahl der täglich in einer Fernsprechzentrale ankommenden Ferngespräche

Häufigkeitskurven wurden auch bereits im Kapitel 4 bei der Erläuterung von empirischen Verteilungen benutzt, so in den Abbildungen 8 bis 17. Eine besondere Form der Häufigkeitskurve ist die Summenkurve, die kumulative Größen beinhaltet. Hierfür ist ein Beispiel in Abbildung 7 gegeben.

Die Entwicklungskurven dienen der grafischen Darstellung der zeitlichen Veränderung des Niveaus einer oder mehrerer Erscheinungen. Die Abszisse enthält bei den Entwicklungskurven im allgemeinen die Zeitangaben, während auf der Ordinate die Größen für das Niveau der Erscheinungen aufgetragen werden. Typische Entwicklungskurven treten zum Beispiel bei der Kennzeichnung des Trends auf (vgl. Abb. 34 bis 45). Ein weiteres Beispiel für eine Entwicklungskurve zeigt Abbildung 51.<sup>42</sup>

<sup>42</sup>Ebenda, S. 337

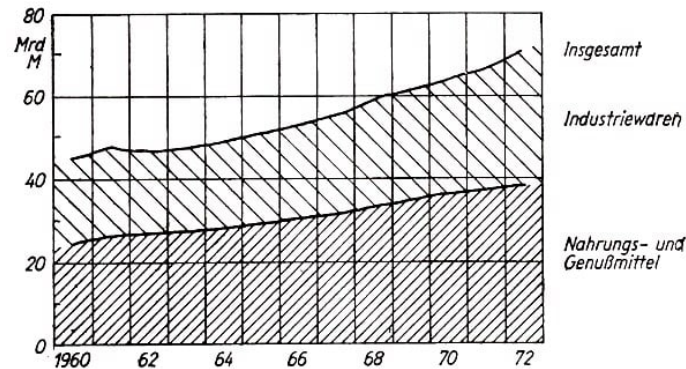


Abb. 51 Einzelhandelsumsatz in der DDR von 1960 bis 1972

Bei der Kurvendarstellung ist immer darauf zu achten, dass dann, wenn in das Diagramm mehrere Kurven einzuzeichnen sind, diese Kurven voneinander abgegrenzt werden. In der Regel geschieht das durch unterschiedliche Linien (vgl. Abb. 53).

Für Abszisse und Ordinate muss der richtige Maßstab gewählt werden, damit die Kurve die typischen Zusammenhänge bzw. die typischen Merkmale der Erscheinungen zeigt. Streng genommen müsste jedes Diagramm vom Nullpunkt des Maßstabes ausgehen. Das würde aber in den Grafiken oft zu großen freien Räumen führen und den Platz für den typischen Kurvenabschnitt einschränken, worunter die Anschaulichkeit litte. Deshalb wird oft auf die Markierung des Nullpunktes verzichtet und nur ein Ausschnitt gezeichnet (vgl. Abb. 50 und 54). Dieser Abschnitt muss aber deutlich gekennzeichnet werden.

Mitunter ist es auch sinnvoll, einen Maßstab zu unterbrechen (vgl. Abb. 14). Das ist vor allem bei Entwicklungskurven der Fall, wenn nicht alle Daten bekannt sind oder ein Vergleich mit einem relativ lange zurückliegenden Zeitpunkt durchgeführt oder ein Zusammenhang besonders deutlich hervorgehoben werden soll.

Für die grafische Darstellung statistischer Ergebnisse existieren zahlreiche Möglichkeiten. Einige wurden hier angedeutet. Entsprechend der praktischen Aufgabenstellung sind die jeweils zweckmäßigsten Darstellungsformen auszuwählen, wobei auch die verschiedenen Formen in geeigneter Weise verändert oder miteinander kombiniert werden können.

## 11 Was sagt eine statistische Analyse aus?

An einer statistischen Erhebung interessiert natürlich vor allem ihre Aussage. In Abhängigkeit von der gegebenen Aufgabenstellung ist zu analysieren, welche Schlussfolgerungen die ermittelten Daten für das untersuchte Problem zulassen.

Wenn wir den Inhalt dieser Formulierung genau durchdenken, stellen wir fest, dass die Analyse in unseren bisherigen Betrachtungen bereits eine wesentliche Rolle spielte. Wir haben die Untersuchungen und Berechnungen nicht abstrakt durchgeführt, sondern immer auf ihre Aussage Wert gelegt. Wir haben in nahezu allen Abschnitten des Buches Analysen vorgenommen und dabei bereits wichtige Prinzipien der Analyse kennengelernt.

Wegen der Bedeutung, die der Analyse bei statistischen Untersuchungen beizumessen ist, soll an dieser Stelle auf einige Schwerpunkte noch einmal zusammenfassend eingegangen werden.

Die Aussage einer statistischen Analyse wird in diesem Abschnitt am Beispiel wirtschaftsstatistischer Analysen erläutert.

Die zu behandelnden Probleme haben jedoch analog auch im naturwissenschaftlichen und technischen Bereich Gültigkeit.

Dabei darf aber keinesfalls verkannt werden, dass bei wirtschaftsstatistischen Betrachtungen die direkte politische und gesellschaftliche Bezugnahme in einem weitaus stärkeren Maß gegeben ist als bei naturwissenschaftlichen bzw. technischen Aufgabenstellungen. Das muss beim Erarbeiten einer Analyse unbedingt beachtet und in ihr zum Ausdruck gebracht werden.

Die statistische Analyse geht von den vorliegenden Zahlenwerten aus, wobei zunächst die Quantität der Erscheinungen analysiert wird. Aus der Quantität von Massenerscheinungen lassen sich aber Schlussfolgerungen über die Qualität ableiten. Das Erlangen von Kenntnissen über die qualitative Seite ist eine wesentliche Aufgabe der Analysetätigkeit.

Dabei wird aber die Qualität der Massenerscheinungen nicht durch die Statistik definiert. Das obliegt anderen Wissenschaften.

Soll z. B. die erreichte Arbeitsproduktivität ermittelt und analysiert werden, definiert die Statistik nicht den Begriff Arbeitsproduktivität, sondern sie hat in diesem Fall vor allem die Aufgabe, die Arbeitsproduktivität exakt zu erfassen und die gewonnenen Daten statistisch auszuwerten.

In diesem Zusammenhang nimmt die Statistik jedoch Einfluss auf die statistische Abgrenzung der betreffenden Kennziffer.

Es muss auch darauf hingewiesen werden, dass aus einer Analyse nicht mehr herausgelesen werden kann und darf, als die Ausgangsdaten zulassen. Deshalb ist der Datenerfassung und der Datenaufbereitung eine große Bedeutung beizumessen, da deren Genauigkeit einen entscheidenden Einfluss auf die Aussage der Analyse hat.

Die Formulierungen in der Analyse müssen äußerst exakt sein, damit sie nicht zu Missverständnissen und zu falschen Schlüssen führen. Zu betonen ist auch, dass für die

inhaltliche Aussage einer Analyse immer der betreffende Fachwissenschaftler verantwortlich ist. Die Statistik entbindet den Fachwissenschaftler nicht von der Verantwortung für den untersuchten Komplex.

Für die statistische Analyse sind exakte Kennziffern eine wichtige Voraussetzung. Dabei ist zu beachten, dass eine Kennziffer allein noch keine große Aussagekraft besitzt, weshalb sie mit anderen Kennziffern verglichen werden muss.

Uns ist bekannt, dass bei Untersuchungsobjekten zwischen sachlichen, zeitlichen und örtlichen Merkmalen unterschieden wird. Bei statistischen Analysen variiert immer nur eines dieser drei Merkmale. Es wird also eine Untersuchung hinsichtlich des sachlichen Merkmals oder hinsichtlich des zeitlichen Merkmals oder hinsichtlich des örtlichen Merkmals vorgenommen.<sup>43</sup>

Wenn sachlich unterschiedliche Untersuchungsobjekte vorliegen, existieren zwei voneinander verschiedene gesellschaftliche Erscheinungen. So sind z. B. die Bevölkerung der DDR und die Fläche der DDR im Jahre 1974 zwei sachlich unterschiedliche Erscheinungen, die sich ihrem Wesen nach unterscheiden.

Sie gleichen sich aber bezüglich des Erfassungszeitraumes (zeitliches Merkmal) und des Erfassungsgebietes (örtliches Merkmal). Wir merken uns, dass ungleichartige Erscheinungen, die sich also in sachlichen Merkmalen unterscheiden, einander gegenübergestellt werden; in unserem Beispiel die Bevölkerung der DDR und die Fläche der DDR im Jahre 1974.

Werden gesellschaftliche Erscheinungen mit gleichen sachlichen Merkmalen hinsichtlich zeitlicher oder örtlicher Merkmale untersucht, dann erfolgt keine Gegenüberstellung, sondern ein Vergleich. Gleichartige Erscheinungen werden also miteinander verglichen.

In der Wirtschaftsstatistik werden drei große Gruppen von Vergleichen unterschieden:

- der Plan - Ist - Vergleich
- der dynamische Vergleich
- der statische Vergleich.

Der Plan-Ist-Vergleich dient zur Plankontrolle. Mit seiner Hilfe werden die Zusammenhänge zwischen Plan und Planerfüllung untersucht. Eine Zielsetzung, die Planzahl, wird mit dem erreichten Ergebnis verglichen. Beim Plan-Ist-Vergleich stimmen die verglichenen Kennziffern hinsichtlich ihrer sachlichen, zeitlichen und örtlichen Abgrenzung überein. Er nimmt somit eine gewisse Sonderstellung ein.

Der dynamische Vergleich gibt Aufschluss über zeitliche Unterschiede eines Untersuchungsobjektes. Ein dynamischer Vergleich liegt vor, wenn die Energieerzeugung mehrerer Jahre untereinander oder die Entwicklung der Arbeitsproduktivität eines Betriebes für mehrere Zeiträume verglichen werden.

Es wird auch von einem dynamischen Vergleich gesprochen, wenn die Aufgliederung eines Jahres in Monate erfolgt und beispielsweise die industrielle Bruttoproduktion oder der Krankheitsstand oder die durchschnittliche Niederschlagsmenge der einzelnen Mo-

---

<sup>43</sup>Vgl. Donda, Herrde, Kuhn, Struck: Statistik. Verlag Die Wirtschaft, Berlin 1972, S. 61



nate miteinander verglichen werden.

Der statische Vergleich zeigt örtliche Unterschiede auf. Beim statischen Vergleich wird z.B. die Energieerzeugung eines Jahres verschiedener Länder miteinander verglichen oder die erreichte Arbeitsproduktivität der Betriebe eines Industriezweiges während eines bestimmten Zeitraumes untersucht.

Ein statischer Vergleich liegt z. B. auch vor, wenn die industrielle Bruttoproduktion eines Jahres der Bezirke der DDR miteinander verglichen wird oder eine Aufgliederung nach Industriezweigen erfolgt. Für den statischen Vergleich wird oft der Begriff Niveauvergleich gebraucht.

Die Analyse untersucht die bestehenden Zusammenhänge. Dabei ist das Untersuchungsobjekt als Teil des Ganzen zu sehen, wobei alle wesentlichen Einflussgrößen zu beachten sind. Bei der Analyse der Planerfüllung eines Jahres z. B. müssen Vergleiche mit den zurückliegenden Planjahren vorgenommen werden.

Es müssen die Auswirkungen solcher Einflussgrößen, wie veränderte Technologie, neue Technik, Materialzuführung, Materialqualität usw., eingeschätzt werden. Darüber hinaus sind auch bereits die Folgen für die nächste Planperiode zu beurteilen und Vergleiche mit anderen Betrieben vorzunehmen.

Durch die statistische Analyse wird nicht nur die Entwicklung einer gesellschaftlichen Erscheinung und ihre Beziehung zu anderen gesellschaftlichen Erscheinungen aufgezeigt, die Analyse soll vor allem der Ermittlung der Ursachen dienen, die eine gesellschaftliche Erscheinung entscheidend beeinflusst haben.

Wird z. B. der Plan eines Betriebes in einem Planabschnitt nicht erfüllt, so sind durch die Analyse die Ursachen hierfür zu ermitteln. Es ist aber auch einzuschätzen, welche Auswirkungen sich daraus für den eigenen Betrieb und für andere Betriebe bzw. Industriezweige ergeben können.

Für den Aufbau einer statistischen Analyse lassen sich folgende wesentliche Faktoren angeben:<sup>44</sup>

- Die statistische Analyse geht von den Beschlüssen der SED und der Regierung der DDR aus. Dadurch können gesellschaftliche Erscheinungen ökonomisch und politisch richtig eingeschätzt werden.

- Durch die Analyse ist das Untersuchungsobjekt - ausgehend von der gegebenen Zielstellung möglichst vielseitig einzuschätzen.

Jede Analyse muss ein für die Leitung aussagekräftiges Ergebnis zum Inhalt haben. Sie darf nicht nur die Entwicklung aufzeigen, sondern muss auch die Ursachen erforschen und Maßnahmen für die Veränderung vorschlagen.

Die Analyse erfordert deshalb gründliche Kenntnisse über das Untersuchungsobjekt und über die statistischen Forschungsmethoden.

- Eine statistische Analyse beginnt in den meisten Fällen mit dem Feststellen allgemeingültiger Fakten, in denen das Wesentliche über das Untersuchungsobjekt ausgesagt wird. Die Analyse über die Erfüllung des Planes eines Betriebes wird von der Gesamter-

---

<sup>44</sup>Vgl. ebenda, S. 64ff.

füllung des Planes ausgehen und dann die Erfüllung der einzelnen Planteile untersuchen, um daraus die Ursachen für den erreichten Stand der Planerfüllung zu erkennen und Schlussfolgerungen ableiten zu können.

- Es folgen Untersuchungen über einzelne Seiten und Züge des Untersuchungsobjektes. Durch sie werden bestimmte spezielle Fragen beantwortet. Zum Beispiel kann der Einfluss der Steigerung der Arbeitsproduktivität auf die Erfüllung des Gewinnplanes oder die Erfüllung des Planes Wissenschaft und Technik auf die Erhöhung der Arbeitsproduktivität eingeschätzt werden.

- Zur Ermittlung wesentlicher Ursachen bestimmter Veränderungen werden die Einflüsse einzelner Seiten besonders herausgearbeitet. Das kann z. B. der Einfluss der Neuererarbeit auf die Planerfüllung sein.

- Es ist oft sinnvoll, die Untersuchungen durch typische Beispiele zu ergänzen, die besonders gute oder schlechte Einzelercheinungen aufzeigen.

- Die Analyse darf sich nicht auf einzelne Seiten des Untersuchungsobjektes beschränken, sondern muss die Gesamtheit der Elemente und ihr Zusammenwirken zum Inhalt haben. Diese Aufgabe ist häufig schwierig zu lösen, vor allem weil der direkte Einfluss einer Größe auf die Gesamtveränderung nicht immer konkret nachweisbar ist.

So ist es beispielsweise schwierig, bei der Analyse der Planerfüllung die direkte Auswirkung verbesserter Arbeits- und Lebensbedingungen einzuschätzen.

- In einer Analyse müssen auch Fakten berücksichtigt werden, die sich zahlenmäßig nicht exakt erfassen lassen, obwohl sie eine quantitative Seite haben. Ein solcher Faktor kann die Qualifikation der Arbeitskräfte und ihr Einfluss auf die Planerfüllung sein.

- Die Ergebnisse einer statistischen Analyse werden in schriftlicher Form zusammengefasst. Inhalt und Aufbau der Analyse werden in einer Disposition dargestellt. Der Textteil beginnt mit einer kurzen Angabe über die Ziele der Analyse.

Es schließen sich die Ergebnisse der Analyse an. Die Textanalyse fasst die gewonnenen Erkenntnisse zusammen und enthält Vorschläge für einzuleitende Maßnahmen. Sie muss auch auf die Quellen des verwendeten Zahlenmaterials hinweisen.

Der Umfang der Textanalyse ist vom Zweck abhängig, für den die Analyse benötigt wird. Soll nur ein grober Überblick über die wesentlichen Seiten einer Erscheinung gegeben werden, ist die Analyse darauf abzustimmen. Werden tiefgehende, detaillierte Kenntnisse über eine Erscheinung benötigt, muss die Analyse auch auf die Details eingehen.

- Eine statistische Analyse enthält Tabellen und Grafiken. Diese können unmittelbar in den Text aufgenommen werden oder einen besonderen Teil der Analyse bilden bzw. einen selbständigen Anlagenband darstellen. Mitunter ist es günstig, die Tabellen nur teilweise mit stark gerundeten Zahlen in die Analyse aufzunehmen.

Der Text soll die Zahlen der Tabellen nicht wiederholen, sondern ihren Inhalt erläutern und daraus Schlussfolgerungen ableiten. Die Grafiken müssen das Verständnis der analytischen Darlegungen und der Tabellen erleichtern.

In der ökonomischen Statistik werden gegenwärtig Untersuchungen durchgeführt, um eine standardisierte Analyse nach dem Baukastensystem zu entwickeln. Die Ergebnisse der Untersuchungen sind abzuwarten, bevor eine Aussage über die praktische Bedeutung dieser Form der Analyse gemacht werden kann.

Nach Abschluss der Auswertung einer statistischen Erhebung liegen Tabellen, Grafiken und eine Textanalyse vor. In ihnen sind die Kennziffern und die daraus abgeleiteten Erkenntnisse konzentriert und anschaulich dargestellt.

Im einzelnen wird das Vorgehen vom Zweck der Analyse beeinflusst und dieser ist bei der Vielzahl gesellschaftlicher Erscheinungen sehr unterschiedlich. Ebenso differenziert muss demzufolge beim Erarbeiten der Analysen vorgegangen werden. Dabei ist immer zu beachten, dass die Statistik niemals Selbstzweck werden darf, sondern immer als Hilfsmittel zum Erkenntnisgewinn anzusehen ist.

## 12 Literaturhinweise

Für das weitere Eindringen interessierter Leser in die Statistik und deren wichtigste Anwendungen empfehlen die Autoren folgende ausgewählte Publikationen:

Adam, J.: Einführung in die Biostatistik, Reaktionskinetik und EDV. VEB Verlag Volk und Gesundheit, Berlin 1972.

Autorenkollektiv: Mathematik in Übersichten. Volk und Wissen - Volkseigener Verlag, Berlin 1973.

Autorenkollektiv: Wahrscheinlichkeitsrechnung und mathematische Statistik. Akademie-Verlag, Berlin 1970.

Clauß, G., und Ebner, H.: Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Volk und Wissen - Volkseigener Verlag, Berlin 1967.

Donda, A.; Herrde, E.; Kuhn, O.; Struck, R.: Statistik-Lehrbuch. Verlag Die Wirtschaft, Berlin 1972.

Fabian, V.: Statistische Methoden. VEB Deutscher Verlag der Wissenschaften, Berlin 1970.

Fischer, P.; Richter, K.-J.; Schneider, H.: Statistische Methoden für Verkehrsuntersuchungen. Transpress VEB Verlag für Verkehrswesen, Berlin 1974.

Flachsmeier, J.: Kombinatorik. VEB Deutscher Verlag der Wissenschaften, Berlin 1972.

Förster, E., und Egermayer, F.: Korrelations- und Regressionsanalyse. Verlag Die Wirtschaft, Berlin 1966.

Forbrig, G., und Janakieff, R.: Grundriss der Industriestatistik, Band I und II. Verlag Die Wirtschaft, Berlin 1967.

Gnedenko, B. W., und Chintschin, A. I.: Elementare Einführung in die Wahrscheinlichkeitsrechnung. VEB Verlag der Wissenschaften, Berlin 1964.

Grenander, U.: Einführung in das Studium der mathematischen Statistik. VEB Fachbuchverlag, Leipzig 1969.

Hasse, M.: Grundbegriffe der Mengenlehre und Logik. B. G. Teubner Verlagsgesellschaft, Leipzig 1970.

Kompanejev, A.S.: Statistische Gesetze in der Physik. B. G. Teubner Verlagsgesellschaft, Leipzig 1972.

Ludwig, R.: Methoden der Fehler- und Ausgleichsrechnung. VEB Deutscher Verlag der Wissenschaften, Berlin 1969.

Maibaum, G.: Wahrscheinlichkeitsrechnung. Volk und Wissen - Volkseigener Verlag, Berlin 1971.

Pawlowski, Z.: Einführung in die mathematische Statistik. Verlag Die Wirtschaft, Berlin 1971.

Rasch, D.: Elementare Einführung in die mathematische Statistik. VEB Deutscher Verlag der Wissenschaften, Berlin 1968.

Rasch, D.; Enderlein, G.; Herrendörfer, G.: Biometrie-Verfahren, Tabellen, angewandte Statistik. VEB Deutscher Landwirtschaftsverlag, Berlin 1973.

Schindowski, E., und Schürz, O.: Statistische Qualitätskontrolle. VEB Verlag Technik, Berlin 1972.

Storm, R.: Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitäts-

kontrolle. VEB Fachbuchverlag, Leipzig 1972.

Vineze, I.: Mathematische Statistik mit industriellen Anwendungen. Akademiai Kiado, Budapest 1971.

Wilenskin, W. J.: Unterhaltsame Mengenlehre. B. G. Teubner Verlagsgesellschaft, Leipzig 1972.

## 13 Kleines Lexikon der Statistik

Damit der Leser sich schnell über einige ausgewählte grundlegende Begriffe der Statistik informieren kann, wird nachstehend deren wesentlicher Inhalt einfach und allgemeinverständlich erläutert.

Abweichung, mittlere quadratische - siehe Standardabweichung

Algorithmus - Rechenvorschrift; exakt formulierte Anweisung zur Lösung einer Klasse von Aufgaben mit einer endlichen Anzahl von Ablaufschritten

Analyse, statistische - Beurteilung der erfassten und aufbereiteten Daten über die statistischen Untersuchungsobjekte. In Abhängigkeit von der Zielstellung muss die Analyse Aufschluss über die untersuchte Massenerscheinung sowie über deren kausales Verhalten zu anderen Untersuchungsobjekten geben. Die statistische Analyse ist ein Hilfsmittel der Leitungstätigkeit. Sie dient der Aufdeckung von Ursachen, die die Erscheinungen oder Prozesse in Natur und Gesellschaft hervorgerufen bzw. beeinflusst haben. Bei statistischen Analysen spielt neben anderen statistischen Methoden der Vergleich eine wesentliche Rolle.

Assoziativkoeffizient - Koeffizient, der als Maß der Assoziation, d. h. der in einer Vierfeldertafel dargestellten Abhängigkeit alternativer Merkmale, dient. Allen A., die sich durch verschiedene Berechnungsformeln unterscheiden, ist gemeinsam, dass ihre Werte zwischen -1 und +1 liegen.

Ausgangsdaten - Zahlenmaterial aus vorhandenen Unterlagen oder spezielle Mess- und Beobachtungswerte, die als Voraussetzung zur Berechnung aufzubereiten sind

Beobachtungswert (empirischer Wert) - durch Zählung oder Messung des zu untersuchenden Merkmals eines Untersuchungsgegenstandes gewonnene Angabe

Bestimmtheitsmaß - ein Maß zur Bestimmung der Stärke des Zusammenhangs zwischen Zufallsvariablen, das eine seiner Dimension nach quadratische Kennziffer darstellt. Es setzt sich zusammen aus der Streuung der Einzelwerte von  $y$  um die Funktionswerte und der Streuung der Funktionswerte um den arithmetischen Mittelwert. Der Wert des B. liegt zwischen 0 und +1. Das B. ist ein Ausdruck für die Bestimmtheit einer Regression, da es den Anteil der Regressionsstreuung an der Gesamtstreuung fixiert.

Binomialverteilung - siehe Verteilungen, theoretische Datenverarbeitung - Erfassen, Aufbereiten, Berechnen, Weiterleiten, Auswerten und Aufbewahren von Daten in dauernder Wiederholung

Dichtefunktion - Funktion  $f(x)$  stetiger Zufallsvariablen; für jedes beliebige Intervall  $(x_1, x_2)$  ist dabei die Wahrscheinlichkeit, dass die Zufallsvariable einen Wert aus diesem Intervall annimmt, gleich dem Flächenstück zwischen dem Abszissenabschnitt  $(x_1, x_2)$  und dem über diesem Abschnitt liegenden Stück des Kurvenbildes. Die Dichtefunktion bestimmt eine Verteilung eindeutig.

Dichtemittel (Gipfelwert, Modalwert, Mode) - das D. einer Verteilung ist - als eine

Maßzahl der Lage - der Wert  $x_i$ , der in der Stichprobe mit der größten Häufigkeit auftritt.

Bei eingipfligen, symmetrischen Verteilungen stimmt das Dichtemittel mit dem Zentralwert überein.

Bei linksseitig schiefen Verteilungen ist das Dichtemittel größer als der Zentralwert und bei rechtsseitig schiefen Verteilungen kleiner als der Zentralwert. Bei mehrgipfligen Verteilungen liegen mehrere Dichtemittel vor.

Dispersion - siehe Streuung

Durchschnitt - das allgemeine, alle einzelnen Größen unter einer bestimmten Zielstellung charakterisierende Niveau, das die Beziehungen zweier Erscheinungen ausdrückt und zahlenmäßig durch die Gegenüberstellung der Gesamtgrößen der beiden Erscheinungen charakterisiert wird (siehe auch Mittel)

Elastizität - Reaktionsempfindlichkeit oder Nachgiebigkeit einer als Wirkung anzusprechenden Erscheinung  $y$  gegenüber einer Ursache  $x$ . Die Messung der E. kann je nach der Aufgabenstellung durch die absolute oder relative Elastizitätsfunktion erfolgen.

Element, statistisches - kleinste statistische Einheit der statistischen Gesamtheit. Das E. ist durch sachliche, örtliche und zeitliche Merkmale, die quantitative oder qualitative Eigenschaften zum Ausdruck bringen, charakterisiert.

Ereignis, zufälliges - Ergebnis eines Versuches, dessen Eintreten vom Zufall abhängig ist

Erwartungswert - Mittelwert einer Zufallsvariablen oder einer Verteilung

Exponentialverteilung - siehe Verteilungen, theoretische

Extrapolation - die Weiterführung bisheriger Entwicklungstendenzen zwischen den Zeitpunkten  $t_0$  und  $t_n$ , die durch Ausgleichsfunktionen (Trendfunktionen) zum Ausdruck gebracht werden, über das zeitliche Intervall hinaus. Die hierbei getroffene Annahme der gleichbleibenden Entwicklungstendenz- auch in der Zukunft- ist aber in der Mehrzahl der praktischen Fälle nicht begründet.

Exzess einer Verteilung (Kurtosis) - Eigenschaft einer Verteilung im Vergleich mit der Normalverteilung. Ein positiver Exzess liegt vor, wenn eine empirische Verteilung steiler ist als die Normalverteilung; von einem negativen Exzess spricht man, wenn die Verteilung flacher ist. Der Exzess unterscheidet sich von der Schiefe einer Verteilung (siehe Schiefeitsmaß und Verteilung, schiefe).

Freiheitsgrad - die Anzahl der voneinander unabhängigen Beobachtungen bzw. die Anzahl der unabhängigen Summanden auf einer Seite einer Bestimmungsgleichung, wenn die andere Seite der Bestimmungsgleichung konstant gehalten wird

Gipfelwert - siehe Dichtemittel

Grundgesamtheit - Menge aller möglichen Realisationen einer Zufallsvariablen

Gruppenbildung - im statistischen Sinn die Untergliederung eines Untersuchungsobjektes in Gruppen, die sich qualitativ oder graduell voneinander unterscheiden, oder die Zusammenführung gleichartiger statistischer Elemente zu Gruppen. Die Gruppenbildung leistet z. B. in der ökonomischen Statistik einen hervorragenden Beitrag zur Erfüllung der Aufgabe der Statistik, wesentliche Seiten und Züge der untersuchten Massenerscheinungen widerzuspiegeln und die objektive Realität erkennen zu helfen. Auf der Grundlage der gebildeten Gruppen werden dann jeweils bestimmte statistische Methoden angewendet.

Häufigkeit, absolute - die Anzahl der Beobachtungen in einer bestimmten Klasse

Häufigkeit, relative - der relative Anteil der auf eine bestimmte Klasse entfallenden Beobachtungen, also im Verhältnis zur Gesamtzahl an Beobachtungen gesehen

Häufigkeitsfunktion - zusammenfassender Ausdruck für die Wahrscheinlichkeitsfunktion und die Dichtefunktion (siehe dort)

Häufigkeitspolygon - grafische Darstellung für eine Häufigkeitsverteilung. Die Häufigkeiten der Klassen werden auf der Abszissenachse aufgetragen und durch Gerade miteinander verbunden. Der auf diese Weise entstandene Kurvenzug gibt eine Näherung für das Bild der Dichtefunktion.

Häufigkeitstabelle - siehe Verteilungstafel, sekundäre

Häufigkeitsverteilung - die Angaben der Klassen mit ihren Häufigkeiten für eine Verteilung. Sie wird meist in der Form einer Tabelle oder eines Häufigkeitspolygons bzw. eines Staffebildes dargestellt.

Histogramm - siehe Staffebild

Hypothese, statistische - Annahme über die unbekannten Parameter bzw. über statistische Gesetzmäßigkeiten einer Grundgesamtheit

Indexzahl - statistische Maßzahl für die Messung der relativen Veränderungen eines Merkmals der zu untersuchenden Erscheinung von Zeitabschnitt zu Zeitabschnitt. Unterschieden werden einfache und zusammengesetzte I.

Interpolation - die näherungsweise Bestimmung des Wertes einer Funktion  $f(x)$  an der Stelle  $x$  aus den Funktionswerten. Es handelt sich dabei um eine Näherungsrechnung zur Bestimmung von Zwischengliedern, z. B. bei einer Regression oder einer statistischen Reihe.

Irrtumswahrscheinlichkeit - siehe Sicherheit, statistische

Klasse - mehrere aufeinanderfolgende Beobachtungswerte einer Beobachtungsreihe; eine Einteilungsmöglichkeit für die Beobachtungswerte

Klassenbreite (Klassenintervall) - die bei der Auswertung von Beobachtungswerten festgelegten Abstände zwischen der unteren und der oberen Klassengrenze. Die K. ergibt sich näherungsweise als Quotient aus der Differenz zwischen dem größten und dem



kleinsten Beobachtungswert dividiert durch die Anzahl der Klassen.

Klassengrenze - die Endpunkte einer Klasse, wobei zwischen der oberen und der unteren K. unterschieden wird

Klassenhäufigkeit - die Häufigkeit der Beobachtungswerte einer Klasse, wobei zwischen absoluter und relativer Häufigkeit unterschieden wird (siehe dort)

Klassenintervall - siehe Klassenbreite

Klassenmitte - das arithmetische Mittel aus dem Betrag der unteren und der oberen Klassengrenze

Klassennummer - die Festlegung von Nummern für die Klassen entsprechend ihrer Größe, und zwar entweder beginnend mit 1 und fortlaufend weiter nummeriert bis zur letzten Klasse, oder man verwendet als Klassennummern die Klassenmitte bzw. einen dieser Klassenmitte entsprechenden kodierten Wert. Im zuletzt erwähnten Falle dienen diese Werte zur vereinfachten Berechnung von Maßzahlen, wie Mittelwerte, der Standardabweichung u. a.

Konfidenzintervall - siehe Vertrauensintervall

Kontingenzkoeffizient - Korrelationsmaß für qualitative Merkmale, und zwar das Maß für den Grad der Abhängigkeit zwischen den Merkmalen einer Kontingenztafel (siehe dort). Der Wert des K. liegt zwischen 0 und +1.

Kontingenztafel - eine  $m$ - $n$ -Tafel zur Ermittlung der Kontingenz zwischen qualitativen Merkmalen. Die K. ist quadratisch und enthält Häufigkeiten.

Korrelation - im weiteren Sinne der Zusammenhang zwischen qualitativen oder quantitativen Merkmalen eines statistischen Untersuchungsgegenstandes und im engeren Sinn nur zwischen quantitativen Merkmalen.

Man spricht beim Vorliegen von zwei Variablen von einer positiven oder direkten K., wenn gleichzeitig die Werte der beiden Veränderlichen zunehmen oder sich vermindern. Weil die beiden Variablen sich in gleicher Richtung bewegen, wird die positive Korrelation auch als Gleichläufigkeit oder Parallelismus bezeichnet.

Eine negative oder indirekte K. ist gegeben, wenn die eine Zufallsvariable in die entgegengesetzte Richtung als die andere strebt. Man bezeichnet diese Erscheinung als Gegenläufigkeit oder Antagonismus.

Korrelation, einfache - liegt vor, wenn der Zusammenhang zwischen zwei Zufallsvariablen untersucht wird.

Korrelation, lineare - siehe Regression, lineare

Korrelation, mehrfache (multiple) - liegt vor, wenn der Zusammenhang zwischen drei und mehr Zufallsvariablen analysiert wird.

Korrelation, nichtlineare - siehe Regression, nichtlineare

Korrelation, partielle - Zusammenhang zwischen zwei Zufallsvariablen unter der Vor-

aussetzung, dass weitere Variablen vorliegen, aber die Abhängigkeit von diesen zum Zwecke der Analyse eliminiert wurde. Die Ordnung der p. K. wird bestimmt nach der Zahl der eliminierten Variablen. Für die p. K. wird ebenfalls ein Korrelationskoeffizient berechnet (siehe auch Regression, partielle).

Korrelationskoeffizient - die am häufigsten gebrauchte Maßzahl der Korrelation ; das Maß für die Strenge (Stärke) eines Zusammenhangs, einer Korrelation. Der K. nimmt Werte zwischen -1 und +1 an.

Korrelationsrechnung (Korrelationsanalyse, Korrelationsstatistik) - Untersuchung, ob eine Korrelation, ein Zusammenhang zwischen Zufallsvariablen, vorhanden ist, und Bestimmung des Grades dieses Zusammenhangs, der mit Hilfe des Korrelationskoeffizienten (siehe dort) gemessen wird

Kovarianz - die Produktsumme aus den Abweichungen der Werte von zwei statistischen Verteilungen oder Reihen. Sie bildet beim linearen Regressions- bzw. Korrelationskoeffizienten den Zähler.

Los (Produktionslos, Partie) - eine Anzahl konstruktiv gleichartiger Einzelteile, die in einem Arbeitsgang in zusammenhängender Folge unter Gewährung einer einmaligen Vorbereitungs- und Abschlusszeit gefertigt werden

Maßzahlen, statistische - charakteristische Kenngrößen messbarer Merkmale. Sie vertreten die Messwerte der Urliste bzw. der Häufigkeitstabelle und dienen zur Beschreibung und zum Vergleich von Häufigkeitsverteilungen. Wichtige statistische Maßzahlen eines messbaren Merkmals sind Mittelwerte und die Varianz; wichtige statistische Maßzahlen zweidimensionaler Verteilungen sind die Kovarianz und der Korrelationskoeffizient.

Median - siehe Zentralwert

Menge - Zusammenfassung von Einzelobjekten zu einem einheitlichen Ganzen

Merkmal, statistisches - kennzeichnet die für eine Untersuchung wesentliche Eigenschaft der einzelnen Elemente der statistischen Gesamtheit. Merkmale werden nach verschiedenen Gesichtspunkten untergliedert, z. B. in quantitative und qualitative Merkmale bzw. in sachliche, örtliche und zeitliche Merkmale.

Merkmalsvariation - konkrete Erscheinungsform des zu untersuchenden Merkmals jedes einzelnen statistischen Elements.

Messwert - siehe Beobachtungswert

Mittel ( Durchschnitt, Mittelwert) - Maßzahl einer statistischen Verteilung oder statistischen Reihe, die ein mittleres Niveau widerspiegelt. Es gibt mehrere Mittel; als Beispiele siehe Mittel, arithmetisches, geometrisches, gewogenes, harmonisches und Dichtemittel.

Mittel, arithmetisches - wichtigste statistische Maßzahl zur Charakterisierung der Lage einer Verteilung oder einer Stichprobe. Das a. M. einer Zufallsvariablen mit der Verteilungsfunktion  $F(x)$  bzw. der Dichtefunktion  $f(x)$  ist der Erwartungswert. Das a. M.

ist vor allem durch zwei Eigenschaften gekennzeichnet:

- Die Summe der Abweichungen der Einzelwerte vom a. M. ist gleich Null.
- Die Summe der Abweichungsquadrate ist beim a. M. ein Minimum im Vergleich mit den anderen Mittelwerten.

Handelt es sich bei den  $n$  Werten um die Beobachtungswerte einer Stichprobe, so entspricht das a. M. dem Stichprobenmittel.

Mittel, geometrisches -  $n$ -te Wurzel aus dem Produkt einer Zahlenfolge von  $n$  positiven Zahlen. Es wird vor allem zur Berechnung der durchschnittlichen relativen Veränderung benutzt.

Mittel, harmonisches - der Reziprokwert des arithmetischen Mittels bezogen auf die reziproken Werte der  $x_1$ . Das harmonische Mittel ist kleiner oder gleich dem arithmetischen Mittel.

Mittelwert - siehe Mittel

Modalwert - siehe Dichtemittel

Mode - siehe Dichtemittel

Nonsensekorrelation - Zusammenhang zwischen Zufallsvariablen, der sinnlos ist. Es besteht kein kausaler Zusammenhang.

Normalverteilung - siehe Verteilung, theoretische Parameter, statistische-charakteristische Größen einer statistischen Verteilung

Partie - siehe Los

Poisson-Verteilung - siehe Verteilung, theoretische

Produktionslos - siehe Los

Prüfverfahren - Verfahren der mathematischen Statistik zur Prüfung der durch Stichproben gelieferten Informationen über Häufigkeitsverteilungen und ihre Parameter in der Grundgesamtheit

Prüfverteilung - Stichprobenverteilung von Maßzahlen, die primär als Prüfwerte von statistischen Tests angewendet werden. Ihre Kenntnis ist vor allem für die statistischen Prüfverfahren von Bedeutung. Wichtige Prüfverteilungen sind z. B. die  $\chi^2$ -Verteilung, die  $t$ -Verteilung und die  $F$ -Verteilung.

Qualitätskontrolle, statistische - Gesamtheit aller Maßnahmen, mittels mathematisch-statistischer Methoden der Analyse von Prüfergebnissen und der Ursachenforschung auf die Qualitätsfestlegung und Qualitätsfertigung einzuwirken. Es werden zwei grundlegende Arten der statistischen Qualitätskontrolle unterschieden: die laufende Kontrolle des Fertigungsprozesses mit mathematischen Methoden (Kontrollkartentechnik) und die Überprüfung bereits gefertigter Erzeugnisse in der Annahme-, Zwischen- und Endkontrolle durch Stichprobenpläne.

Rangkorrelation - Zusammenhang zwischen Urteilen u. ä., in Rangplätzen zum Aus-

druck kommende Einschätzung bzw. die auf der Grundlage von Rangzahlen durchgeführte, nichtparametrische Korrelationsrechnung. Als Maß der Übereinstimmung wird der Rangkorrelationskoeffizient (siehe dort) ermittelt.

Rangkorrelationskoeffizient - Maßzahl für die Übereinstimmung von zwei Rangordnungen. Der R. wird benutzt, wenn keine Beobachtungswerte ermittelt, sondern lediglich Rangplätze festgelegt werden können.

Regression - Bezeichnung für den mit Hilfe einer Regressionsfunktion beschriebenen Zusammenhang zwischen Zufallsvariablen, wobei die Werte einer Zufallsvariablen als gegeben betrachtet werden.

Regression, einfache - siehe Korrelation, einfache

Regression, lineare - Charakterisierung des Zusammenhangs beim Bestehen linearer Beziehungen zwischen den in die Untersuchung einbezogenen Variablen

Regression, mehrfache - siehe Korrelation, mehrfache

Regression, nichtlineare - Charakterisierung des Zusammenhangs beim Bestehen nichtlinearer Beziehungen zwischen den in die Untersuchung einbezogenen Variablen

Regression, partielle - eine Teilregression innerhalb der mehrfachen Regression (siehe analog hierzu auch Korrelation, partielle). Die p. R. unterscheidet sich von der totalen R., bei der nur die einfache Regression zwischen zwei Variablen betrachtet wird, dadurch, dass drei oder mehr Variablen vorliegen und die bei der p. R. nicht betrachteten Variablen eliminiert werden.

Regressionsfunktion (Regressionsgleichung) - mathematische Funktion zur Wiedergabe der Regression von  $y$  auf  $x$ , d. h. der stochastischen Abhängigkeit der Variablen  $y$  von der Variablen  $x$ . Sie wird meist mit der Methode der kleinsten Quadrate berechnet, wobei zuvor ein Funktionstyp zu bestimmen ist, der dem Inhalt des untersuchten Zusammenhangs bestmöglich Rechnung trägt.

Regressionskoeffizient - Parameter der Regressionsfunktion. Der R. wird mit der Methode der kleinsten Quadratsumme bzw. mit der Maximum-Likelihood-Methode auf der Grundlage der Beobachtungswerte ermittelt.

Regressionsrechnung (Regressionsanalyse, Regressionsstatistik) - Untersuchungen zur Bestimmung einer Ausgleichsfunktion für die Zufallsvariablen  $X$  und  $Y$ , wobei die Ausgleichsfunktion der Bedingung genügen muss, dass die Summe der Quadrate der Abweichungen der empirischen Werte von den Funktionswerten ein Minimum darstellt.

Reihe, statistische - Menge der Beobachtungswerte. Die statistische R. ist der Ausgangspunkt für die Anwendung mathematisch-statistischer Methoden. Analog den statistischen Merkmalen sind örtliche, sachliche und zeitliche statistische R. zu unterscheiden.

Schätzverfahren, statistische - Verfahren, deren Aufgabe es ist, die Größe der unbekannten Parameter der Grundgesamtheit anhand der Maßzahlen von Stichproben zu schätzen. Es wird zwischen der Punkt- und der Intervallschätzung unterschieden. Die

Punktschätzung dient dazu, für einen Parameter einen ganz bestimmten Wert festzulegen.

Bei der Intervallschätzung wird ein aus Stichprobendaten berechnetes Intervall angegeben, das den gesuchten Parameter mit einer bestimmten Wahrscheinlichkeit einschließt.

Scheinkorrelation (mittelbare Korrelation) - Zusammenhang zwischen Zufallsvariablen, wobei weder die eine noch die andere Zufallsvariable die Ursache bzw. die Wirkung darstellt, sondern eine dritte Erscheinung die Ursache ist, von der beide Zufallsvariablen abhängen. Es besteht also kein kausaler Zusammenhang.

Schiefheitsmaß - Maß, mit dem die Asymmetrie einer Verteilung bestimmt wird. Der höchste Wert beträgt bei kontinuierlichen Verteilungen 1. Es gibt verschiedene Schiefheitsmaße. Die Schiefe unterscheidet sich vom Exzess einer Verteilung (siehe dort).

Schwankung, periodische - bei der Untersuchung von Entwicklungsprozessen regelmäßig wiederkehrende Schwankungen, die auf sich wiederholende unterschiedliche Einflussintensität der die Erscheinung hervorrufenden Ursachen zurückzuführen ist

Sicherheit, statistische - Wahrscheinlichkeit dafür, dass ein Wert im dazugehörigen Vertrauensintervall liegt. Die statistische Sicherheit ergänzt sich mit der Irrtumswahrscheinlichkeit (Wahrscheinlichkeit dafür, dass ein Wert außerhalb des Vertrauensintervalls liegt) zu 1 bzw. zu 100%.

Spannweite (Variationsbreite, Variationsweite) - Maßzahl der Streuung, Differenz zwischen  $x_{\max}$  und  $x_{\min}$  einer Verteilung

Staffelbild (Histogramm) - grafische Darstellung von Beobachtungswerten, das entsteht, indem über jedem Wert oder über jeder Klasse auf der Abszissenachse ein Rechteck errichtet wird. Die Fläche oder die Höhe dieser Rechtecke sind entweder der absoluten bzw. relativen Häufigkeit der Beobachtungswerte oder der Klassenhäufigkeit proportional.

Standardabweichung (mittlere quadratische Abweichung) - Maßzahl der Streuung der Beobachtungswerte um den Mittelwert, die als Quadratwurzel aus der Streuung (siehe dort) gebildet wird

Statistik, mathematische - Zweig der angewandten Mathematik, der ein vorwiegend auf der Wahrscheinlichkeitsrechnung beruhendes Instrumentarium zur Verfügung stellt, um Massenerscheinungen in Natur, Gesellschaft und Technik zu untersuchen und wissenschaftlich zu beurteilen. Diese vor allem der Auswertung statistischer Daten dienenden Methoden sind ein besonders wirksames Hilfsmittel in den Naturwissenschaften und in der Technik, um Gesetzmäßigkeiten erkennen zu können.

Stichprobe, statistische - Teilmenge, deren Elemente durch Zufallsauswahl aus einer Grundgesamtheit ausgewählt werden und diese repräsentieren sollen

Stichprobenfehler - maximale Abweichung zwischen den Maßzahlen der Stichprobe und den entsprechenden Parametern der Grundgesamtheit, die unter den gegebenen Bedingungen mit einer bestimmten Wahrscheinlichkeit erwartet werden darf. Der Stichpro-

benfehler ist berechenbar.

Streubereich - der Bereich der Beobachtungswerte von  $x_{\min}$  bis  $x_{\max}$ .

Streuung (Varianz, Dispersion, quadratische Streuung) - Maßzahl für die Messung der Streuung von Beobachtungswerten um ihren Mittelwert. Für eine empirische Verteilung oder eine statistische Reihe wird die S. ermittelt als Quotient der Summe der Abweichungsquadrate für jeden Beobachtungswert bezogen auf den Mittelwert dieser Verteilung und dividiert durch die Anzahl der Beobachtungswerte.

Streuungsdiagramm - grafische Darstellung der Punkteschar einer empirischen Verteilung in einem Koordinatensystem. Das S. gestattet Schlussfolgerungen über den Typ der Regressionsfunktion.

Strichliste - siehe Verteilungstafel, primäre

Summenhäufigkeit, relative - die auf die Gesamtzahl der Beobachtungswerte, die gleich 100% gesetzt wird, bezogenen kumulativen Häufigkeiten. Die grafische Darstellung für die r. S. ergibt die Summenkurve.

Summenkurve - siehe Summenhäufigkeit, relative

Testverfahren, statistische - siehe Prüfverfahren, statistische

Trend - Grundrichtung der Entwicklung. Bei Zeitreihenuntersuchungen wird der T. durch die den jeweiligen Entwicklungsprozess widerspiegelnde Trendfunktion ausgedrückt.

Trendfunktion - Ausgleichsfunktion, die die Entwicklungsrichtung einer statistischen Zeitreihe widerspiegelt und der Forderung genügt, dass die Summe der Abweichungsquadrate zwischen Beobachtungs- und Funktionswerten ein Minimum ist

Urliste - erste schriftliche Zusammenstellung von Beobachtungswerten entsprechend der Reihenfolge ihrer Gewinnung. Die Werte sind hier noch ungeordnet.

Variabilitätskoeffizient - siehe Variationskoeffizient

Variable, diskrete (Zufallsveränderliche, diskrete) - die sich bei statistischen Untersuchungen ergebenden Zufallsgrößen, die nicht jeden Wert innerhalb eines bestimmten Intervalls annehmen können, z. B. wenn die Werte nur ganzzahlig sein dürfen

Variable, stetige (Zufallsveränderliche, stetige) - die sich bei statistischen Untersuchungen ergebenden Zufallsgrößen, die innerhalb eines bestimmten Intervalls, innerhalb einer gewissen Spanne, jeden Wert annehmen können

Varianz - siehe Streuung

Variationsbreite (-weite) - siehe Spannweite

Variationskoeffizient (Variabilitätskoeffizient) - Vergleichsmaß, das in der Form einer dimensionslosen Größe die Standardabweichung zum arithmetischen Mittel ins Verhältnis setzt, wobei der sich ergebende Wert in Prozent ausgedrückt wird

Verteilung, bivariable - zweidimensionale Verteilung. Es werden zwei Variablen analysiert.

Verteilung, eingipflige - eine empirische Verteilung mit einem Gipfel

Verteilung, empirische - die Verteilung der in der Praxis tatsächlich gemessenen oder gezählten Werte, d.h. der Realisationen der Zufallsvariablen  $X$

Verteilung, flache - Verteilung mit negativem Exzess (siehe dort)

Verteilung, gemischte - eine Verteilung, die sich aus mehreren, jeweils mit einem bestimmten Gewicht versehenen Verteilungen zusammensetzt, so dass ihre Verteilungsfunktion lautet:  $F(x) = p_1 F_1(x) + \dots + p_n F_n(x)$ . Für eine gemischte empirische Verteilung gilt analog, dass die Elemente mit der Wahrscheinlichkeit  $p_1$  aus der ersten Verteilung, mit der Wahrscheinlichkeit  $p_2$  aus der zweiten Verteilung und mit der Wahrscheinlichkeit  $p_n$  aus der  $n$ -ten Verteilung stammen können.

Verteilung, hypergeometrische - siehe Verteilungen, theoretische

Verteilung, mehrgipflige - eine empirische Verteilung, die mehrere Dichtemittel hat und die sich bei ihrer grafischen Darstellung durch mehrere Gipfel auszeichnet. Oft handelt es sich hierbei um sogenannte gemischte Verteilungen.

Verteilung, monovariabel - eindimensionale Verteilung. Es wird eine Variable analysiert.

Verteilung, schiefe - eine asymmetrische Verteilung mit einem sich von Null unterscheidenden Schiefheitsmaß (siehe dort). Schiefe Verteilungen sind dadurch charakterisiert, dass das arithmetische Mittel und der Zentralwert sich zahlenmäßig unterscheiden. Eine große Differenz zwischen Mittelwert und Zentralwert deutet auf eine erhebliche Schiefe hin.

Verteilung, steile - Verteilung mit positivem Exzess (siehe dort)

Verteilung, theoretische - eine der Verteilungsfunktion einer Zufallsgröße adäquate Verteilung. Sie gibt die Aufteilung der Wahrscheinlichkeitssumme 1 auf einzelne Merkmalsvariationen entsprechend bestimmten Ausgangsbedingungen in einer mathematischen Funktion an. Dazu gehören u. a. die Binomialverteilung, die Poisson-Verteilung, die hypergeometrische Verteilung, die Normalverteilung, die Exponentialverteilung und die Prüfverteilungen.

Verteilung, zweigipflige - eine empirische Verteilung, die sich bei ihrer grafischen Darstellung durch zwei Maxima auszeichnet

Verteilungsfunktion - Funktion, die die Wahrscheinlichkeit dafür angibt, dass die Zufallsvariable  $X$  einen Wert annimmt, der kleiner als der Wert  $x$  ist;  $F(x) = P(X < x)$ . Durch die Verteilungsfunktion wird eine Verteilung eindeutig bestimmt.

Verteilungstafel, primäre - Aufstellung der der Größe nach geordneten beobachteten Werte einer empirischen Verteilung in der Form einer Tabelle. Sie ist oft mit der Strichliste identisch, d. h., in der primären Verteilungstafel wird für jeden beobachteten Wert

in der Spalte oder Zeile, die seiner Größe entspricht, ein Strich angebracht.

Verteilungstafel, sekundäre ( Häufigkeitstabelle) - Zusammenfassung der beobachteten Werte einer empirischen Verteilung in Größenklassen entsprechend ihrer Häufigkeit in einer Tabelle

Vertrauensintervall - bei Stichprobenuntersuchungen ermitteltes Intervall, in dem mit einer bestimmten Wahrscheinlichkeit (der statistischen Sicherheit) der Wert eines Parameters in der Grundgesamtheit zu erwarten ist. Es wird auch als Konfidenzintervall bezeichnet.

Wahrscheinlichkeitsdichte - die Dichte einer stetigen Zufallsgröße  $X$ , ausgedrückt durch die Dichtefunktion  $f(x)$ , wobei stets  $f(x) \geq 0$  gilt

Wahrscheinlichkeitsfunktion - Funktion  $f(x) = p_k$ , die bei diskreten Verteilungen die Wahrscheinlichkeiten der einzelnen möglichen Werte der Zufallsvariablen angibt

Wahrscheinlichkeitsverteilung - Die W. gibt an, nach welcher Gesetzmäßigkeit die möglichen Werte einer Zufallsvariablen angenommen werden.

Wirtschaftsstatistik - Anwendung statistischer Methoden zur Untersuchung von wirtschaftlichen Massenerscheinungen. Grundlagen der sozialistischen W. sind die politische Ökonomie des Sozialismus und der dialektische und historische Materialismus.

Zeitreihe - Ordnung der Beobachtungswerte verschiedener Zeitpunkte oder Zeitintervalle entsprechend der zeitlichen Reihenfolge. Wesentlichste Komponenten einer Z. sind der Trend und die periodischen Schwankungen.

Zeitreihenanalyse - Zerlegung einer Zeitreihe in ihre Komponenten und deren Quantifizierung mit Hilfe statistischer Methoden

Zentralwert (Median) - der mittelste Wert einer der Größe nach geordneten Zahlenreihe. Der Z. einer Verteilung wird bei einer geraden

Zahl von Messwerten als das arithmetische Mittel aus den beiden in der Mitte liegenden Werten der geordneten Beobachtungswerte gebildet, bei einer ungeraden Zahl von Messwerten ist es der mittelste Wert. Es gibt also ebenso viele größere wie kleinere Beobachtungswerte.

Zufallsauswahl - die Auswahl der Elemente für die Stichprobe erfolgt zufällig. Jedes Element der Grundgesamtheit muss die gleiche Möglichkeit haben, in die Stichprobe zu gelangen. Die Zufallsauswahl ist eine Voraussetzung der Stichprobentheorie. Sie wird durch unterschiedliche Auswahltechniken realisiert.

Zufallsgröße - siehe Zufallsvariable

Zufallsvariable - auch Zufallsgröße genannt, ist eine Variable, die Werte in bestimmten vorgegebenen Intervallen der Zahlengeraden mit bestimmten Wahrscheinlichkeiten annimmt. Die angenommenen Werte heißen Realisationen der Zufallsvariablen. Eine Z. heißt diskret, wenn es nur endlich oder abzählbar unendlich viele Realisationen gibt; eine



Z. heißt stetig, wenn sie in einem bestimmten Intervall jeden beliebigen Wert annehmen kann.