Carl von Ossietzky
Universität
Oldenburg

Risikomanagement für Finanzdienstleister (M.Sc.)

Quantitative Methoden

Christiane Goodfellow

Center für
lebenslanges
Lernen

Berufsbegleitender Masterstudiengang

Risikomanagement für Finanzdienstleister (M.Sc.)

Center für lebenslanges Lernen

Christiane Goodfellow

Quantitative Methoden

Carl von Ossietzky Universität Oldenburg, 2021

Impressum

Autorin: Prof. Dr. Christiane Goodfellow

Herausgeber: Carl von Ossietzky Universität Oldenburg - Center für lebenslanges Lernen C3L

Auflage: 7. Auflage 2021 (Erstauflage 2015)

Oldenburg, März 2021

Prof. Dr. Christiane Goodfellow

Akademischer Werdegang

- 1999-2001 Wissenschaftliche Mitarbeiterin bei der Deutschen Bundesbank, Frankfurt am Main
- 2001-2003 Analystin bei der Bank of England, London
- 2003-2004 Doktorandin an der Universität Ulm
- 2005-2006 Doktorandin an der Europa-Universität Viadrina, Frankfurt (Oder)
- 2006-2008 Doktorandin an der Universität Münster
- April 2008: Promotion zum Dr. rer. pol. mit einer Arbeit zur empirischen Kapitalmarktforschung
- 2008-2011 Elternzeit für zwei Kinder (Universität Münster)
- 2012-2017 Professur für Allgemeine BWL, insbesondere Versicherungs-, Bank- und Kreditwirtschaft an der Jade Hochschule, Studienort Wilhelmshaven
- Seit 2017 Professur für Allgemeine BWL und Statistik an der Jade Hochschule, Studienort Wilhelmshaven

Schwerpunkte in Forschung und Lehre

In meinem ersten Jahr als Doktorandin habe ich mich an der Universität Ulm mit statistischen Methoden zur Vorhersage von Unternehmensinsolvenzen beschäftigt und gemeinsam mit zwei Koautoren ein Bayesianisches Verfahren hierauf angewendet. Im Anschluss daran habe ich die Handelsqualität an einer Börse von einer anonymen elektronischen Handelsplattform (bspw. Xetra) mit einem nicht-anonymen Parketthandel wiederum mit statistischen Verfahren verglichen und daraus Handelsempfehlungen für Privatanleger abgeleitet. Unterschiede im Investorenverhalten zwischen Privatanlegern und institutionellen Investoren haben mich seither beschäftigt. Neuere Publikationen fallen in den Bereich der Markteffizienz und Behavioural Finance. Die aktuelle Liste meiner Veröffentlichungen steht auf
https://sites.google.com/site/christianegoodfellow/

Da mein Forschungsschwerpunkt die empirische Kapitalmarktforschung ist, liegt es nahe, mich auch in der Lehre auf Statistik und Ökonometrie zu konzentrieren. An der Jade Hochschule biete ich in den Bachelor-Studiengängen „Wirtschaft“ und „Tourismuswirtschaft“ sowohl die Grundlagenveranstaltung „Statistik“ als auch die Fortgeschrittenenvorlesung „Angewandte Statistik und Ökonometrie“ an.

Einführung in die Quantitativen Methoden

Das Modul „Quantitative Methoden“ gibt Ihnen das Handwerkszeug mit, auf das Sie in nahezu allen anderen Modulen Ihres Studiums zurückgreifen werden. Wie wichtig dieses Modul ist, zeigt ein Blick auf das Regelwerk zur Bankenaufsicht „Basel III“ (weitgehend wörtlich entnommen aus Goodfellow und Salm (2012)):

Ursprünglich umfasste das Regelwerk zur Bankenaufsicht lediglich die Anforderung, risikobehaftete Aktiva zu 8 % mit Eigenkapital zu unterlegen, wobei Hypotheken nur hälftig angerechnet wurden. Im Bankgeschäft sind diese risikobehafteten Aktiva vor allem Kundenkredite. Die Eigenkapital-Unterlegung ist für Banken teuer, da Eigenkapital zugleich Haftungskapital ist und die Eigenkapitalgeber eine Vergütung dieses Risikos einfordern. Wenn die Bank unabhängig vom Risiko – und damit auch unabhängig vom Ertragspotential – der Aktiva Eigenkapital vorhalten muss, dann lohnen sich aus betriebswirtschaftlicher Sicht vor allem solche Aktivgeschäfte, die mit hohen Ertragschancen ausgestattet sind. In aller Regel werden dies jedoch riskante Geschäfte sein.

Diese Anreizproblematik heißt Regulierungsarbitrage und wohnt jedem bankaufsichtlichen Regelwerk inne, das riskante und weniger riskante Geschäfte mit gleichen Eigenkapitalanforderungen „bestraft“. Tatsächlich sollten riskante Geschäfte höhere Eigenkapitalanforderungen haben als weniger riskante: Banken, die auf der Aktivseite eine sichere Geschäftsstrategie verfolgen, sollten mit geringen Eigenkapitalanforderungen belohnt werden. Dieser Gedanke wird in Basel II umgesetzt. Außerdem unterliegen operationelle und Marktpreisrisiken nunmehr ebenfalls aufsichtsrechtlichen Regeln.

In dem Regelwerk Basel II werden Marktpreisrisiken mit dem Value-at-Risk-Ansatz erfasst. Der Wert einer (Aktiv-)Position ist danach eine Zufallsvariable, die einer statistischen Verteilung folgt. Wir werden darauf im Laufe der Veranstaltung zurückkommen. Ausfallrisiken werden entweder mit externen Ratings quantifiziert oder über ein internes Verfahren ermittelt, das die Bankenaufsicht genehmigen muss. Ratings werden mit statistischen Verfahren erstellt.

Gerade als dieses Basel-II-Regelwerk 2007 in der Europäischen Union rechtskräftig wurde, gerieten Geschäftsbanken in den USA in Folge der Subprime-Immobilienkrise in Schwierigkeiten. Vor dem Hintergrund der globalen Finanzkrise, die sich 2008 zu einer Wirtschaftskrise ausweitete, legte der Basler Ausschuss für Bankenaufsicht im Dezember 2010 ein wiederum überarbeitetes Regelwerk vor. Dieses Basel-III-Rahmenwerk zur Stärkung der globalen Eigenkapital- und Liquiditätsvorschriften besteht aus zwei Veröffentlichungen des Basler Ausschuss für Bankenaufsicht („Basel III: Ein globaler Regulierungsrahmen für widerstandsfähigere Banken und Bankensysteme“ (bcbs189) und „Basel III: Internationale Rahmenvereinbarung über Messung, Standards und Überwachung in Bezug auf das Liquiditätsrisiko“ (bcbs188)) und wird nun schrittweise umgesetzt. Ziel der Vorschriften ist es, mit strengeren Regeln für Eigenkapital und Liquidität die Widerstandsfähigkeit des Bankensektors gegenüber Stress-Situationen im Finanzsektor und in der Wirtschaft zu verbessern und zukünftig die Gefahr von Banken- und Finanzkrisen sowie deren Auswirkungen auf die Realwirtschaft zu verringern.

Die besondere Bedeutung statistischer Verfahren in der Bankenaufsicht und deshalb auch im Risikomanagement wird deutlich, wenn Sie sich beispielsweise im oben angesprochenen Basel-III-Dokument bcbs 189 die Absätze 97 bis 99 anschauen: Eine Bank muss außerbörsliche Derivate nicht nur für das Ausfallrisiko des Kontrahenten mit Eigenkapital unterlegen, sondern auch für das Marktpreisrisiko durch Bonitätsverschlechterung des Kontrahenten. Dieses Verlustrisiko durch die Verschlechterung der Kontrahentenbonität wird mit einem „credit value adjustment (CVA)" erfasst, das wiederum auf dem Value-at-Risk-Ansatz beruht. Noch nie beinhaltete ein bankaufsichtliches Regelwerk so viele komplexe Formeln und Berechnungsmethoden wie Basel III. Dieses Modul bereitet Sie darauf vor, die quantitativen aufsichtsrechtlichen Anforderungen zu verstehen und deren Umsetzung selbstständig in Ihren Instituten voranzutreiben.

Im Wesentlichen gab es zwei Konstellationen, die 2007/08 bei Banken zu Schieflagen geführt haben: Erstens haben US-amerikanische Geschäftsbanken zu leichtfertig, zu billig und in zu hohem Umfang Baufinanzierungen vergeben, die zu erheblichen Teilen nicht mehr bedient werden konnten (Ausfallrisiko). Zweitens haben sich Fristentransformation und die Verflechtungen auf dem Interbankenmarkt, insbesondere durch Verbriefungen, als erhebliches Liquiditätsrisiko erwiesen. Als sich abzeichnete, dass Banken durch Kreditausfälle insolvent werden, mochten sie sich gegenseitig kein Geld mehr leihen, so dass der Interbankenmarkt zusammenbrach. Wegen der Fristentransformation war aber gerade der Interbankenmarkt als kurzfristige Finanzierungsquelle unerlässlich.

Hier setzen die Liquiditätsanforderungen von Basel III an, indem Banken ausreichend Mittelzuflüsse sicher in Aussicht haben müssen, um einen 30-tägigen Liquiditätsengpass am Markt zu überstehen. Darüber hinaus wird das Ausmaß der Fristentransformation beschränkt, indem eine mittel- und langfristige Refinanzierung verlangt wird. Auch diese Anforderungen werden im oben angesprochenen Dokument bcbs 188 in Formeln ausgedrückt (insbesondere Absätze 15/16 und 120/121). Allerdings werden die Faktoren für Mittelzu- und -abflüsse für die Liquiditätsanforderungen vorgegeben, so dass die in der Bank erforderlichen Berechnungen eher deskriptiver Natur sind, während die Eigenkapitalanforderungen teilweise auf statistischen Verteilungen beruhen, die wir im 3. Kapitel kennenlernen.

Auch für Versicherer gelten Mindestkapitalanforderungen. Diese können entweder nach einer Standardformel oder nach einem internen Ansatz berechnet werden, wobei die internen Modelle offengelegt und von der Aufsicht genehmigt werden müssen. In die Standardformel fließen Risiken ein, die versicherungstypspezifisch sind (bspw. Mortalitätsrate bei Lebensversicherern), aber auch operationelle Risiken. Bei der Ermittlung der Eigenkapitalanforderung wird von Annahmen hinsichtlich der Verteilung der Einzelrisiken und bezüglich der statistischen Abhängigkeiten zwischen diesen Einzelrisiken ausgegangen. Ein wesentlicher Kritikpunkt besteht darin, dass lediglich lineare Abhängigkeiten berücksichtigt werden. Auf Abhängigkeitsmaße und statistische Unabhängigkeit werden wir in den Kapiteln 1 und 5 zu sprechen kommen.

Sowohl in Banken als auch in Versicherungsunternehmen muss das Risikomanagement gewisse formale Anforderungen erfüllen. Dazu gehört auch, dass die Personen, die mit Risikomanagement- und Revisionsaufgaben betraut sind, entsprechend qualifiziert sein müssen. Mit Ihrem Studium schaffen Sie die Voraussetzungen für die Erfüllung derartiger aufsichtlicher Anforderungen und reduzieren zudem das operationelle Risiko, dem in der Aufsicht immer größere Bedeutung beigemessen wird.

Wir beginnen dieses Modul mit der deskriptiven Statistik. Sie beschreibt die Informationen in einem Datensatz, fasst diese zusammen und veranschaulicht sie. Zunächst lernen wir einige Kennzahlen, die Datensätze charakterisieren. Die wichtigste Kennzahl, den Mittelwert, kennen Sie schon, möglicherweise als Durchschnitt. Darüber hinaus werden wir uns mit Zusammenhängen zwischen zwei Variablen beschäftigen. Sie können danach Fragen beantworten wie „Verursachen Raucher höhere Kosten in der privaten Krankenversicherung?“ Oder „Wenn das Bruttoinlandsprodukt sinkt, um wie viel Prozent steigen die Unternehmensinsolvenzen unter den Kreditnehmern einer Bank?“.

Danach befassen wir uns mit Wahrscheinlichkeiten. Zunächst legen wir ein paar theoretische Grundlagen, bevor wir uns der induktiven Statistik zuwenden. Sie ist das Herzstück der Statistik und zieht Schlüsse von einer Stichprobe auf die Grundgesamtheit. Diese Schlüsse sind exakt nicht möglich; wir werden dabei Fehler machen, d. h. den wahren Wert der Grundgesamtheit verfehlen. Aber das Besondere ist, dass wir die Fehlerhäufigkeit bzw. -wahrscheinlichkeit quantifizieren können. Ein bestimmtes Konfidenzintervall wird beispielsweise mit 95%iger Wahrscheinlichkeit den wahren, aber unbekannten Wert überdecken (Kapitel 7). In diesen Zusammenhang gehören auch verschiedene statistische Testverfahren (Kapitel 8), beispielsweise „Ist davon auszugehen, dass innerhalb des nächsten Jahres ein Jahrhundertereignis auf die Gebäudeversicherung zukommt?“.

Wenn wir Übungsaufgaben bearbeiten, die anscheinend keinen Bezug zur Finanzwelt haben, dann liegt dies darin begründet, dass Ihnen viele statistische Konzepte aus dem Alltag bekannt sind und ich Ihnen den Zugang zu diesen statistischen Inhalten über Ihre eigenen Alltagserfahrungen ermöglichen möchte. Natürlich können und sollten Sie einschlägige Beispiele aus der Praxis in die Präsenzphasen einbringen.

Unvollständige Tabellen im Text sollen Sie ermuntern, das Beispiel fortzuschreiben und damit die Tabelle selbstständig zu vervollständigen. Dieses Studienmaterial ist als Arbeitsbuch konzipiert, d. h., Sie sollen durchaus darin Rechnungen nachvollziehen und Tabellen ausfüllen.

Mein besonderer Dank gilt Herrn Prof. Dr. Dietmar Pfeifer, der mir großzügig seine Folien zu dieser Veranstaltung überlassen hat. Deshalb profitiert auch dieses Handbuch von seinem reichen Erfahrungsschatz über die Anwendung statistischer Verteilungen in Versicherern. Die Beispiele und Übungsaufgaben sind zum sehr großen Teil entnommen aus Barrow (2009) oder Zucchini et al. (2009). Dies ist nicht für jede Aufgabe einzeln erwähnt.

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

KAPITEL 1: DESKRIPTIVE STATISTIK

Lernziele:

- Arithmetisches Mittel, Median und Modalwert berechnen und interpretieren
- Varianz und Standardabweichung berechnen und interpretieren
- Quantile bestimmen
- Datensätze im Histogramm grafisch darstellen
- Empirische Verteilungsfunktion grafisch darstellen
- Zusammenhänge zwischen zwei Variablen erkennen und quantifizieren
- Regressionsschätzung einer Geraden mit der Methode der Kleinsten Quadrate durchführen und Schätzergebnisse interpretieren

1 DESKRIPTIVE STATISTIK

Die deskriptive oder beschreibende Statistik fasst Informationen zusammen, die in einem u. U. sehr großen Datensatz stecken. Hierzu gibt es sowohl numerische Methoden (Kennzahlen, die einen Datensatz charakterisieren, wie beispielsweise der Mittelwert) als auch graphische Methoden (u. a. das Histogramm, das wir bald kennenlernen werden). Wenn Sie einen kleinen Datensatz betrachten, dann überblicken Sie oftmals die Informationen, die darin stecken, unmittelbar und brauchen deshalb nicht auf die Berechnung von Kennzahlen oder die graphische Aufbereitung zurückzugreifen.

Stellen Sie sich vor, Sie interessieren sich für das Alter der Studierenden in Ihrer Gruppe. Da die Gruppe so klein ist, überblicken Sie die Altersangaben Ihrer Kommilitonen sofort. Wenn Sie sich aber für das Alter aller Studierenden der Universität Oldenburg interessieren, oder sogar für das Alter aller Einwohner Oldenburgs, dann wäre der Datensatz so umfangreich, dass Sie nicht unmittelbar erfassen können, wie alt die betrachteten Personen im Durchschnitt sind oder wie stark die Beobachtungen streuen. Für solche Fälle lernen wir zunächst Lage- und Streuungsmaße kennen.

1.1 Lagemaße

Die Lagemaße sind Kennzahlen, die die Lage der Verteilung erfassen, d. h. also, wie weit links oder rechts auf der Abszisse (das ist die horizontale Achse) sich die Verteilung befindet. Das einfachste und gebräuchlichste Lagemaß ist das arithmetische Mittel, das Sie ganz sicher unter dem Begriff „Durchschnitt" bereits kennen und das auch als „Mittelwert" bezeichnet wird.

Nehmen Sie an, es wären drei Studierende in Ihrer Gruppe, deren Alter 29, 30 bzw. 31 Jahre sein soll. Wie hoch ist das Durchschnittsalter in der Gruppe? Was passiert, wenn ich mich zusätzlich zur Gruppe zählen möchte: Steigt oder sinkt das Durchschnittsalter? Dieses Gedankenexperiment zeigt Ihnen, wie der Mittelwert Auskunft über die Lage der Verteilung gibt.

Wir gehen nun von Ihrer Berechnung des Durchschnittsalters aus und übertragen Ihr Verfahren auf den allgemeinen Fall: Sie müssen zunächst die einzelnen Beobachtungen aufsummieren (im Beispiel: 29+30+31) und diese Summe danach durch die Anzahl der Beobachtungen dividieren (im Beispiel: 3 Studierende). Aus dieser Berechnung ergibt sich das Durchschnittsalter (im Beispiel: 30 Jahre).

Die Beobachtungen nennen wir x_i, deren Summe für n Beobachtungen ist dann $\sum_{i=1}^{n} x_i$, wobei i die Beobachtungen durchnummeriert. Für die erste Beobachtung ist i=1, für die zweite Beobachtung ist i=2 und für die letzte (die n-te) Beobachtung ist i=n. In unserem Beispiel ist n=3, $x_1 = 29$, $x_2 = 30$ und $x_3 = 31$, also $\sum_{i=1}^{3} x_i = 29 + 30 + 31 = 90$. Diese Summe muss nun noch durch die Anzahl der Beobachtungen (n=3) dividiert werden: 90/3=30, also unser Durchschnittsalter.

Formal ausgedrückt erhalten wir

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

wobei $\overline{x}$ unser Mittelwert ist.

Grundsätzlich unterscheiden wir zwischen dem Mittelwert für die Grundgesamtheit und dem für eine Stichprobe. Im Normalfall werden wir mit Stichproben arbeiten. Stellen Sie sich vor, Sie interessieren sich für das Alter der Bevölkerung in Deutschland. Die Grundgesamtheit umfasst alle Beobachtungen, die es gibt, also das Alter für jede der rund 80 Millionen Personen. Meistens ist es nicht möglich, in einer Untersuchung die gesamte Grundgesamtheit zu befragen (dies wäre eine Art Volkszählung wie in der Weihnachtsgeschichte nach Lukas). Stattdessen wird ein möglichst repräsentativer Ausschnitt befragt; dieser Ausschnitt ist die Stichprobe.

In jedem Fall werden alle Beobachtungen aufsummiert und durch die Anzahl der Beobachtungen dividiert. Lediglich die Notation ändert sich: Für die Stichprobe erhalten wir $\overline{x}$, während wir den Mittelwert für die Grundgesamtheit mit μ bezeichnen. Die Anzahl der Beobachtungen ist in der Stichprobe n und in der Grundgesamtheit N , so dass sich für den Mittelwert der Grundgesamtheit ergibt:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Aus dieser Beispielrechnung wird deutlich, dass Sie für die Berechnung des Mittelwertes jede einzelne Beobachtung kennen müssen. Bei gruppierten Daten ist dies nicht der Fall. Im Beispiel wüssten wir dann nur, dass es zwei Studierende im Alter von 21 bis 30 Jahren und eine(n) Studierende(n) im Alter von 31 bis 40 Jahren gibt. In unserem Beispiel ist es unrealistisch, nur gruppierte Daten vorliegen zu haben. Aber stellen Sie sich die Altersverteilung für alle Einwohner Oldenburgs vor – dann hätten Sie 160.000 Beobachtungen, d. h. Altersangaben. Um diese Datenfülle übersichtlicher darzustellen, würde man die Beobachtungen in einer Tabelle zusammenfassen:

Tabelle 1: Beispiel für eine Altersverteilung

Alter in Jahren: Beobachtung	Anzahl der Personen (im Beispiel)
0 bis 10	keine
11 bis 20	keine
21 bis 30	zwei
31 bis 40	eine
usw.	usw.

Jetzt wissen Sie nur, dass zwei Personen in das Intervall 21 bis 30 Jahre und eine Person in das nächste Intervall fallen, aber sie kennen die exakten Beobachtungen (Alter in Jahren) nicht. In dieser Situation müssen Sie eine Annahme bezüglich der Beobachtungen treffen, um einen Mittelwert ausrechnen zu können. Im Allgemeinen geht man davon aus, dass die Beobachtungen innerhalb eines Intervalls nahezu gleich verteilt sind, so dass der Mittelpunkt des Intervalls eine geeignete Näherung für die tatsächliche Beobachtung ist, die Sie nicht kennen. Dann hätten wir zwei Mal 25,5 Jahre und ein Mal 35,5 Jahre in unserem Datensatz. Mit diesen „hypothetischen" Beobachtungen lässt sich der Mittelwert wie oben beschrieben berechnen: (25,5+25,5+35,5)/3=86,5/3=28,8 Jahre. Durch die Gruppierung der Daten verlieren Sie Informationen, so dass das Ergebnis ungenauer ist als bei exakten Beobachtungen.

Wenn Sie in der Praxis Mittelwerte berechnen, sollten Sie sich zunächst überlegen, welche Angabe die Beobachtungen sind (dies ist die Ausprägung des Merkmals, für das Sie sich interessieren, im Beispiel also das Alter in Jahren) und welche die Anzahl bzw. Häufigkeit. Im obigen Beispiel beobachten Sie im 3. Intervall zwei Personen, während im vierten nur eine Person vorkommt. Berechnen Sie also nicht die durchschnittliche Anzahl Personen pro Intervall, sondern tatsächlich das mittlere Alter!

Grafisch können Sie sich den Mittelwert als Schwerpunkt der Verteilung vorstellen. Wenn Sie die Verteilung als Modell aus Holz oder Pappe vor sich hätten und dieses auf einem Stift balancieren sollten, dann liegt der Mittelwert genau dort, wo Sie den Stift ansetzen müssten, damit das Modell nicht wegkippt. Wir kommen darauf später (Kapitel 3) zurück.

Aufgabe 1

Wie viele Fernseher hat jede Familie im Durchschnitt?

Es gibt 10 Familien mit nur einem Gerät im Haushalt, 12 Familien mit 2 Fernsehern und 3 Familien mit 3 Fernsehern.

Aufgabe 2

Ein Autofahrer notiert seine Benzinkäufe auf einer längeren Reise:

Tankstelle	1	2	3
Anzahl Liter	33	40	25
Preis pro Liter	1,62	1,55	1,65

Berechnen Sie den durchschnittlichen Preis pro Liter auf dieser Reise.

Übrigens erwarten Statistiker immer den Mittelwert. Deshalb ist der Mittelwert zugleich der Erwartungswert: Wie viele Fernseher erwarten wir pro Familie? Welchen Preis pro Liter Benzin erwarten wir?

Eine vergleichbare Situation wie bei den gruppierten Daten finden wir im folgenden Beispiel vor.

Tabelle 2: Kosten pro Schüler nach Schultyp

	Grundschule	Sekundarstufe I	Sekundarstufe II
Kosten pro Schüler und Jahr in GBP	1.750	3.100	3.820
Anzahl der Schüler	8.000	7.000	3.000
Anteil der Schüler			

Die politisch interessante Frage ist: Wie viel Geld wird im Durchschnitt über die Stufen pro Schüler ausgegeben? Es ist aus Tabelle 2 unmittelbar ersichtlich, dass die Anteile der Schüler über die Stufen nicht gleich sind. Berechnen Sie zunächst die Anzahl der Schüler insgesamt und vervollständigen Sie mit dem Ergebnis die letzte Zeile der Tabelle. Danach ermitteln Sie die Anteile der Schüler, die auf die drei Stufen entfallen. Diese Anteile sind dann die Gewichte, mit denen Sie die Kosten in jeder Stufe multiplizieren müssen, um den gewichteten Durchschnitt zu berechnen. Zu Ihrer Kontrolle: Die Summe der Gewichte ist 1. Das Ergebnis lautet genau 2.620 Pfund.

Wir haben nun das arithmetische Mittel kennengelernt und können Durchschnitte im „Standardfall“ und für gruppierte Daten berechnen und sind außerdem mit dem gewichteten Durchschnitt vertraut. Den Erwartungswert als Mittelwert haben wir ebenfalls kurz eingeführt.

Es gibt außer dem Mittelwert noch weitere Lagemaße, von denen wir zwei kennenlernen möchten: Median und Modalwert.

Stellen Sie sich vor, Sie sortieren alle Beobachtungen ihrer Größe nach. Diejenige, die in der Mitte steht, ist der Median. Er teilt die Verteilung (einfacher: den Datensatz) in zwei gleiche Teile, d. h. in zwei gleich große Hälften.

In unserem ersten Beispiel sortieren wir zunächst die Studierenden nach dem Alter: 29, 30 und 31. Die mittlere Beobachtung ist 30, d. h. der Median liegt bei 30 Jahren. Der Median ist nicht etwa 2, weil die 2. Person in der Mitte steht, sondern vielmehr die Merkmalsausprägung der 2. Person. Da die drei Studierenden eine symmetrische Altersverteilung abgeben, liegen Mittelwert und Median aufeinander, sind also gleich. Bei einer stark asymmetrischen Verteilung ist dies nicht so.

Nehmen Sie nun an, wir hätten drei Studierende im Alter von 29, 30 und 45 Jahren. Das arithmetische Mittel beträgt – bitte nachvollziehen! – 34,7 Jahre; dieses Maß reagiert daher sehr sensitiv auf den einen 45-jährigen „Ausreißer“ in der Gruppe.

Dies ist oftmals unerwünscht, weil das Lagemaß die Lage der gesamten Verteilung ausdrücken soll und nicht etwa die Lage einzelner extremer Beobachtungen. In dem Fall mit dem einen Ausreißer ist daher der Median als Lagemaß aussagekräftiger als der Mittelwert: 30 Jahre erscheint repräsentativer für die Gruppe als knapp 35 Jahre.

Noch deutlicher wird der Unterschied zwischen Mittelwert und Median bei der Vermögensverteilung. Diese ist in westlichen Ländern zumeist stark asymmetrisch: viele Menschen haben ein jeweils geringes Vermögen, während einige wenige Personen richtig viel Geld haben. Wenn Sie über alle Personen das durchschnittliche Vermögen ausrechnen, erhalten Sie einen überraschend hohen Wert. Das liegt daran, dass der Mittelwert sensitiv auf Ausreißer reagiert, d. h. nach oben „verzerrt“ wird durch die wenigen Reichen. Bei solchen asymmetrischen Verteilungen sollten Sie als Lagemaß den Median verwenden.

Im März 2013 hat die Deutsche Bundesbank eine Untersuchung zum Nettovermögen der privaten Haushalte in Deutschland veröffentlicht (Panel on Household Finances, PHF). Dort wird als Lagemaß der Median (und nicht das arithmetische Mittel!) verwendet. Sie können sich die Untersuchung unter
http://www.ecb.europa.eu/pub/pdf/other/ecbsp2en.pdf ansehen.

Als zweite Alternative zum Mittelwert lernen wir den Modalwert – oder: Modus – kennen. Dies ist der häufigste Wert; deshalb wird er auch als typischer Wert einer Verteilung bzw. einer Stichprobe bezeichnet.

Wenn Sie den Modalwert für gruppierte Daten ermitteln, dann werden in breiteren Intervallen typischerweise mehr Beobachtungen „stecken“ als in schmaleren Intervallen. Deshalb können Sie in dem Fall unterschiedlicher Klassenbreiten nicht einfach die Beobachtungen abzählen und dann die am häufigsten vorkommende Klasse verwenden, sondern Sie müssen zunächst die Häufigkeitsdichte für jedes Intervall berechnen. Diese Dichte ist „bereinigt“ um den Effekt der Klassenbreite, indem die Anzahl der Personen im Intervall durch die Klassenbreite dividiert wird. Je breiter die Klasse, umso geringer die Häufigkeitsdichte bei konstanter Personenzahl.

Zum Abschluss dieses Abschnittes üben wir alle drei Lagemaße am Beispiel der Vermögensverteilung in Großbritannien im Jahr 1979.

Beispiel: Vermögensverteilung in Großbritannien im Jahr 1979

Berechnen Sie Mittelwert, Median und Modalwert und interpretieren Sie die Ergebnisse.

Tabelle 3: Vermögensverteilung in Großbritannien im Jahr 1979

Vermögen in GBP	Anzahl in Tausend
0 bis	1.606
1.000 bis	2.927
3.000 bis	2.562
5.000 bis	3.483
10.000 bis	2.876
15.000 bis	1.916
20.000 bis	3.425
50.000 bis	621
100.000 bis	170
200.000 bis	59
Summe	

Sie erkennen, dass die Klassen unterschiedlich breit sind. Hierauf müssen Sie bei der Ermittlung des Modalwertes besonders achten. Es fällt auf, dass zahlreiche Personen ein sehr geringes Vermögen haben, während einige wenige Personen über ein großes Vermögen verfügen. Dies ist typisch für Einkommens- und Vermögensverteilungen in westlichen Ländern.

Beispiel: Vermögensverteilung in Großbritannien im Jahr 1979

Wie viele Personen wurden insgesamt befragt? Ermitteln Sie die Summe der rechten Spalte. Dies ist Ihr n im ersten Beispiel.

Zunächst zum Mittelwert:

Vermögen in GBP	Anzahl in Tausend f_i	Mittelpunkt des Intervalls x_i	x_i * f_i
0 bis	1.606	500	803.000
1.000 bis	2.927	2.000	5.854.000
3.000 bis	2.562	4.000	usw.
5.000 bis	3.483	7.500	
10.000 bis	2.876	12.500	
15.000 bis	1.916	17.500	
20.000 bis	3.425	35.000	
50.000 bis	621	75.000	
100.000 bis	170	150.000	
200.000 bis	59	300.000	17.700.000
Summe	19.645	nicht sinnvoll	322.157.500

Wie sind die 803.000 zu interpretieren, was bedeuten die 322.157.500?

Im letzten Schritt teilen Sie das gesamte Vermögen (322.157.500 GBP) auf die 19.645 befragten Personen auf:

322.157.500/19.645 = 16.399. Ein Durchschnittsbrite verfügt über ein Vermögen von knapp 16.400 Pfund.

Für den Median sortieren Sie die Personen nach der Höhe ihres Vermögens. Es gibt 19.645 Personen, die Mitte liegt bei der 9.823. Person. In welches Intervall fällt diese Person? Sie müssen die Häufigkeiten kumulieren:

Vermögen in GBP	Anzahl in Tausend f_i	Kumulierte Häufigkeit
0 bis	1.606	1.606
1.000 bis	2.927	1.606+2.927=4.533
3.000 bis	2.562	4.533+2.562=7.095
5.000 bis	3.483	7.095+3.483=10.578
10.000 bis	2.876	usw.
15.000 bis	1.916	
20.000 bis	3.425	
50.000 bis	621	
100.000 bis	170	
200.000 bis	59	19.645
Summe	19.645	nicht sinnvoll

Die 9.823. Person fällt in das Intervall 5.000 bis 9.999,99 Pfund. Wir gehen wiederum davon aus, dass die Vermögen innerhalb eines Intervalls gleichverteilt sind, so dass anzunehmen ist, dass das Vermögen dieser Person auf die Intervallmitte fällt: rund 7.500 Pfund.

Im Vergleich zum Mittelwert von 17.000 Pfund ist das recht gering; der Mittelwert reagiert sensitiv auf die wenigen Wohlhabenden und ist deshalb nach oben verzerrt.

Wenden wir uns nun dem Modalwert zu:

Vermögen in GBP	Klassenbreite (rund)	Anzahl in Tausend f_i	Häufigkeitsdichte: Personenzahl/Klassenbreite
0 bis	1.000	1.606	1.606/1.000=1,6
1.000 bis	2.000	2.927	2.927/2.000=1,5
3.000 bis	2.000	2.562	2.562/2.000=1,3
5.000 bis	5.000	3.483	usw.
10.000 bis	5.000	2.876	
15.000 bis	5.000	1.916	
20.000 bis	30.000	3.425	
50.000 bis	50.000	621	
100.000 bis	100.000	170	
200.000 bis	200.000	59	
Summe		19.645	nicht sinnvoll

Hier müssen wir eine Annahme treffen, wo das letzte Intervall enden soll. Es zeigt sich bei den letzten Intervallen ein gewisses Muster bezüglich der Klassenbreite: 50-100-?? Ich schlage daher 200 vor; das Intervall endet demnach bei 400.000 Pfund. Wenn Sie einen anderen Wert vorschlagen und diesen plausibel begründen können, wäre dies natürlich ebenso richtig.

Das Intervall mit der höchsten Häufigkeitsdichte ist der Modalwert – also das erste Intervall von 0 bis 999,99 Pfund. Der Intervallmittelpunkt liegt bei rund 500 Pfund.

Fassen wir unsere Ergebnisse zusammen:

Tabelle 4: Lagemaße für die Vermögensverteilung in Großbritannien im Jahr 1979

	in Pfund, gerundet
Mittelwert	16.400
Median	7.500
Modalwert	500

Der Mittelwert reagiert sensitiv auf die wenigen Wohlhabenden in unserer Stichprobe und fällt deshalb recht hoch aus. Weil seine Aussagekraft in einer so asymmetrischen Verteilung eingeschränkt ist, haben wir darüber hinaus noch Median und Modalwert ermittelt. Der Median teilt die Stichprobe in zwei gleich große Teile (nach Personenzahl, nicht nach Vermögen!). Am häufigsten beobachten wir die untere Vermögensklasse, in der die Personen annahmegemäß nur 500 Pfund besitzen. Bei stark asymmetrischen Verteilungen fallen die drei Lagemaße weit auseinander.

Aufgabe 3

Die Tabelle zeigt die Anzahl der Betriebe im produzierenden Gewerbe in Großbritannien im Jahr 1991/1992 geordnet nach Zahl der Beschäftigten:

Anzahl der Beschäftigten	Anzahl der Betriebe
1 bis	95.409
10 bis	15.961
20 bis	16.688
50 bis	7.229
100 bis	4.504
200 bis	2.949
500 bis	790
1.000 bis	332
Gesamt	

Berechnen Sie das arithmetische Mittel, den Median und den Modalwert und interpretieren Sie Ihre Ergebnisse.

1.2 Quantile

Sie kennen bereits den Median. Er teilt die Verteilung nach Anzahl der Beobachtungen in zwei gleich große Teile (Hälften). Stellen Sie sich vor, Sie wollten die Verteilung nicht in zwei, sondern in vier gleich große Teile aufteilen, wiederum nach der Anzahl der Beobachtungen.

Spielen Sie ein Musikinstrument? Dann wissen Sie bestimmt, wie die vier gleich großen Teile der Verteilung heißen. Wenn vier Personen gemeinsam musizieren, dann spielen sie im Quartett. Die vier gleich großen Teile der statistischen Verteilung heißen Quartile.

Und wie wäre es bei fünf gleich großen Teilen? Fünf Musiker spielen im Quintett, also sind die Teile Quintile.

Und jetzt für zehn gleiche Teile? Dies sind Dezile.

Und abschließend für 100 gleiche Teile? Sie heißen Perzentile.

Im obigen Beispiel haben Sie sich mit der Vermögensverteilung in Großbritannien beschäftigt. Nehmen Sie nun an, Sie wollten wissen, wie groß das Vermögen der unteren 10 % der Bevölkerung maximal ist. Sie würden also 10 % am linken Verteilungsrand abschneiden und suchen die Stelle, d. h. das Vermögen, an dem Sie schneiden müssen.

Insgesamt wurden 19.645 Personen befragt. Davon 10 % sind 1.965 Personen. Die 1.965. Person fällt in das zweite Intervall, also 1.000 bis 2.999,99 Pfund. Wenn wir den Intervallmittelpunkt als Näherung annehmen und runden, haben die unteren 10 % ein Vermögen von maximal 2.000 Pfund.

Aufgabe 4

Wie viel Vermögen haben die oberen 25 % mindestens? (Vermögensverteilung von 1979)

Wir kommen auf die Quantile erneut zu sprechen, wenn wir konkrete Wahrscheinlichkeitsverteilungen kennenlernen. Beim Value at Risk schneiden wir ebenfalls eine gegebene Fläche unter der Verteilung ab und interessieren uns für die Stelle, an der zu schneiden ist.

1.3 Streuungsmaße

Nachdem wir drei Lagemaße kennengelernt haben, wenden wir uns nun der Frage zu, wie breit die Verteilung streut. Diese Streuung ist ein wesentliches Charakteristikum einer Verteilung.

Das einfachste Streuungsmaß ist die Streubreite. Sie ist die Differenz zwischen größter und kleinster Beobachtung. Dies ist problematisch, weil:

- die größte Beobachtung nicht immer bekannt ist (siehe beispielsweise Tabelle 4 oder Aufgabe 4);
- die Streubreite auf den zwei extremsten Beobachtungen beruht und deshalb wenig aussagekräftig ist;
- keine Informationen über die Form der Verteilung in die Betrachtung einfließen.

Das gebräuchlichste Streuungsmaß ist die Varianz bzw. die Standardabweichung. In deren Berechnung geht jede einzelne Beobachtung ein. Das Ziel ist die Erfassung des Abstands einer Beobachtung vom Mittelwert der Verteilung. Für jede Beobachtung wird ein solcher Abstand berechnet und schließlich werden die Abstände über die Beobachtungen aufsummiert. Das Aufsummieren und Dividieren durch die Anzahl der Beobachtungen n kennen Sie bereits aus der Durchschnittsberechnung. Die Varianz wird mit σ^2 (Grundgesamtheit) bzw. s^2 (Stichprobe) bezeichnet.

Schauen wir uns die Formel für die Varianz an:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Sie erkennen unmittelbar, dass Sie für die Varianzberechnung zunächst den Mittelwert μ kennen bzw. ermitteln müssen. Sie berechnen dann für jede Beobachtung x_i die Differenz zum Mittelwert. Diese Differenz müssen Sie quadrieren. Was würde passieren, wenn Sie die Differenzen nicht quadrieren? Es gibt sowohl positive als auch negative Abweichungen, denn es gibt Beobachtungen, die größer als der Mittelwert sind und andere, die kleiner als der Mittelwert sind. Wenn Sie diese positiven und negativen Abweichungen dann aufsummieren, könnten Sie bei null landen! Das Quadrieren hat außerdem den Effekt, dass Abweichungen, die betragsmäßig kleiner als eins sind, (noch) kleiner werden, während solche, die betragsmäßig größer als eins sind, größer werden. Insofern reagiert die Varianz, ebenso wie der Mittelwert, sensitiv auf extreme Beobachtungen. Nachdem Sie die quadrierten Abweichungen aufsummiert haben, müssen Sie die Summe noch durch die Anzahl der Beobachtungen, d. h. durch die Anzahl der quadrierten Abweichungen, dividieren. Sie erhalten dadurch eine durchschnittliche quadrierte Abweichung.

Da dies ein quadriertes Maß ist, wird sie auch in einer quadrierten Einheit gemessen: Jahre^2, Pfund^2 usw. Dies ist natürlich nicht interpretierbar, gibt aber einen Eindruck davon, wie breit die Beobachtungen streuen („Streuungsmaß“).

Um zu einem interpretierbaren, nicht-quadrierten Streuungsmaß zu gelangen, ziehen wir abschließend die Quadratwurzel aus der Varianz und erhalten die Standardabweichung:

$$\sigma = \sqrt{\sigma^2}$$

Die Standardabweichung wird in regulären Einheiten gemessen (Jahre, Pfund usw.) und kann unmittelbar interpretiert werden. Sie benötigen für die Berechnung der Standardabweichung immer erst die Varianz und dafür immer erst den Mittelwert.

Wenn Sie sich für das Durchschnittsalter der deutschen Bevölkerung interessieren, dieses aber für die Grundgesamtheit nicht kennen, würden Sie eine Stichprobe ziehen, das Durchschnittsalter für die Stichprobe berechnen und hoffen, dass Sie möglichst nahe am wahren Wert für die Grundgesamtheit liegen.

Der wahre Mittelwert für die Grundgesamtheit heißt μ, die dazugehörige Varianz σ^2 und die Standardabweichung entsprechend σ. Für die Stichprobe sind es $\overline{x}$ für den Mittelwert, s^2 für die Varianz und s für die Standardabweichung.

Die Formel für die Stichprobe lautet entsprechend:

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

bzw.

$$s = \sqrt{s^2}$$

Wenn Sie die Varianz für die Stichprobe berechnen und zuvor den Mittelwert ebenfalls aus der Stichprobe ermitteln, also den wahren Wert μ nicht kennen, sollten Sie bei der Berechnung von s^2 nicht durch n dividieren, sondern durch *(n-1)*.

Die Formel (sie heißt auch „Schätzer“) für s^2 trifft im Mittel den wahren Wert der Grundgesamtheit, also σ^2, wenn wir durch *(n-1)* dividieren, während der Schätzer, der durch n dividiert, im Mittel neben (unter) dem wahren Wert liegt. Schätzer, die im Mittel den wahren Wert treffen, heißen „erwartungstreu“ oder „unverzerrt“ (auf Englisch: unbiased). Entsprechend liegen verzerrte Schätzer im Mittel daneben. Sie haben trotzdem ihre Existenzberechtigung, wenn sie zwar im Mittel daneben liegen, aber nur wenig, während andere Schätzer im Mittel richtig liegen, aber mit einer recht hohen Wahrscheinlichkeit weit daneben landen (also breit streuen).

In dem Tabellenkalkulationsprogramm Microsoft Excel sind deshalb jeweils zwei Formeln für die Varianz und die Standardabweichung hinterlegt: Einmal die für die Grundgesamtheit (VAR.P) – hier kennen Sie μ – und einmal die für die Stichprobe – hier schätzen Sie μ durch $\overline{x}$ –, die dann erwartungstreu ist (VAR.S). Auf Englisch ist „population“ die Grundgesamtheit und „sample“ die Stichprobe, daher die Abkürzungen in Excel. Sie sollten wissen, wann Sie welche Formel anzuwenden haben.

Aufgabe 5

Berechnen Sie für die Vermögensverteilung in Großbritannien die Standardabweichung. Benutzen Sie als Mittelwert $\overline{x}$ *=16.399 Pfund (Beispiel oben). Sie haben hier gruppierte Daten, so dass Sie den Intervallmittelpunkt zugrunde legen müssen.*

Vermögen in GBP	Anzahl in Tausend
0 bis	1.606
1.000 bis	2.927
3.000 bis	2.562
5.000 bis	3.483
10.000 bis	2.876
15.000 bis	1.916
20.000 bis	3.425
50.000 bis	621
100.000 bis	170
200.000 bis	59
Summe	

Aufgabe 6

Berechnen Sie für die Betriebsgröße in Großbritannien die Standardabweichung. Der Mittelwert $\overline{x}$ *wurde in Aufgabe 3 aus der Stichprobe geschätzt. Runden Sie auf ganze Mitarbeiter.*