HWZ-Wissen

Andri Signorell

Statistische Datenanalyse

Theorie und Praxisanwendung mit Excel

HWZ

VERLAG SKV

Andri Signorell
HWZ Hochschule für Wirtschaft Zürich

Statistische Datenanalyse

VERLAG SKV

1. Auflage 2024

Andri Signorell
HWZ Hochschule für Wirtschaft Zürich

ISBN 978-3-286-34931-5

Das Werk erscheint als E-Book unter der ISBN 978-3-286-11870-6 (PDF)
und als ePub unter der ISBN 978-3-286-11871-3 (EPUB)

Umschlagbild: Pasuwan/Shutterstock.com

Haben Sie Fragen, Anregungen oder Rückmeldungen?
Wir nehmen diese sehr gerne per E-Mail an feedback@verlagskv.ch entgegen.

Inhaltsverzeichnis

Vorwort

«Das H in unserem Namen steht für Hochschule». Das waren die Worte meines Mathematikdozenten, als er mir mit ernster Miene im ersten Semester des Bachelorstudiums die erste von ihm korrigierte Prüfung überreichte. Bei einem Klassendurchschnitt von 3.4 war mein Ergebnis von 2.3 ein besonderes Lerngeschenk. Für mich und meine Frustrationstoleranz.

Nach dem Unterricht ging ich zu ihm und fragte ihn, ob es für mich überhaupt Sinn mache, mit dem Studium weiterzufahren. Seine Antwort war für mich wegweisend: «Sie sind einer, der gut auswendig lernt. Versuchen Sie das nächste Mal, zuerst die Frage zu verstehen, wenn Sie sich ans Lernen machen. Mathematische und statistische Methoden sind im realen Leben zuerst einmal Werkzeuge zur Lösungsfindung. Die Berechnungen und angewandten Methoden sind dann das Resultat dieser Denkweise, probieren Sie es aus.»

Oft hört man von Studierenden, dass man nicht alles zu wissen braucht, sondern nur, wo die Antwort zu finden ist. Mehr und mehr liefern uns nun aber das Internet und neue Anwendungen sehr rasch Antworten und es wird zur Kernkompetenz, ihnen kluge Fragen zu stellen. Gerade mit dieser Entwicklung sollten wir uns darauf besinnen, dass unser Wissen über mathematische und statistische Methoden uns befähigen, nicht nur um Anwendungen zu nutzen oder Antworten weiterzutragen. Vielmehr sind sie die Grundlage, um den Frage- und Erkenntnisprozess zu verstehen.

Daten als Entscheidungsgrundlage sind letztlich Methodenentscheidungen, die ihren Ursprung in der Mathematik und vor allem der Statistik haben.

Für den Bezug zu unserem Leben und unseren Entscheidungssituationen ist dabei die angewandte Statistik die zentrale Grundlage. Mit dem Kauf dieses Buches haben Sie in Ihre wissenschaftliche Mündigkeit und Ihre Kernkompetenzen investiert.

Herzlichen Glückwunsch dazu! Der Spass kann beginnen!

Prof. Dr. MBA Georges-Simon Ulrich,
Direktor Bundesamt für Statistik und Chair of the Statistical Comission at the United Nations

Vorwort des Autors

Unser Leben ist von Unsicherheit geprägt. Zufällige Ereignisse begegnen uns laufend und überall. Wie viel Schnee wird im nächsten Winter fallen? Wie viel verdiene ich nach Abschluss meines Studiums? Wie gross ist der Umsatz unseres Unternehmens im laufenden Jahr? Werden wir heiraten? Wie viel Zins zahlt die Pensionskasse Ende Jahr auf unser Vorsorgekapital? Auf alle diese Fragen können wir keine exakten Antworten liefern, aus heutiger Perspektive erscheinen sie zufällig. Meistens fehlt uns das notwendige Wissen, um die Phänomene deterministisch beschreiben zu können. Gewissen Ausgängen werden wir aber nichtsdestotrotz, vielleicht aufgrund von Erfahrung oder theoretischen Überlegungen, eine höhere Wahrscheinlichkeit zusprechen als anderen.

Andernorts wiederum sehen wir uns der Kraft der Anschaulichkeit anekdotischer Beobachtungen ausgesetzt und sind versucht, Muster zu sehen, wo keine sind. Manche Zeitgenossen vermeinen unser aller Schicksal in den Sternen lesen zu können, andere sind überzeugt, dass Krankheiten mit homöopathischen Zuckerkügelchen geheilt werden können. Hypothesen sind schnell gebildet. So könnte z. B. behauptet werden, dass erstgeborene Kinder bevorzugt zu spät geboren werden, also die Schwangerschaft beim ersten Kind länger als bei Nachgeborenen dauern soll. In lebhaften Diskussionen meinen dann andere, das sei ein Mythos und da sei nichts dran, und wiederum andere sagen, es sei genau umgekehrt. Oft werden dabei eigene Beobachtungen zur Untermauerung der Behauptungen angeführt. Das hört sich dann etwa so an: «Meine beiden Freundinnen, die vor kurzem ihre ersten Babys zur Welt gebracht haben, haben beide ihre Kinder fast 2 Wochen übertragen.» oder «Unser erstes Kind kam 2 Wochen nach Termin, das zweite war 4 Tage zu früh da.»

Berichte wie diese werden als «anekdotische Belege» bezeichnet, weil sie auf unveröffentlichten und meist persönlichen Daten beruhen. Als Unterhaltung an der Party ist gegen Anekdoten natürlich nichts einzuwenden. Für die ernsthafte Prüfung der Fragestellung sind sie allerdings aus mehreren Gründen ungeeignet. Typischerweise ist bei derartigen Aussagen die Anzahl Beobachtungen klein. Eine kleine Anzahl Beobachtungen taugt aber nicht dazu, tatsächlich existierende Effekte zu erkennen. Weiter sind die herangezogenen Fälle oft nicht repräsentativ, die argumentierenden Personen neigen dazu, Beispiele beizusteuern, die ihre These bestätigen, und jene auszublenden, die ihr widersprechen. Wenn schliesslich die Anekdoten persönliche Geschichten sind, werden sie oft ungenau erinnert und verfälscht wiedergegeben.

Die Grenzen von Anekdoten lassen sich mit Hilfe der Instrumente der Statistik überwinden. Diese umfassen Methoden der Datenerhebung, mit denen sichergestellt werden kann, dass genügend viele Daten mit guter Repräsentativität in die Analyse miteingeschlossen werden. Dann geht es um die Verwendung von Kennzahlen, Schätzungen und Tests, um zu prüfen, ob ein Effekt zufällig zustande gekommen sein könnte oder nicht. Damit gelangen wir zu Schlussfolgerungen, die besser zu rechtfertigen sind und mit grösserer Wahrscheinlichkeit richtig sind. In diesem Buch

geht es um die Art und Weise, wie wir mit Daten umgehen und Schlussfolgerungen daraus ziehen.

In der beruflichen Welt von heute wird solches statistisches Basiswissen zunehmend unabdingbar. Die Vermittlung der notwendigen Grundkenntnisse und des kompetenten Umgangs mit Werkzeugen der Statistik hat deshalb ihren festen Platz in den Lehrplänen vieler Disziplinen auf Ebene Fachhochschule/Universität gefunden. Angewandte Datenanalyse beinhaltet die Synthese verschiedener Elemente. Nachdem man erst einmal den Zweck und den Umfang der Analyse gründlich geklärt hat, muss man die Merkmale der verwendeten Daten verstehen und deren Schwächen erkennen. Man nutzt die Theorie, um ein Modell des Prozesses zu erstellen, der die Daten hervorgebracht haben könnte. Mit Hilfe der statistischen Methoden, die sich auf die Wahrscheinlichkeitstheorie stützen, werden dann die Daten zusammengefasst, z. B. in Schätzungen. Dieser Vorgang erfordert auch Fähigkeiten im Umgang mit Software, seien dies Tabellenkalkulationen oder spezialisierte Statistikprogramme. Abschliessend muss man dann in der Lage sein, die Statistiken oder Schätzungen im Hinblick auf ihren ursprünglichen Zweck und die Theorie zu interpretieren.

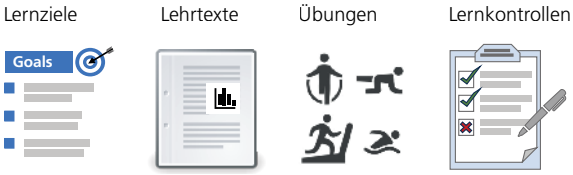
Das vorliegende Buch ist als Einführung in die Statistik gedacht. Es enthält die typischen Inhalte für einen zweisemestrigen Statistik-Grundkurs an einer Fachhochschule. Das Buch ist modular aufgebaut, wobei darauf geachtet wurde, die einzelnen Module umfangmässig so zu gestalten, dass sie in einem 4-stündigen Vorlesungs-Block inklusive einer Nachbearbeitungszeit vergleichbaren Umfanges bearbeitet werden können.

Die Module 1–5 behandeln die gängigen Methoden der deskriptiven Statistik. Hier geht es um die Frage, wie man empirisch gewonnene Daten aufbereitet, wie man sie organisiert, wie man sie darstellt, wie man sie auswertet, wie man das Ergebnis dieser Auswertung interpretiert. Zentral ist hier die univariate und bivariate Beschreibung der Daten mittels Kennzahlen und grafischen Darstellungen. Das insbesondere im Umfeld der ökonomischen Fragestellungen wichtige Konzept der Konzentration rundet diesen ersten Block ab. Die Folge bilden die Grundlagen der Wahrscheinlichkeitsrechnung in den Modulen 6–10. In diesen Modulen werden nach Einführung der Grundbegriffe und Rechenregeln die wichtigen Konzepte von Zufallsvariablen und ihren theoretischen Wahrscheinlichkeitsverteilungen entwickelt.

Darauf aufbauend werden in den Modulen 11–15 die fundamentalen Konzepte der schliessenden Statistik eingeführt. Dabei handelt es sich um stichprobengestützte Verfahren für Punktschätzungen, Konfidenzintervalle und Hypothesentests.

Den Abschluss bilden die für die praktische Analysearbeit wichtigen Modelle der linearen Regressionsrechnung. Dabei wird die Zielsetzung verfolgt, die Studierenden so weit zu bringen, dass sie damit ihre eigenen realen datenanalytischen Probleme zufriedenstellend modellieren können.

Die einzelnen Module haben stets dieselbe Struktur, die die folgenden Elemente umfasst:



Einer allgemeinen Einordnung der Thematik folgt die Vorstellung der Lernziele für das jeweilige Modul. Danach werden die inhaltlichen Themen anhand von Lehrtexten und konkreten Beispielen eingeführt. Themen, die im Zusammenhang als relevant betrachtet werden, aber über den üblichen Stoff in Grundlagen-Kursen hinausgehen, werden mit einem Sternchen (*) gekennzeichnet. An vielen Stellen finden sich Hinweise auf die Behandlung der Problemstellungen mit dem PC. Nützliche Excel-Funktionen für den jeweiligen Aufgabenbereich werden eingeführt und beschrieben. Den Abschluss eines jeden Moduls bildet eine umfassende Sammlung von Übungsaufgaben.

Die Lehrtexte sind so ausführlich wie nötig und so knapp wie möglich formuliert. Dabei wurde auf Textverständlichkeit geachtet, ohne es an der notwendigen Exaktheit mangeln zu lassen. Im Sinne eines praxisorientierten Ansatzes wurde der Fokus auf die Frage «Wie?» gelegt. Gemeint ist damit, dass mathematische Formalismen und Herleitungen auf das absolut Notwendige beschränkt wurden. Für die angewandte Datenanalyse ist es vielfach nicht zwingend notwendig, die Mathematik hinter einem Verfahren zu beherrschen, um es adäquat und sinnvoll praktisch anzuwenden. Häufig genügt es, die einzelnen Verfahren zu kennen und zu wissen, für welche Fragestellungen sie geeignet sind, wie sie durchgeführt werden, welche Voraussetzungen erfüllt sein müssen und vor allem, was sich aus den Ergebnissen für die inhaltliche Fragestellung ableiten lässt. Die statistische Arithmetik wird im Buch daher ausschliesslich an jenen Stellen zum Thema gemacht, an denen sie im Sinne der Didaktik angebracht erscheint und dem Lernenden grundlegende Einsichten zu verschaffen verspricht.

Beim Erarbeiten des Stoffes ist es eine gute Idee, sich von Konfuzius leiten zu lassen. Seine Ideen haben in all den Jahren nichts an Aktualität verloren:

> *«Ich höre und vergesse. Ich sehe und behalte. Ich handle und verstehe.»*
>
> *Konfuzius (ca. 500 v. Chr.)*

Statistik ist ein verständnisintensiver Stoff, der sich den meisten von uns nur erschliesst, wenn er aktiv erarbeitet wird. Den Übungen kommt deshalb in diesem Zusammenhang zentrale Bedeutung zu. Die Module umfassen eine Vielzahl an Aufgabenstellungen, anhand derer der Stoff möglichst vielfältig erarbeitbar und erfassbar gemacht werden soll. Den Studierenden soll die Möglichkeit geben werden, ihren individuellen Lernweg durch die komplexe Materie zu finden. Die Aufgaben wurden nach Möglichkeit basierend auf (potenziell) realen Problemstellungen for-

muliert, auch in der Absicht, dass die Methoden damit anschliessend besser in die Praxis transferiert werden können.

Ein weiteres – für den Lernprozess entscheidendes – Element sind PC-gestützte formative Lernkontrollen. Damit können Sie selbstgesteuert Ihren Lernstand prüfen und entwickeln. Die interaktive Form der Aufgabenstellungen und Prüfungsfragen legt schnell individuelle Wissenslücken offen, die dann zielgerichtet gestopft werden können.

Versuchen Sie den Empfehlungen der modernen Didaktik zu folgen: Um die Lernziele zu erreichen (und die Prüfungen zu bestehen), müssen Sie viel selbstständig arbeiten. Betrachten Sie die Übungen in diesem Zusammenhang als OBLIGATORISCH! Versuchen Sie den Stoff in kleine Portionen aufzuteilen und vermeiden Sie lange Lernsequenzen (die sich oft als ineffizient erweisen). Bereiten Sie sich auf Vorlesungen vor, indem Sie die Lehrtexte vorgängig lesen. Wiederholen Sie anschliessend kontinuierlich den Stoff der Vorlesungen.

Eine heutzutage zentrale Rolle in der Datenanalyse spielt der PC. Rechenkapazität ist im Überfluss vorhanden. Die breitflächige und niederschwellige Verfügbarkeit von statistischen Applikationen hat zur Folge, dass Datenauswertungen praktisch nicht mehr «von Hand» gemacht werden (müssen). Dieser Umstand wurde auch in den Lehrtexten ausgiebig berücksichtigt. Das Rechnen wird an den allermeisten Stellen dem Computer überlassen. Damit verlagern sich die Ansprüche an die Benutzer von der Arithmetik zur methodischen Kompetenz. Wichtig ist, dass Sie die geeigneten Verfahren für eine bestimmte Fragestellung auswählen, die hierfür notwendigen Parameter angemessen festlegen und die von den Systemen gelieferten Resultate hinsichtlich der ursprünglichen inhaltlichen Fragestellung interpretieren können.

Die Entscheidung für eine geeignete Software ist nicht leicht (und wird oft kontrovers diskutiert). Es gilt viele Aspekte zu berücksichtigen. Für die Motivation der Studierenden kann es entscheidend sein, ob ein Tool eingesetzt wird, welches ihnen berufliche Chancen eröffnet oder ein Produkt, das in ihrem beruflichen Umfeld eher unbekannt ist. Dedizierte Statistik-Software wie SAS, SPSS, Stata oder R weist naturgemäss eine gewisse Komplexität auf. Eine Einführung kann sehr viel Zeit und Ressourcen kosten. Es erweist sich deshalb als sinnvoll, sich in Einführungskursen in die Statistik eines weniger komplexen Werkzeugs zu bedienen, mit dem die Studierenden schon mehr oder weniger vertraut sind. (Dies gilt vor allem, für weniger technisch orientierte Studienrichtungen, wie z. B. Betriebsökonomie). Aufgrund der breiten Verfügbarkeit und der einfachen Bedienung bietet sich das Tabellenkalkulationssystem Microsoft Excel an. Es ist sehr weit verbreitet und die meisten Studierenden auf Fachhochschulstufe haben leidliche Vorkenntnisse, die es auch aufgrund knapper Zeitbudgets im Umfeld der beruflichen Bildung zu nutzen gilt. Die statistischen Möglichkeiten des Systems sind zwar limitiert, reichen aber für den vorliegenden Stoffumfang gerade noch so aus. Auf Nachteile oder Stolpersteine im Umgang mit Excel – die durchaus existieren (!) – wird an entsprechender Stelle im Buch aufmerksam gemacht.

Beachten Sie: Das Buch ist kein Einführungskurs in Excel, sondern zeigt, wie Excel für datenanalytische und statistische Aufgabenstellungen optimal genutzt werden kann. Stellen Sie vorgängig sicher, dass Sie über die notwendigen Grundkenntnisse in Excel wie den Umgang mit Dateien, Formeln, Bezügen und Funktionen verfügen. Allfällige Defizite lassen sich gut mit vielfältig auf dem freien Markt verfügbarer Literatur oder entsprechenden Kursen ausräumen.

An der Entstehung eines Buches sind stets viele Personen beteiligt. Ich hatte das Vergnügen und die Ehre, von der Kompetenz und der Anregung der nachstehenden Fachleute zu profitieren. Ich bedanke mich an dieser Stelle bei allen Beteiligten herzlich für all das Engagement, für die geteilte Fachkompetenz und für die stets angenehme und anregende Zusammenarbeit!

Intensiv mit der Struktur und dem didaktischen Aufbau des Texts auseinandergesetzt hat sich dipl. Ing. ETH Beat Scherrer. Die auf seiner langjährigen Lehrtätigkeit als Hochschul-Dozent basierenden Anmerkungen führten an überaus vielen Stellen des Buchs zu signifikanten Verbesserungen. Dass die Übungen sinnvoll und sachlich logisch formuliert sind, hat lic. iur. Beat Moser sichergestellt. Er stellte zudem in minuziöser Kleinarbeit sicher, dass die in den Aufgabenstellungen erforderlichen Kompetenzen im Theorieteil ausreichende und sprachlich verständliche Erwähnung fanden. Dass die mathematischen Formulierungen und Behauptungen im Text ihre Richtigkeit haben, hat Lukas Graz (Ms. Sc. ETH in Statistik) vom statistischen Beratungsdienst der ETH Zürich detailliert geprüft.

Meinen Kollegen Heidi Dux und Dominik Vordermann verdanke ich inhaltliche Diskussionen über Stoff, Umfang und Tiefe. Sie gaben mir viele Anregungen für die Auswahl, Zusammenstellung und Ordnung der Inhalte.

Das Projekt initiiert, finanziert und kompetent geleitet hat Dr. Daniel Schmid aus der HWZ-Schulleitung.

Von Seiten des Verlags hat sich Gianni Cocchiarella und sein Team in langwieriger und geduldiger Kleinstarbeit meinen Layout-Forderungen gestellt und die meisten davon möglich gemacht. Das frische und aufgeräumte Druckbild ist massgeblich auf deren Layoutvorlage zurückzuführen.

Die öffentliche Statistik in der Schweiz wird massgeblich durch das Bundesamt für Statistik (BfS) geprägt. Es ist mir eine Freude, dass der Direktor des BfS Prof. Dr. Georges Ulrich meinem Buch ein Vorwort gewidmet hat.

Meine Frau Silvia entwickelte zum Buch nie eine wirkliche Liebesbeziehung. Wie sollte sie auch? Die Produktionsphase war meinerseits von unzähligen Absagen an soziale Verpflichtungen geprägt, die sie dann allein wahrnehmen musste. Dafür, dass sie dies geduldig und verständnisvoll ertrug und ich sie auch nach Abschluss des Projekts fest an meiner Seite weiss, bin ich ihr von Herzen dankbar.

Als letztes bedanke ich mich bei Ihnen, geschätzter Leser, geschätzte Leserin! Ich hoffe, das vorliegende Buch leiste einen signifikanten Beitrag zu Ihrer datenanalytischen Ausbildung. Erstrebenswert wäre, dass Sie anschliessend das Leben auch aus einer anderen, quantitativeren Perspektive betrachten. In jedem Fall wünsche ich

Ihnen viele Lernerfolge und – vor allem – Spass an der quantitativen Datenanalyse in den vor Ihnen liegenden Statistik-Stunden.

Wir alle, die am Projekt beteiligt waren, haben uns redlich bemüht, Fehler zu vermeiden, resp. zu erkennen und auszumerzen. Es ist nicht garantiert, dass uns dies überall und an allen Stellen gelungen ist. Falls Sie Fehler entdecken, sind wir dankbar, wenn Sie dies per E-Mail an die Adresse andri@signorell.net melden. Die Studierenden von morgen werden es Ihnen ebenfalls danken!

Formales

Aus Gründen der besseren Lesbarkeit wird in diesem Buch vielfach das generische Maskulinum verwendet. Dies impliziert immer beide Formen, schliesst also die weibliche Form (und alle weiteren) mit ein.

Folgende Formate werden verwendet:

Format	
Funktionen	**SUMME()**
Tastatureingaben	<Ctrl+A>
Excel-Menus	**Datei | Optionen | Add-Ins**
Datensätze	**titanic.xlsx**
Excel-Hinweis	 Wie wird die Aufgabe mit Excel gelöst?
Buttons	OK

Folgende Notationen werden verwendet:

Format	
$\bar{x}$	Empirischer Mittelwert
s, s^2	Empirische Standardabweichung und Varianz
F(x)	(Empirische) Verteilungsfunktion
X, Y, Z	Zufallsvariablen
E(X)	Erwartungswert der Zufallsvariablen X
Var(X), sd(X)	Varianz und Standardabweichung der Zufallsvariablen X
X ~ B(n, p)	«X ist binomialverteilt mit n und p»
B(n, p)	Binomialverteilung mit den Parametern n und p
H(N, K, n)	Hypergeometrische Verteilung mit N, K und n
G(p)	Geometrische Verteilung mit p
nB(b, p)	Negativ-Binomialverteilung mit den Parametern b und p
Pois(λ)	Poissonverteilung mit l
N(μ, σ)	Normalverteilung mit Mittelwert m und Standardabweichung σ
U(a, b)	Uniforme Verteilung mit a und b
Exp(λ)	Exponentialverteilung mit λ
t(df)	Student-t-Verteilung mit df
F(df_1, df_2)	F-Verteilung mit df_1 und df_2
χ^2(df)	Chi-Quadrat-Verteilung mit df
df	Freiheitsgrade (engl: degrees of freedom)
$q_{B(n, p), \alpha}$	α-Quantil der Binomialverteilung mit n und p
$q_{N(\mu, \sigma), \alpha}$	α-Quantil der Normalverteilung mit μ und σ
P(X=k)	Wahrscheinlichkeit, dass die diskrete Zufallsvariable X genau den Wert k annimmt
P(a ≤ X ≤ b)	Wahrscheinlichkeit, dass die kontinuierliche Zufallsvariable X einen Wert zwischen a und b annimmt
$\beta 0, \ldots \beta_i$	b-Koeffizienten eines linearen Regressionsmodells

1 Einleitung

1.1 Einordnung

Ob in Technik, Wirtschaft, Gesellschaft oder Politik: Die Welt wird immer quantitativer und Daten-orientierter. Viele Fachgebiete streben heute danach, Entscheidungen unter Berücksichtigung der Unsicherheit in den zur Verfügung stehenden Daten zu treffen. Statistik als Fachgebiet ist der Schlüssel dazu und bietet die Anleitungen, wie man geeignet Daten sammelt, sie analysiert und unter Unsicherheit Schlüsse daraus zieht oder Entscheidungen fällt. Sie umfasst Methoden, mit denen in Daten Strukturen gesucht und erkannt werden können, um Unternehmens-, Produktions- und Marketingprozesse zu verbessern und zu überwachen. (BFS, 2009)

Der Statistik haftet da und dort der Ruf an, manipulierbar zu sein («Ich traue keiner Statistik, die ich nicht selbst gefälscht habe.» oder «Mit Statistik lässt sich alles beweisen.»). Diese Auffassung ist falsch. Bei der Statistik handelt es sich um eine präzise Wissenschaft mit starker Anbindung an die Mathematik und Informatik; die Grenzen sind fliessend.

Sie findet heute Anwendungen in praktisch allen empirischen Wissenschaften, von den Wirtschaftswissenschaften über die Medizin und Psychologie bis hin zu den Sprachwissenschaften zur Beschreibung und Beurteilung der erhobenen oder gemessenen Daten.

1.2 Begriffe

Der Ursprung des Begriffs Statistik geht auf das 18. Jahrhundert zurück, als man darunter vorwiegend die «Lehre von den Staatsmerkwürdigkeiten» verstand, insbesondere die Beschreibung und Sammeln von Daten zur Anzahl der Einwohner, der Soldaten, des Steueraufkommens (Achenwall, 1743).

Heute wird der Begriff in drei Bedeutungen verwendet. Einerseits werden einfache Datensammlungen als Statistik bezeichnet (z. B. Meldestatistik, Unfallstatistik). Gemeint sind damit Zahlenkolonnen, Tabellen oder auch grafische Darstellungen mit denen Datenmengen zusammengefasst werden.

Unter Statistik wird – in einer weiteren Bedeutung – die statistische Methodenlehre mit den Methoden der mathematischen Statistik, einem Teilgebiet der angewandten Mathematik subsummiert. In der Universität oder Fachhochschule besucht man das Fach «Statistik».

Und zuletzt wird innerhalb der mathematischen Statistik eine Schätzfunktion als Statistik bezeichnet (z. B. die Chi-Quadrat-Statistik). Damit gemeint ist dann der auf der Basis von vorhandenen empirischen Daten ermittelte Schätzwert der Funktion.

In jüngerer Zeit haben sich im Umfeld der Statistik und der Datenanalyse weitere Begriffe etabliert, vor allem aus dem Bereich der Computerwissenschaften und Informatik. Oft gehört ist der Begriff «Data Mining». Darunter versteht man das Analysieren grosser Datensätze mit dem Ziel, Muster und Beziehungen zu erkennen. Mit Hilfe von Data-Mining-Techniken und -Tools sollen zukünftige Trends vorhergesagt und fundierte, zahlenbasierte Geschäftsentscheidungen getroffen werden können.

Ebenfalls eng verwandt ist das Thema «Machine Learning» (ML). Beim maschinellen Lernen geht es primär darum, bekannte Muster vom Computer automatisch in neuen Daten wiederzuerkennen. In ähnlicher Bedeutung wird auch der Begriff «Artificial Intelligence» (AI) verwendet. Alle diese Begriffe sind allerdings nicht scharf definiert und werden in Literatur und Presse oft synonym oder mindestens mit stark überlappender Bedeutung verwendet.

Im Umfeld der quantitativen Wirtschaftswissenschaften hat sich der Begriff «Ökonometrie» etabliert. Dieses Fachgebiet fokussiert auf die statistische Modellierung und quantitative Analyse ökonomischer Phänomene. Die Statistik ist daher neben der Wirtschaftstheorie die wichtigste Komponente der Ökonometrie, deren materieller Teil als empirische Wirtschaftsforschung oder angewandte Ökonometrie bezeichnet wird. Die Ökonometrie fusst stark auf Regressions- und Klassifikation-Methoden.

1.3 Gebiete der Statistik

Das Themengebiet der Disziplin «Statistik» gliedert sich in verschiedene Teilbereiche. Neben der deskriptiven (oder «beschreibenden») und induktiven (oder «schliessenden») Statistik wird heute auch die explorative Statistik dem Fach zugerechnet. Die Wahrscheinlichkeitsrechnung bildet die Grundlage für die induktive Statistik.

Die deskriptive Statistik befasst sich mit der Aufbereitung und Darstellung von Daten mit dem Ziel, Zusammenhänge und Abhängigkeiten zu erkennen und zu messen. Die explorative Statistik geht auf die Konzepte von John W. Tukey[1] zurück und begreift die Exploration als eine der drei Grundaufgaben der Statistik. Tukey legte besonderen Wert auf die möglichst direkte Darstellung und auf das Verständnis empirischer Daten (im Gegensatz zu theoretischen Modellen). Die Abgrenzung zwischen deskriptiver Statistik und explorativer Datenanalyse ist aber nicht wirklich scharf. Die explorative Datenanalyse ist eher konzipiert zur Suche nach Strukturen und Besonderheiten in den Daten und verwendet hierbei auch Ansätze der induktiven Statistik. Sie kann so zu neuen Fragestellungen und Hypothesen in den jeweiligen Anwendungen führen. Sie wird daher typischerweise eingesetzt, wenn die Fragestellung nicht genau definiert ist oder auch die Wahl eines geeigneten statistischen Modells unklar ist.

Die Methoden der deskriptiven Statistik umfassen eine Vielzahl von statistischen Instrumenten, um Zahlen quantitativ zu beschreiben. Um die zentrale Lage zu repräsentieren, verwendet man z. B. oft den arithmetischen Mittelwert (umgangssprach-

1 John Wilder Tukey (1915–2000), US-amerikanischer Statistiker

lich: «Durchschnitt»). In Ergänzung dazu gibt die Standardabweichung an, wie gross die Streuung einer Zahlenreihe ist. Häufigkeitsverteilungen zeigen detailliertere Muster innerhalb der Zahlen und können gut grafisch mit Diagrammen visualisiert werden.

Die schliessende Statistik umfasst Methoden, mit denen von einer Stichprobe auf die Gesamtheit geschlossen werden kann. Damit sind statistische Schätz- und Testverfahren gemeint, mit denen vermutete Strukturen bestätigt oder verworfen werden können. Die Grundlage hierfür bildet die Wahrscheinlichkeitsrechnung, die die theoretischen Modelle beschreibt, die dann in der induktiven Statistik für die Beschreibung der realen Fragestellungen verwendet werden.

Die beiden Begriffe können auch mit folgendem Ansatz gut auseinandergehalten werden. Die Wahrscheinlichkeitsrechnung geht von einem typischerweise vereinfachten, reduzierten Modell aus und quantifiziert die Wahrscheinlichkeiten für bestimmte vom Zufall beeinflusste Beobachtungen. Die induktive Statistik geht von der Beobachtung aus und beschreibt darauf basierend die (unbekannte) Grundgesamtheit möglichst treffend. Zu wissen, wie sicher oder unsicher Ihre Schätzungen sind, ist ein wichtiger Bestandteil der Statistik.

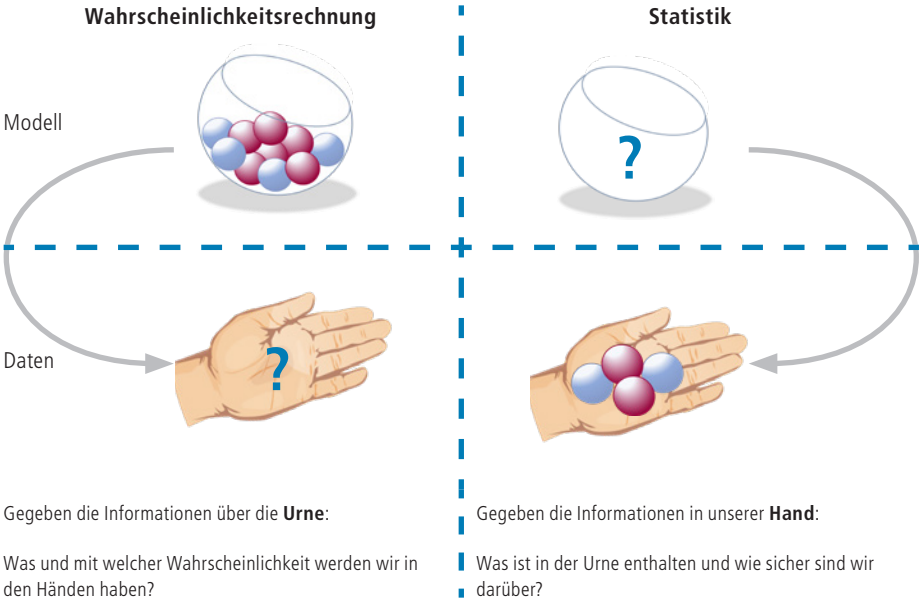


Abb. 1-1: Begriffe «Wahrscheinlichkeitsrechnung» versus «Statistik», (nach Meier, 2020)

Hie und da begegnet man der Bezeichnung «Stochastik» als übergeordneten Begriff für die Wahrscheinlichkeitsrechnung und Statistik. Neben der mathematischen Modellierung zufälliger Ereignisse mit Hilfe der Wahrscheinlichkeitsrechnung umfasst die Stochastik damit auch den Umgang und die Auswertung von Datensätzen (Statistik). Die Statistik ist also so gesehen ein Teilbereich der Stochastik.

1.4 Geschichte der Statistik

Die Anfänge der Statistik liegen weit zurück. Bereits die ersten Bauern in Mesopotamien, schätzten, ob genügend Getreide für den Winter vorrätig war. Im alten Ägypten waren die Pharaonen an den Mengen Korn interessiert, die die Landwirtschaft zu produzieren in der Lage war und erhoben dazu erste Daten. Diese Art der Fragestellung zog sich dann durch die weitere Geschichte bis ins 16. Jahrhundert, wo die ersten ernsthafteren Entwicklungen der Wahrscheinlichkeitsrechnung erfolgten.

Wirklichen Aufschwung erlebte die Statistik dann aber erst am Anfang des 20ten Jahrhunderts, als fundamentale Zusammenhänge und Theorien entwickelt wurden. So gesehen ist die Statistik eine vergleichsweise junge Wissenschaft.

Eine gute Übersicht findet sich in der folgenden Zeitstrahl-Grafik[2], die mit den mathematischen Grundlagen im 16ten Jahrhundert startet und die wichtigen Eckpunkte und Meilensteine der Entwicklung im weiteren Verlauf aufführt.

1.5 Bedeutung der Statistik

Gegenwärtig verändern gerade riesige Mengen digitaler Daten die Welt und damit auch die Art und Weise, wie wir in ihr leben. Wenn wir Buchempfehlungen von Amazon, basierend auf unseren letzten Einkäufen erhalten, wenn Netflix weiterführende Serien anbietet, Shazam uns aufklärt, welchen Song wir gerade hören, wenn die Autos autonom zu fahren beginnen, wenn Drohnen in Kriegsgebieten ihre Ziele selbstständig auswählen und treffen, ja sogar wenn es in Partnerschaftsbörsen zu einem «Match» kommt, in allen Fällen ist Statistik an den entscheidenden Stellen mit im Spiel. Ihre Bedeutung hat in den letzten Jahrzehnten mit den Fortschritten in der elektronischen Datenverarbeitung (Digitalisierung) und den immer umfangreicheren Datenbeständen enorm zugenommen.

Die Treiber für diese Entwicklung, sozusagen das Benzin für die Datenanalyse, sind die immer günstiger verfügbare Kapazität an Rechenleistung kombiniert mit riesigen Mengen an billigem Speicherplatz für die Lagerung der Daten. Die Anlage «SuperMUC-NG» am Leibniz Rechenzentrum ist ein eindrückliches Beispiel für die Grössenordnungen. Sie schaffte es im 2019 mit einer Rechenleistung von 19 Petaflops in die Top 10 der schnellsten Supercomputer. Ein Petaflops entspricht einer Billiarde Rechenoperationen pro Sekunde. Die Top-Position hielt zu dieser Zeit IBM mit dem Supercomputer «Summit», der es auf 148.6 Petaflops brachte (also nochmals 8-mal mehr).[3]

Wir erleben vor diesem Hintergrund auch eine stürmische Entwicklung der statistischen Methoden. Algorithmen, wie jener der zum Einsatz kommt, wenn Google innerhalb von Sekundenbruchteilen auf ein Stichwort hin entsprechende Links liefert, wurden erst im Laufe der letzten 20–30 Jahre entdeckt und entwickelt[4].

2 https://www.statsref.com/timeline.pdf
3 (onlinepc.ch, 2019)
4 (PageRank – Wikipedia, n. d.)

speziell gekennzeichnet werden[9]. Moderne Systeme können problemlos mit fehlenden Werten umgehen.

Je nach Studie können die gleichen Daten mehrfach erhoben werden, z. B. zu mehreren Zeitpunkten. In solchen Fällen werden die Datensätze typischerweise durch eine weitere Variable (z. B. *Messung*) gekennzeichnet und als weitere Zeile geführt. Mehrfachmessungen sollten nicht nebeneinander als zusätzliche Spalten erfasst werden.

In der Literatur findet man diese Struktur manchmal auch als «lange Form» (im Gegensatz zu «breit») bezeichnet.

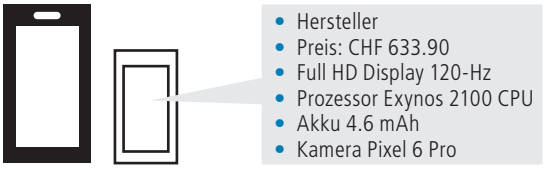
2.5 Datentypen und Skalen

Die weitere Diskussion wird vereinfacht, wenn wir ein paar grundlegende allgemeine Begriffe und Bezeichnungen einführen.

Generell interessieren wir und für bestimmte *Objekte*. An diesen Objekten beobachten wir *Eigenschaften*, die wir allgemein als *Merkmale* bezeichnen. Ein Merkmal kann von Objekt zu Objekt unterschiedliche Formen oder Grade annehmen (*Ausprägung*).

Beispiel

Wenn für einen Preisvergleich die 30 besten Handys 2023 geprüft werden, dann entsprechen die unterschiedlichen Handys den Objekten. An diesen beobachtet man relevante Merkmale wie den Preis, die Leistungsfähigkeit des Chips, den Kameratyp, die Grösse und den Typ des Displays, Akkulaufzeit, Durchsatz des Modems, Garantiekonditionen, etc.



Im Umgang mit Daten hat sich auch der Begriff *Variable* anstelle von Merkmal eingebürgert. Betrachtet man nur eine abstraktere Menge an Elementen, Zahlen oder Begriffe, spricht man von einem *Datenvektor*.

Der Charakter einer Variable ist für die Analyse entscheidend. Sowohl die Wahl von passenden Kennzahlen für die deskriptive Beschreibung wie auch die Wahl allfälliger statistischer Modelle hängt massgeblich davon ab.

Variablen besitzen ein bestimmtes *Skalenniveau*. Abhängig von diesem weisen sie einen unterschiedlich detaillierten Informationsgehalt auf. Viele Variablen verfügen über eine naturgegebene Skala. Um beim Handy-Beispiel zu bleiben, ist die Variable «Hersteller» eine *qualitative* Variable. Rechnen kann man damit nicht. Die «Akku-

9 wie dies in Vergangenheit bei bestimmten Statistik-Systemen, wie SPSS, üblich war.

laufzeit» hingegen ist eine *quantitative* Variable, von der man z. B. den Mittelwert bestimmen kann.

Die wichtigste Unterteilung der Variablentypen ist die Unterscheidung zwischen qualitativen und quantitativen Variablen.

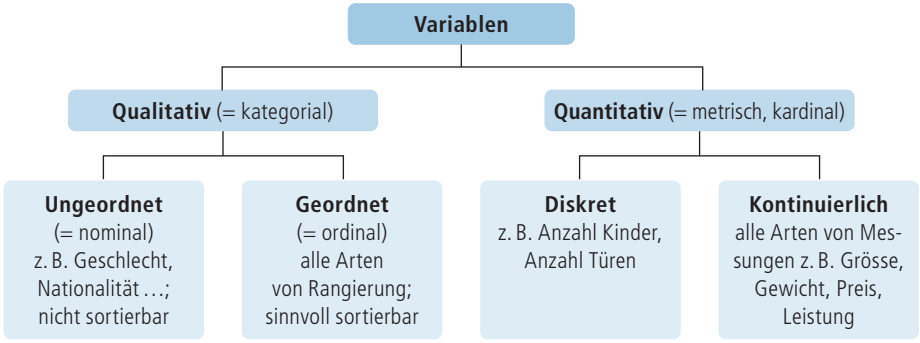


Abb. 2-3: Variablentypen und -skalen

Qualitative Variablen bezeichnet man auch als *kategorial* und quantitative Variablen als *metrisch* (oder auch *kardinal*).

Kategoriale Variablen lassen sich weiter unterteilen in solche, deren Kategorien geordnet (*ordinale* Variablen) und solchen deren Kategorien ungeordnet sind (*nominale* Variablen). So wäre z. B. die Qualität des Handy-Displays geordnet (von gut bis schlecht), während der Hersteller ungeordnet wäre (keine sinnvolle, inhaltliche Reihenfolge vorhanden). Eine Variable ist ordinal, wenn die Kategorien logisch von der kleinsten bis zur grössten in einem für die vorliegende Frage sinnvollen Sinne geordnet werden können (wobei unsinnige Ordnungen wie das Alphabet ausschliessen); andernfalls ist sie ungeordnet.

Ein für die Praxis wichtiger Spezialfall sind Variablen, die genau zwei Ausprägungen haben. Eine solche könnte die Variable Rabatt sein. Sie enthält nur die Information WAHR, falls ein Rabatt gewährt wurde und FALSCH, wenn nicht. Solche Variablen werden auch als *dichotome* oder *binäre* Variablen bezeichnet. Die technische Kodierung kann auf unterschiedliche Weise erfolgen, beim Geschlecht wird meistens mit M und W kodiert, wenn's darum geht das Kündigungsverhalten zu beschreiben wird 0 und 1 oder wahr/falsch verwendet.

Quantitative Variablen werden danach unterschieden, ob die Variable alle Zwischenzahlenwerte (alle reellen Zahlen) oder nur bestimmte Werte (z. B. ganze Zahlen) annehmen kann. Erstere bezeichnet man als *kontinuierliche (oder stetige)* Variablen. Zuweilen spricht man auch von *Messdaten*. Die Akkulaufzeit ist beispielsweise eine kontinuierliche Grösse, man kann sie mit Hilfe einer Uhr gut messen. Die Anzahl Kameras ist demgegenüber eine diskrete Variable. Ein handelsübliches Handy hat entweder 0, 1, 2 oder 3 Kameras. Handys mit 1.123 Kameras gibt's nicht. Solche Variablen werden auch als *Zähldaten* bezeichnet, weil sie oft durch Zählprozesse entstehen. Obwohl man bei Messdaten so tut, als wäre jede reelle Zahl als

Rahmen eines Excel-Grundkurses nachholen. Im weiteren Verlauf dieses Buchs werden diese Grundkenntnisse vorausgesetzt.

Folgende knappe Übersicht umfasst die wichtigsten für die Datenorganisation nützlichen Funktionen:

Name	Beschreibung
SUMME(…)	Berechnet die Summe eines Zellbezugs
ANZAHL()	Berechnet wie viele Zellen in einem Bereich **Zahlen** enthalten, Texte, Wahrheitswerte etc. werden ignoriert.
ANZAHL2()	Zählt die Anzahl nicht leerer Zellen in einem Bereich (z. B. Texte)
ANZAHLLEEREZELLEN()	Liefert die Anzahl leere Zellen
MAX(), MIN()	Liefert den grössten, resp. den kleinsten Wert eines Zellbezugs
KGRÖSSTE(), KKLEINSTE()	Liefert die k-grössten, resp. k-kleinsten Elemente eines Zellbereichs
WENN()	Prüft eine Bedingung und liefert je nach Ergebnis entweder den ersten oder zweiten Wert
SUMMEWENN()	Summiert einen Zellbereich in Abhängigkeit einer Bedingung
ZÄHLENWENN()	Zählt Elemente eines Zellbereich in Abhängigkeit einer Bedingung
RANG.MITTELW()	Berechnet Ränge nach der in der Statistik üblichen Mittelrang-Methode
XVERWEIS()	Die XVERWEIS-Funktion durchsucht einen Bereich oder eine Anordnung und gibt dann das Element zurück, das der ersten gefundenen Übereinstimmung entspricht.

2.6.2 Ränge

Viele Methoden in der modernen Statistik verwenden Ränge. Für die Berechnung von Rängen muss eine Eigenheit beachtet werden, die die Verwendung von Rängen in der Statistik von jener beispielsweise im Sport unterscheidet. Die Sportrangierung sieht üblicherweise für Athleten mit der gleichen Leistung die gleichen Ränge vor, meistens den jeweils höheren.

Für numerische Verfahren hingegen ist es von Bedeutung, dass die Rangsumme für eine bestimmte Stichprobengrösse konstant bleibt. Bei der Sportrangierung weist die Skala Leerstellen auf, wenn gleiche Ränge[11] auftreten (z. B.: nicht vergebene Bronzemedaille bei zwei zweiten Plätzen). Um die Rangsumme zu bewahren, müssen Ränge in der Statistik deshalb als mittlere Ränge bestimmt werden, anstatt einzelne Ränge zu überspringen.

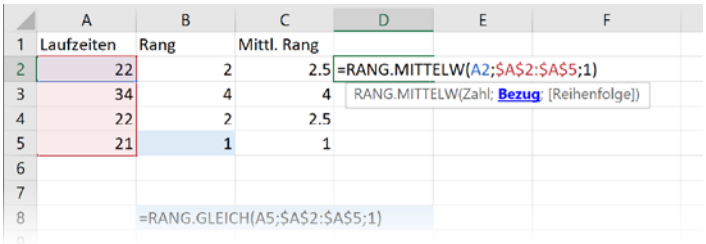


Abb. 2-8: Mittlere Ränge anstatt Sportrangierung

11 Gleiche Werte in einer Variablen werden auch Bindungen (engl: ties) genannt.

Die mittleren Ränge bestimmen sich als Mittelwert der zur Debatte stehenden Ränge, d. h. wenn 2 Läufer auf Rang 2 zu liegen kommen, würde ihr mittlerer Rang (2 + 3) / 2 = 2.5 betragen.

Die Funktion **RANG.MITTELW()** liefert die so berechneten Ränge.

2.6.3 Namen für Zellbereiche

In der Datenanalyse werden meistens rechteckige Datenstrukturen wie oben bereits eingeführt verwendet. Für solcherart organisierte Daten bietet Excel die Möglichkeit, Zellbereiche mit sprechenden Namen zu bezeichnen. Auf diese Namen kann anschliessend in Funktionen oder Berechnungen alternativ zu den sonst verwendeten Zellbezügen verwiesen werden.

Für die Datenanalyse hat das den Vorteil, dass ganze Variablen mit einem Namen angesprochen werden können. Dadurch vermeidet man langes und unkontrolliertes Scrolling mit der Maus über lange Tabellen, wenn Kennzahlen berechnet werden sollen. Bei grossen Datensätzen erweist sich dies als grosse Erleichterung und als fehlervermeidenden, robusten Ansatz.

Um die Namen der Variablen zu definieren wird als erstes der gewünschte Datenbereich markiert. Das geht am besten so, dass man den Zellcursor innerhalb des Listenbereichs platziert und die Tastenkombination <Ctrl+A> drückt. Damit wird der ganze, von leeren Spalten und Zeilen begrenzte, Listenbereich um den Zellcursor herum selektiert. Danach wählt man den Menüpunkt **Formeln | Aus Auswahl erstellen**.

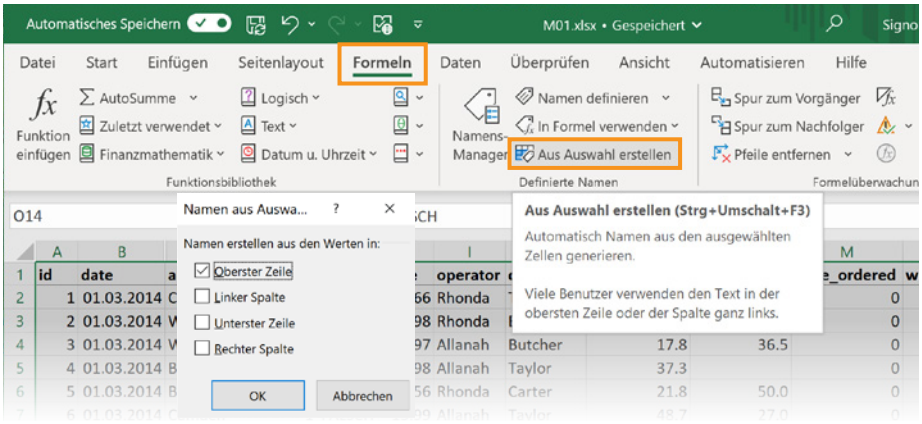


Abb. 2-9: Namen aus Auswahl erstellen

Im erscheinenden Dialog wird die Option «Oberster Zeile» gewählt. Klick auf OK beendet den Prozess. Unmittelbar danach stehen die neu definierten Namen in der Namensliste zur Verfügung. Werden Sie dort ausgewählt, wird der zugehörige Zellbereich im Tabellenblatt markiert.

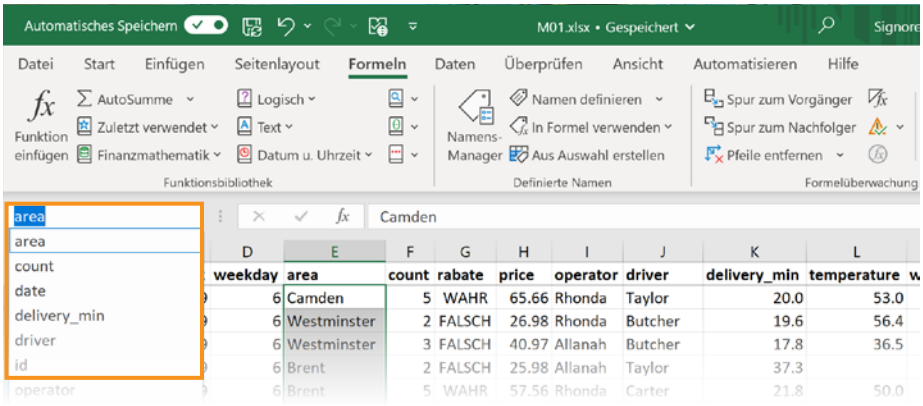


Abb. 2-10: Namensfeld in der Excel-Benutzeroberfläche

Die Namen können jetzt auch in Funktionen verwendet werden. Soll die Summe der Variable *price* berechnet werden, kann **=SUMME(price)** direkt in eine Zelle eingetragen werden.

2.6.4 Text in Spalten

Beim Arbeiten mit Daten ergibt sich oft die Situation, dass man den Text in einer Zelle auf mehrere aufteilen möchte, z. B. wenn ein in ein Tabellenblatt eingefügter Text Trennzeichen enthält (**benzin.txt**).

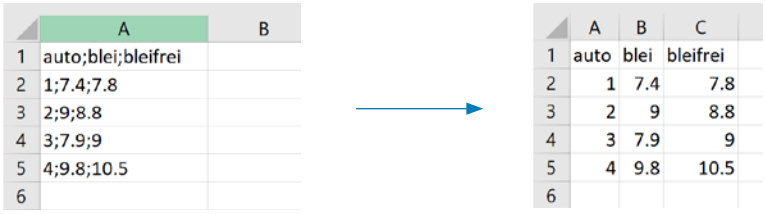


Abb. 2-11: Text mit Trennzeichen in Spalten aufteilen

Hierzu geht man wie folgt vor:

- Zellen auswählen, deren Inhalt aufgeteilt werden soll.
- Menü **Daten | Datentools | Text in Spalten** wählen:

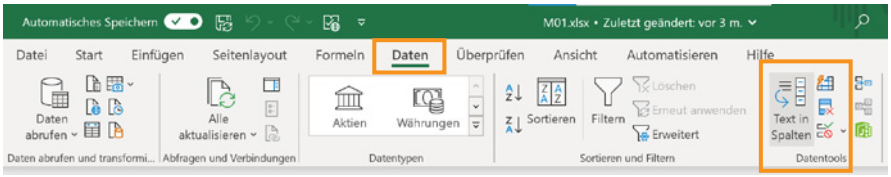


Abb. 2-12: Menü Text in Spalten

d) Fügen Sie dem Datensatz eine neue Variable mit dem Namen *popdens* hinzu, in der Sie die Einwohnerdichte (Einwohner pro Fläche) für jeden Staat berechnen.

e) Erzeugen Sie eine neue Variable *rnk_frost*, die den Rang der Staaten nach der Anzahl der Frosttage (*frost*) enthält. Der Rang lässt sich mit Hilfe der Excel-Funktion **RANG.MITTELW()** berechnen.Überlegen Sie sich, warum der Rang des ersten Staates (Alabama) mit 5.5 angegeben wird.

f) Verwenden Sie die Operatoren > oder < um eine neue Kategorie-Variable zu erzeugen (z. B. *clima*). Es sollen die «kalten» von den «warmen» Staaten unterschieden werden. Definieren Sie das Kriterium als:
warm: frost < 70 und kalt: frost ≥ 70.

Tipp Verwenden Sie die Funktion **WENN()**.

g) Erzeugen Sie einen neue Spalte *cold_rich*, die die Staaten, die sowohl über ein mittleres Einkommen von mehr als 5'000 $ verfügen als auch unter mehr als 120 Frosttage leiden, enthält.

Tipp Verwenden Sie die Funktion **WENN()** und **UND()**.

h) Verwenden Sie verschachtelte **WENN()**-Funktionen, um eine Kategorie mit mehreren Ausprägungen zu erzeugen. Die Staaten mit bis zu 49 Frost-Tagen sollen als «warm», die mit 50–150 Frosttagen als «kalt», die mit mehr als 150 Frost-Tage als «eisig» bezeichnet werden.

i) Zeichen Sie ein Punkt XY-Diagramm, das den Zusammenhang zwischen Einkommen und Frosttage beleuchtet. Die Anzahl der Frosttage sollen auf der x-Achse und das Einkommen auf der y-Achse aufgetragen werden.

j) Über welchen Bereich spannt sich die Variable Einkommen (*income*). Lassen Sie sich den grössten und den kleinsten Wert ausgeben.

Tipp Funktionen **MIN()**, **MAX()**.

k) Lassen Sie sich mit den Funktionen **KGRÖSSTE()**/**KKLEINSTE()** die flächenmässig jeweils 3 grössten und die 3 kleinsten Staaten ausgeben.

l) Überprüfen Sie mit der Funktion, ob alle Variablen komplett befüllt sind oder ob Daten fehlen.

Tipp **ANZAHLLEEREZELLEN()**.

m) Suchen Sie einen Ansatz, wie man mit der Funktion **ZUFALLSZAHL()** zufällig 5 Staaten aus dem Datensatz auswählen könnte.

n) Wie viele Staaten und wie viele Einwohner haben die einzelnen Regionen? Bestimmen Sie alle Ausprägungen der Variable region.

Tipp **EINDEUTIG()**

Berechnen Sie die Anzahl und erzeugen Sie ein Säulendiagramm (Barplot).

Tipp **SUMMEWENN()**, **ZÄHLENWENN()**

3.6.11 Radar

Auf Antrag lärmgeplagter Bewohner wurde in der Stadt Zürich in einem Quartier eine Radar-Anlage aufgestellt. Die gemessenen Geschwindigkeiten finden sich in der Datei **radar.xlsx**.

In Fragestellungen wie diesen sind speziell die Anteile von Bedeutung. Es interessiert besonders, wie viele Lenker zu schnell fuhren, aber auch wie viele Lenker in einem interessierenden Zwischenbereich fuhren.

a) Erzeugen Sie die zugehörige empirische kumulative Verteilungsfunktion. Verwenden Sie die einzelnen beobachteten Werte als Basis für die Häufigkeiten wie in der folgenden Tabelle:

	x	h_i	H_i	F_i
1	35	1	1	2 %
2	39	1	2	4 %
3	43	2	4	8 %
…				

b) Erzeugen Sie die entsprechende Grafik!

c) Wie hoch ist der Anteil der Fahrer, die höchstens die vorgeschriebene Geschwindigkeit von 50 km/h fuhren? Schätzen Sie grob anhand der Grafik.

d) Wie hoch ist der Anteil der Fahrer, die mit einer Geschwindigkeit von 60 km/h oder mehr fuhren?

e) Wie hoch ist der Anteil der Fahrer, die mit einer Geschwindigkeit zwischen 55 km/h und 65 km/h fuhren?

f) Schätzen Sie anhand der Grafik das 20 %-Quantil, das 80 %-Quantil und das 90 %-Quantil der gemessenen Geschwindigkeit.

3.6.12 Klassen-Review

Die folgenden Klasseneinteilungen haben alle einen oder mehrere Konstruktionsfehler. Finden Sie diese!

a)

Klassengrenzen	**27–32**	**33–38**	**39–44**	**45–49**	**50–55**
Anzahl	33	44	53	25	10

b)

Klassengrenzen	**5–9**	**9–13**	**13–17**	**17–20**	**20–24**
Anzahl	3	4	5	2	0

c)

Klassengrenzen	**23–27**	**28–32**	**38–43**	**42–47**	**48–52**
Anzahl	9	3	2	5	2

3.6.13 Körpergrösse

Die Körpergrösse ist ein zuweilen entscheidender Einflussfaktor im menschlichen Leben. So zeigen etliche Studien z. B. einen Zusammenhang der Körpergrösse mit beruflichem und politischem Erfolg (Hanna & Gerhard, 2021).

Die Datei **heights.xlsx** enthält die Körpergrösse für eine Auswahl von bekannten Persönlichkeiten aus Sport, Politik und Gesellschaft (ohne allerdings den Anspruch auf Repräsentativität zu erheben).

a) Erzeugen Sie ein Histogramm für eine Klassenbreite von 5 cm. Gibt es Ausreisser?

b) Stellen Sie die Summenhäufigkeitsverteilung bezüglich der Klasseneinteilung von (a) graphisch dar.

c) Bestimmen Sie die Anzahl der Männer und der Frauen, sowie das Total der Personen im Datensatz und berechnen Sie die relativen Häufigkeiten.

d) Bestimmen Sie die Ausprägungen der Variable *group* und geben Sie an, wie viele Personen sich in der jeweiligen Gruppe befinden.

3.7 Lernkontrolle auf Moodle

Zu diesem Modul ist eine formative Lernkontrolle auf Moodle verfügbar. Über einen Klick auf den Button rechts gelangen Sie zum entsprechenden Test. (Für Registrierung siehe Anleitung im Anhang.)

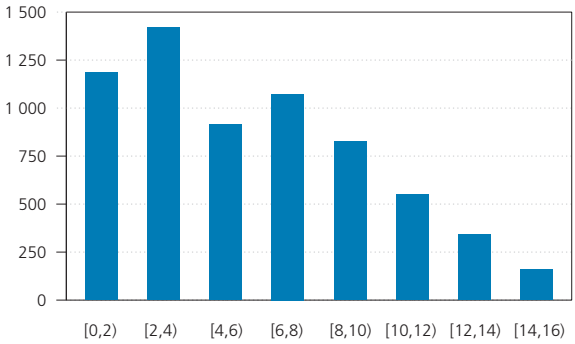


Abb. 4-18: Altersverteilung Fahrzeuge Autoscout24

Um diese Eigenschaft zu beschreiben, gibt es die Kennzahl *Schiefe* (engl. *skewness*), die den Grad der Abweichung von der Symmetrie beschreibt.

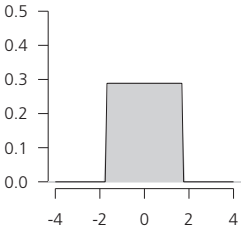
Eine gängige Definition der Schiefe ist:

$$\frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{4-10}$$

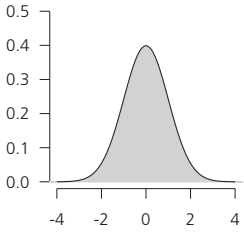
Verteilen sich die Beobachtungen gleichmässig um die Mitte, so spricht man von einer symmetrischen Verteilung, die Schiefe wird 0. Für rechtsschiefe Verteilungen ist die Schiefe positiv, für linksschiefe negativ.

Für die Berechnung mit Excel existiert die Funktion:

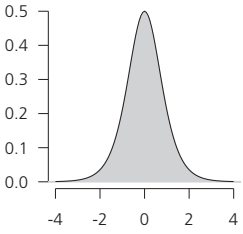
Kurtosis Neben der Schiefe einer Verteilung existiert auch eine Kennzahl zur Beschreibung der Wölbung (spitzig «leptokurtisch» oder breitgipflig «platykurtisch») einer Verteilung. Die *Kurtosis* (engl.: kurtosis) gibt an, ob eine Verteilung sich spitziger aufwölbt als die Normalverteilung[36] oder weniger spitzig. Je grösser der Wert wird, umso spitziger ist die Verteilung.



Uniform(min = $-\sqrt{3}$ max = $-\sqrt{3}$)
kurtosis = –1.2



Normal($\mu = 0$, $\sigma = 1$)
kurtosis = 0



Logistic($\alpha = 0$, $\beta = 0.5$)
kurtosis = 1.2

36 Siehe Modul 11 «Kontinuierliche Verteilungen»

7.9 Übungen

7.9.1 Wohnung

Zur Beschreibung des Wohnungsmarktes in Zürich wurde erfasst, ob die entsprechende Wohnung einen Balkon hat (Ereignis B), einen Glaskeramik-Herd hat (H) und ob die Wohnung einen Garagenplatz hat (Ereignis G).

Stellen Sie die nachfolgenden Ereignisse durch geeignete Verknüpfungen der Ereignisse B, H und G dar.

Eine Wohnung besitzt …

a) … einen Balkon und einen Garagenplatz

b) … einen Garagenplatz aber keinen Balkon

c) … weder Balkon noch Glaskeramik-Herd

d) Welche Wohnungen sind durch folgende Ereignisse beschrieben:

H / G, $H^C \cap G^C$, $B^C \cup H^C$, $G \cap (B \cup H)^C$

7.9.2 Roulette

Beim Roulette landet eine Kugel auf einer der Zahlen {0, 1,... 36}. Der Spieler kann auf einzelne Zahlen oder auf einfache Chancen setzen, also auf pair (engl: *even*) d. h. gerade Zahlen in {1, ..., 36}, oder impair (engl: *odd*), das Komplement (wieder ohne 0), *analog manque* {1, ..., 18} bzw. *passe* {19, ..., 36}, resp. *rouge* {1, 3, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, 36} bzw. *noir* die restlichen nicht roten Zahlen in {1, ..., 36}. Die Null (*zero*) ist grün!

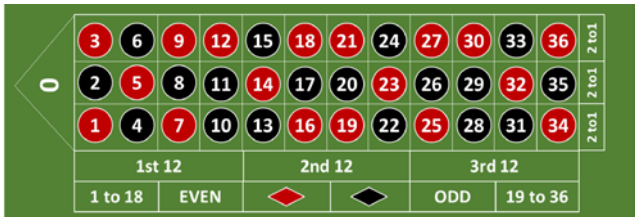


Abb. 7-13:
Tableau für Europäisches Roulette

a) Wie gross ist die Wahrscheinlichkeit, dass eine rote Zahl gewinnt?

b) Wie gross ist die Wahrscheinlichkeit, dass keine schwarze Zahl gewinnt?

c) Wie gross ist die Wahrscheinlichkeit, dass eine der Zahlen von 1 bis 12 gewinnt?

d) Wie gross ist die Wahrscheinlichkeit, dass eine rote oder eine schwarze Zahl gewinnt?

e) Angenommen eine rote Zahl aus dem Bereich von 1 bis 12 gewinnt, wie gross ist die Wahrscheinlichkeit, dass es sich um eine gerade Zahl handelt?

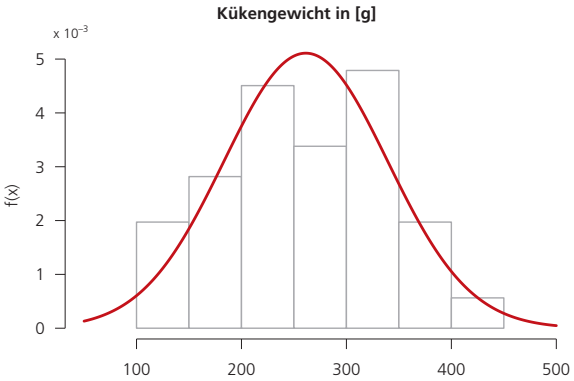


Abb. 15-6: Anpassung an eine Normalverteilung

Wir prüfen die Hypothese mit einem Anpassungstest. Mit den in der Grafik nahegelegten Klassengrenzen bestimmt man die beobachteten Häufigkeiten b_i (hier = h_i).

Um die erwarteten Häufigkeiten e_i zu erhalten, berechnet man die Wahrscheinlichkeiten p_{norm_i}, dass eine $N(\mu, \sigma^2)$-verteilte Zufallsvariable in den Bereich der entsprechenden Klasse fällt. Die Parameter μ und σ werden üblicherweise aus den Stichprobendaten bestimmt, falls keine weiteren Kenntnisse vorliegen.

MW	261.31	=MITTELWERT(weight)			
Std.Abw.	78.07	=STABW.S(weight)			
(von	bis]	h_i	H_i	p_norm_i	e_i
100	150	7	7	7.7%	5.47
150	200	10	17	13.9%	9.88
200	250	16	33	22.6%	16.06
250	300	12	45	=NORM.VERT(E14;E6;E7;WAHR)	
300	350	17	62	-NORM.VERT(D14;E6;E7;WAHR)	
350	400	7	69	9.0% NORM.VERT(x; Mittelwert; Standabwn; Kumuliert)	
400	450	2	71	3.8%	2.69
				100.0%	71.00

(Für die Berechnung des Mittelwerts und der Standardabweichung wurde auf den vorher definierten Bereichsnamen **weight** referenziert.)

Die erwarteten Werte ergeben sich dann aus den Wahrscheinlichkeiten und der totalen Anzahl Werte (hier: n = 71) in der Stichprobe $e_i = p_{norm\_i} \cdot n$.

(von	bis]	h_i	H_i	p_norm_i	e_i
100	150	7	7	7.7%	=H11*G17
150	200	10	17	13.9%	9.88
200	250	16	33	22.6%	16.06
250	300	12	45	24.7%	17.57
300	350	17	62	18.2%	12.93
350	400	7	69	9.0%	6.40
400	450	2	71	3.8%	2.69
				100.0%	71.00

20 Anhang

20.1 Referenzen

Achenwall, G. (1743). Notitia politica vulgo statistica.

Aitkin, M. (1978). The Analysis of Unbalanced Cross-Classifications. Journal of the Royal Statistical Society. Series A (General), 141(2), 195. https://doi.org/10.2307/2344453

Alliance SwissPass. (2019). Das nationale Schwarzfahrerregister im öV nimmt den Betrieb auf. https://www.allianceswisspass.ch/de/asp/News/Newsmeldung?newsid=139

Alphabet und Buchstabenhäufigkeit: Deutsch. (n. d.). Retrieved June 1, 2023, from https://www.sttmedia.de/buchstabenhaeufigkeit-deutsch

Alzheimer Schweiz. (2022). Rückzug von Biogen bzgl. Aducanumab. Medienmitteilung (08.06.2022). https://www.alzheimer-schweiz.ch/de/medien/medienmitteilung-08062022-rueckzug-von-biogen-bzgl-aducanumab

Beecher, H. K. (1959). Measurement of subjective responses: quantitative effects of drugs. In Published in 1959 in New York NY) by Oxford university press. New York (N.Y.) : Oxford university press, 1959. https://lib.ugent.be/catalog/rug01:000071673

BFS, B. für S. (2009). Berufsbild Statistiker Statistikerin. https://www.stat.ch/images/resources/pdfs/StatistikerRZD.pdf

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: data from berkeley. Science (New York, N. Y.), 187(4175), 398–404. https://doi.org/10.1126/SCIENCE.187.4175.398

blutspende.ch. (n. d.). Blutgruppen. Retrieved June 1, 2023, from https://www.blutspende.ch/de/wissen-ueber-blut/blutgruppen

Box, G. E., Hunter, S. J., & Hunter, W. G. (2005). Statistics for Experimenters. In Wiley (Vol. 21, Issue 3). https://doi.org/10.1080/00401706.1979.10489788

Brown, L. D., Cai, T. T., & Das Gupta, A. (2001). Interval estimation for a binomial proportion. Statistical Science, 16(2), 101–117. https://doi.org/10.1214/ss/1009213286

Bryant, P. G., & Smith, M. A. (1995). Practical data analysis: case studies in business statistics.