A. Strukturen und Werkzeuge zur Statistik & Methodenlehre

1 Einführung und Überblick

In diesem Kapitel wird ein Überblick über die Struktur der gesamten Materialien gegeben. Um die Definitionen und Abgrenzungen besser nachvollziehen zu können, seien zwei Beispiele vorangestellt.

1.1 Von der Konzeption zur Auswertung

1.1.1 Konzeption empirischer Erhebungen - field research

Vor jeder statistischen Auswertung basiert auf Daten. Diese können bereits existieren wie z.B. Umsatzdaten aus dem Rechnungswesen eines Unternehmens, die dann i.d.R. nur noch datentechnisch für die Auswertung mittels einer Software aufbereitet werden müssen. In diesem Fall spricht man auch von „**desk research**“. Im anderen Fall existieren keine Daten, so dass diese erst noch in irgendeiner Form – z.B. Fragebogen, Interview usw. -erhoben werden müssen. In diesen Fällen spricht man auch von „**field research**“.

Die Konzeption empirischer Erhebungen ist häufig sehr komplex. Leider beginnen viele Autoren direkt mit der Formulierung der Fragen, ohne den Gesamtkontext zu strukturieren. Die Nachteile dieser Vorgehensweise zeigen sich dann häufig bei der Auswertung, weil bestimmte Fragestellungen im Vorfeld nicht bedacht und die notwendigen Daten nicht erhoben wurden. Die mangelhafte Herleitung der Fragen entspricht darüber hinaus auch nicht dem üblichen wissenschaftlichen Standard.

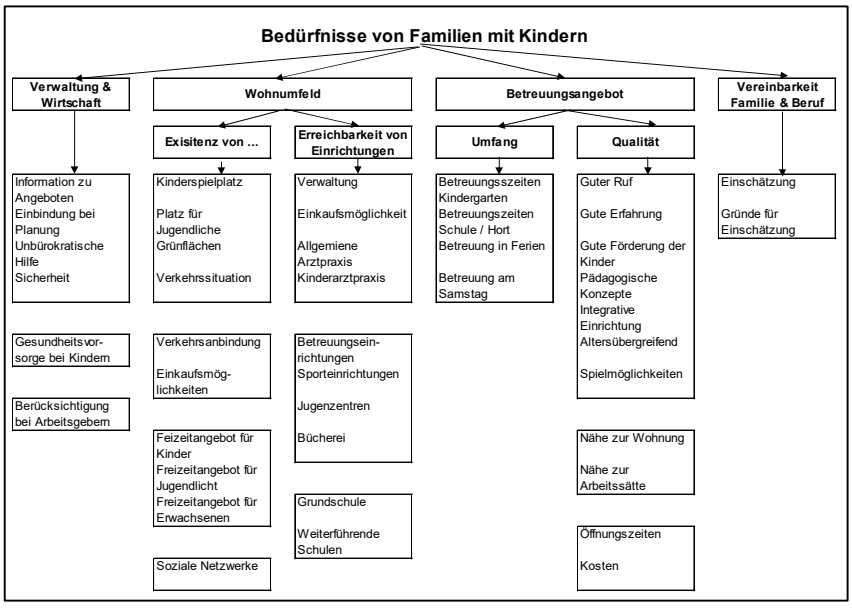
Bei einer exakten wissenschaftlichen Vorgehensweise wird zuerst der zu untersuchende Zusammenhang in Form einer **Hypothese** formuliert. Üblich sind dabei so genannter je-desto-Hypothesen, die aus einer **unabhängigen** und einer **abhängigen Variablen** bestehen. Bei der folgenden Hypothese sind die Bedürfnisse von Familien mit Kindern die unabhängige Größe, von der die Einschätzung der Familienfreundlichkeit abhängig ist.

Hypothese

Je mehr die Bedürfnisse von Familien mit Kinder berücksichtigt werden

desto familienfreundlicher wird eine Kommune eingeschätzt

Diese Hypothese wird dann in ihre einzelnen Bestandteile differenziert, was als die **Operationalisierung** einer Hypothese bezeichnet wird. Bei dem nachstehenden Beispiel wurde die unabhängige Variable „Bedürfnisse von Familien mit Kindern“ auf der **ersten Ebene** in die Aspekte „Verwaltung und Wirtschaft“, „Wohnumfeld“, „Betreuungsangebot“ und „Vereinbarkeit Familie und Beruf“ differenziert. Die weitere Operationalisierung ist der nachstehenden Grafik zu entnehmen.



In dieser Phase einer empirischen Untersuchung ist die Darstellung des gesamten Kontextes wesentlich, da hierdurch alle Aspekte und alle möglichen Fragenstellungen dargestellt werden. Häufig ist es aus Gründen des Umfanges nicht möglich, den gesamten Kontext in einem Interview oder Fragebogen zu erheben. Notwendig ist dann aber die Beschränkung auf bestimmte Aspekte und deren Gründen für Dritte nachvollziehbar offen zu legen. Nur dieses entspricht einer exakten wissenschaftlichen Vorgehensweise.

Erst im Anschluss an diese Operationalisierung erfolgt die Formulierung der Fragen. Die nachsehende Abbildung zeigt dieses exemplarisch für die obige Problemstellung.

2. Wie stehen Sie zu folgenden Aussagen Ihre Gemeinde/ nähere Umgebung betreffend?

	stimme voll und ganz zu	stimme eher zu	Unentschieden	stimme eher nicht zu	stimme überhaupt nicht zu
Meine Gemeinde ist kinderfreundlich	❑	❑	❑	❑	❑
Ich fühle mich über Angebote für Familien gut informiert	❑	❑	❑	❑	❑
Ich fühle mich mit meiner Familie in meiner Gemeinde sicher	❑	❑	❑	❑	❑
Bei Planungen der öffentlichen Hand werden die Bedürfnisse von Familien berücksichtigt	❑	❑	❑	❑	❑
Auf den Ämtern wird bei Familienangelegenheiten unbürokratisch geholfen	❑	❑	❑	❑	❑
Die Gesundheitsvorsorge für meine Kinder ist gut organisiert	❑	❑	❑	❑	❑
Der Arbeitgeber nimmt Rücksicht auf die Bedürfnisse von Familien	❑	❑	❑	❑	❑
Das Spektrum geeigneter Freizeitangebote für Kinder und Jugendliche ist gut	❑	❑	❑	❑	❑

1.1.2 Realisierung empirischer Erhebungen

Datenerhebung

Ausgangspunkt ist i.d.R. ein Fragebogen. Die nachstehende Abbildung ist ein Auszug einer anderen Untersuchung, in diesem Fall einer Befragung von Kommunen zum demographischen Wandel. Die Abbildung zeigt einen Ausschnitt des Fragebogens. Die Bedeutung der Spalten „**Var.**“ für die Numerierung der Variablen und „**Cod.**“ für die Kodierung der Antwort erschließt sich erst in den folgenden Absätzen.

1. Fragenblock: Allgemeine Fragen zum demographischen Wandel

	gar nicht wichtig 1	2	3	4	sehr wichtig 5	Var.	Cod.
1.1. Wie wichtig ist das Thema „Demographischer Wandel“ zurzeit für Ihre Kommune?	❑	❑	❑	❑	❑	1	1-5

	vollkommen unzureichend 1	2	3	4	vollkommen ausreichend 5		
1.2. Wird das Thema „Demographischer Wandel“ in der Gesellschaft ausreichend behandelt?	❑	❑	❑	❑	❑	2	1-5
1.3. Wird das Thema „Demographischer Wandel“ in Ihrer Kommune ausreichend behandelt?	❑	❑	❑	❑	❑	3	1-5

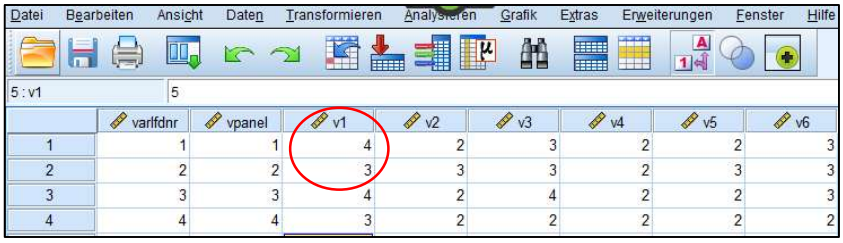
	gar nicht 1	2	3	4	vollkommen 5		
1.4. Ist die Bundesebene in der Lage, mit den Auswirkungen des dem Bitte sagen Sie das noch einmal umzugehen?	❑	❑	❑	❑	❑	4	1-5

Datenerfassung

Nach der Durchführung der Befragung erfolgt die statistische Auswertung. In der Regel verwendet man dafür ein Programm, z.B. SPSS oder R. Für jede Frage wird dort ein Speicherfeld – eine sogenannte Variable – eingerichtet. Die Antwort auf die Frage 1.1 „Wie wichtig ist das Thema demographischer Wandel zurzeit für ihre Kommune?“ wird zum Beispiel in der Variablen 1 mit der Bezeichnung „v1“ gespeichert. Vor der Erfassung aller Daten müssen erst einmal sämtliche Variablen des Fragebogens eingerichtet werden. Für das vorliegende Beispiel nachstehend einen Auszug der Variablenliste in SPSS. Diese Ansicht der Struktur der Befragung wird in SPSS als Variablenansicht bezeichnet. Jede Zeile entspricht hier einer Variablen.

	Name	Typ	Breite	Dezimal...	Beschriftung	Werte
1	varlfdnr	Numerisch	5	0	laufende Nummer	Keine
2	vpanel	Numerisch	5	0	Panelkennung	Keine
3	v1	Numerisch	1	0	Relevanz für Kommune	{1, nicht wichtig}...
4	v2	Numerisch	1	0	Behandlung in Gesellschaft	{1, 1 unzureichend}...
5	v3	Numerisch	1	0	Behandlung in Kommune	{1, 1 unzureichend}...
6	v4	Numerisch	1	0	Auswirkungen Bundesebene	{1, gar nicht}...
7	v5	Numerisch	1	0	Auswirkungen Landesebene	{1, gar nicht}...
8	v6	Numerisch	1	0	Auswirkungen Kommunale Ebene	{1, gar nicht}...

Sind alle Variablen eingerichtet, kann die Erfassung der Antworten erfolgen. In SPSS muss dafür von der Variablenansicht in die Datenansicht gewechselt werden. In der nachstehenden Grafik stellt jetzt jede Zeile die Antworten eines Probanden da.



	varlfdnr	vpanel	v1	v2	v3	v4	v5	v6
1	1	1	4	2	3	2	2	3
2	2	2	3	3	3	2	3	3
3	3	3	4	2	4	2	2	3
4	4	4	3	2	2	2	2	2

In Abhängigkeit davon, welche Ausprägung zwischen 1 und 5 von dem Probanden angekreuzt wurde wird entsprechend der Codierung der Wert (1 bis 5) in der Variablen erfasst. Der Proband 1 hat beispielsweise bei der Frage 1.1, deren Antworten in der Variablen v1 gespeichert werden, den Wert 4 angekreuzt; der Proband 2 hat bei der gleichen Frage den Wert 3 angekreuzt.

Dem obigen Auszug aus dem Fragebogen ist zu entnehmen, dass die Antwort auf die Frage 1.2 „wird das Thema demographischer Wandel in der Gesellschaft ausreichend behandelt“ in der Variablen v2 gespeichert. Der Proband 1 hat hier den Wert 2 angekreuzt, so dass in der ersten Zeile (Proband 1) bei der Variablen v2 der Wert 2 eingegeben wurde.

In der Matrix für die Datenerfassung stellen folglich die Variablen die Spalten dar. Jede Zeile stellt die Antworten eines Probanden dar, die als Datensatz bezeichnet werden.

1.1.3 Auswertung empirischer Erhebungen

Auswertung

Nach der Datenerfassung erfolgt die statistische Auswertung der Daten. Eine erste, einfache Form ist die der **Häufigkeitsverteilung**, bei der die Häufigkeit des Auftretens eines Merkmals - z.B. der Einschätzung „nicht wichtig“ - durch einfaches Zählen bestimmt wird. Anhand der nachstehenden Tabelle ist zu erkennen, dass in fünf Kommunen die Relevanz des Themas „Demographischer Wandel“ als „nicht wichtig“ eingestuft wird.

Relevanz für Kommune

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	nicht wichtig	5	,8	,8	,8
	2	33	5,1	5,1	5,9
	3	145	22,4	22,4	28,3
	4	251	38,7	38,9	67,2
	sehr wichtig	212	32,7	32,8	100,0
	Gesamt	646	99,7	100,0	
Fehlend	System	2	,3		
Gesamt		648	100,0		

1.2 Struktur der Statistik

Deutlich wird an diesen Beispielen, dass erst bei der Auswertung der Daten die Statistik zum Tragen kommt. Die zentralen Problemstellungen der Statistik und die daraus abgeleiteten Strukturen sollen wiederum an zwei einfachen Beispielen erläutert werden. Da die Fragestellung dieser beiden Beispiele

sehr einfach ist, wird hier auf die Formulierung einer Hypothese; der Operationalisierung und der Ableitung der Fragen verzichtet. Das **erste Beispiel** befasst sich mit den Englisch- und Mathematiknoten von mehreren Studierenden. Die Befragung würde dann auch nur in den Fragen nach diesen beiden Noten bestehen. Das **zweite Beispiel** beinhaltet die sogenannte „Sonntagsfrage", also der Frage nach der Wahl der Partei eines Probanden, wenn am kommenden Sonntag Bundestagswahl wäre. In diesem Fall besteht die empirische Erhebung nur aus einer Frage, nämlich danach welche Partei von den Probanden gewählt würde.

Beschreibende Statistik: Urliste, Häufigkeitsverteilung und Parameter

Das erste Beispiel ist der Notenspiegel einer Klausur, der die Anzahl der Schüler mit einer bestimmten Note darstellt. Die nachstehende Abbildung zeigt die Einzelnote für jeden Probanden, die man als **Urliste** bezeichnet.

	ID_Student	Englisch	Mathematik
1	1	4,0	2,0
2	2	3,0	1,0
3	3	3,0	3,0
4	4	2,0	3,0
5	5	4,0	4,0
6	6	3,0	4,0
7	7	3,0	3,0
8	8	1,0	3,0
9	9	3,0	2,0
10	10	5,0	3,0

Aus der Urliste wird durch einfaches Auszählen welche Note wie häufig auftritt der **Häufigkeitsverteilung** erzeugt. Die Anteilswerte in Prozent ergeben sich dann durch Prozentrechnung, z.B. für die Note 1 durch

$$p = \frac{Anzahl\ für\ Note}{Gesamtanzahl} * 100 = \frac{8}{70} * 100 = 11{,}4286\%$$

Häufigkeitsverteilung

Note Mathematik

		Häufigkeit	Prozent	Kumulierte Prozente
Gültig	1,0	8	11,4	11,4
	2,0	13	18,6	30,0
	3,0	27	38,6	68,6
	4,0	18	25,7	94,3
	5,0	4	5,7	100,0
	Gesamt	70	100,0	

Parameter

Statistiken

Note Mathematik

N	Gültig	70
	Fehlend	0
Mittelwert		2,957
Std.-Abweichung		1,069

Neben dieser Häufigkeitsverteilung sind aber auch die bereits angesprochenen **Parameter** zur Beschreibung der Verteilung von Bedeutung. So ist zum Beispiel zur Einordnung der eigenen Leistung der **Mittelwert** von 2,957 von Bedeutung. Ist die eigene Note beispielsweise 2,0, so ist der Schüler besser als der Durchschnitt. Des Weiteren ist auch von Interesse, wie stark die Noten um den Mittelwert schwanken. Dieser Wert wird als **Standardabweichung** bezeichnet. Wenn alle Schüler in der Klausur die Note 3 bekommen hätten, wäre die Standardabweichung 0. In diesem Beispiel ist die Standardabweichung 1,069; d.h., dass die Noten im Mittel um 1,069 nach oben und nach unten vom Mittelwert abweichen.

Schließende Statistik: Stichprobe, Grundgesamtheit und Wahrscheinlichkeitsrechnung

Des Weiteren wird hier mit dem Politbarometer des ZDFs auf ein geläufiges Beispiel aus der Wahlforschung Bezug genommen.

Politbarometer: 10.09.2021

[1]

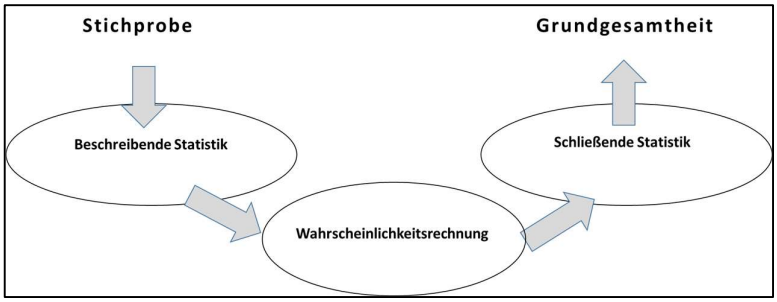
Auf der entsprechenden Seite findet man auch Hinweise zum Verfahren und zur Qualität der Aussagen:

> „Da es sich um eine Zufallsstichprobe handelt, kann für jedes Stichprobenergebnis ein Vertrauensbereich angegeben werden, innerhalb dessen der wirkliche Wert des Merkmals in der Gesamtheit mit einer bestimmten Wahrscheinlichkeit liegt. Unter Berücksichtigung des Stichprobendesigns und des Gewichtungsmodells ergeben sich bei einem Stichprobenumfang von n = 1.250 folgende Vertrauensbereiche: Der Fehlerbereich beträgt bei einem Anteilswert von 40 Prozent rund +/- drei Prozentpunkte und bei einem Anteilswert von 10 Prozent rund +/- zwei Prozentpunkte.“[2]

Die Anzahl der befragten Bürger, der Probanden, beläuft sich hier auf n=1.250. Diese Menge bezeichnet man als **Stichprobe**. Anhand der Befragungsergebnisse lassen sich die prozentualen Anteile für die jeweilige Partei ermitteln. Im Gegensatz zum ersten Beispiel ist hier nur der Parameter „Anteilswert“ von Bedeutung. Diese Anteilswerte gelten erst einmal nur für die Stichprobe. Mit Hilfe der **Wahrscheinlichkeitsrechnung** und der **schließenden Statistik** lassen sich diese Werte auf die wahlberechtigte Bevölkerung, die **Grundgesamtheit** übertragen. Dabei wird in der Grundgesamtheit natürlich nicht der exakt gleiche Wert wie in der Stichprobe auftreten. Der Wert wird in der Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit in einem Intervall um den Wert der Stichprobe liegen. Hier wird dieses Fehlerintervall bei einem Anteilswert von 10% mit 2% angegeben. D.h., dass der Prozentsatz in der Grundgesamt in dem Intervall von 8% bis 12% liegen wird.

Teilgebiete der Statistik

Anhand dieser Beispiele lassen sich auch die mathematischen Teilgebiete identifizieren.



Mit Hilfe der **beschreibenden Statistik** werden die Parameter der Stichprobe berechnet. Der Sinn solcher Stichprobenuntersuchungen ist sehr häufig die Übertragung der Erkenntnisse auf die Grundgesamtheit. Wie bereits erläutert, werden die Parameterwerte der Stichprobe nicht exakt denen der Grundgesamtheit entsprechen, sondern in der Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit in einem Intervall liegen. Dieses Gebiet wird als **schließende Statistik** bezeichnet. Die Verfahren der schließenden Statistik basieren dabei auf den Gesetzen der **Wahrscheinlichkeitsrechnung**, was auch aus den Überlegungen des vorherigen Abschnitts hervorgeht: die Ergebnisse der Stichprobe können eben nur mit einer bestimmten Wahrscheinlichkeit auf die Grundgesamtheit übertragen werden.

[1] Internet: https://www.zdf.de/politik/politbarometer/210910-video-100.html; 16.09.21
[2] Internet: https://www.forschungsgruppe.de/Rund_um_die_Meinungsforschung/Methodik_Politbarometer/; Stand 16.09.21

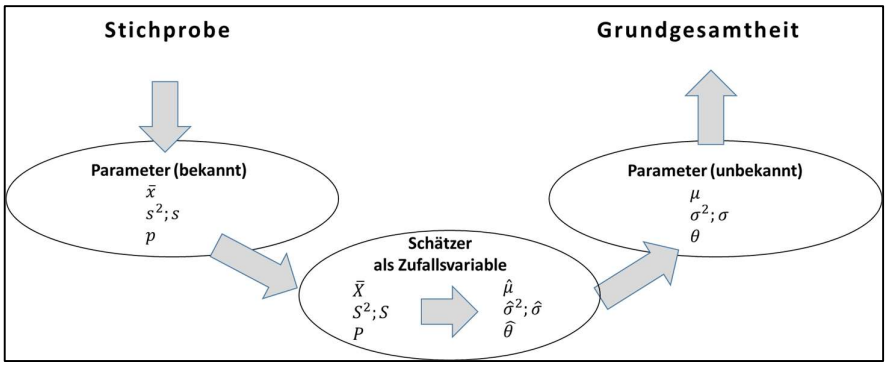
Die schließende Statistik und die Wahrscheinlichkeitstheorie werden auch unter dem Begriff **Stochastik** zusammengefasst.

> „Als **mathematische Statistik** bezeichnet man das Teilgebiet der Statistik, das die Methoden und Verfahren der Statistik mit mathematischen Mitteln analysiert beziehungsweise mit ihrer Hilfe erst begründet. Meist weitgehend synonym werden die Begriffe **induktive Statistik**, **beurteilende Statistik** und **Inferenzstatistik** (**schließende Statistik**) gebraucht, die den zur beschreibenden Statistik komplementären Teil der Statistik charakterisieren.
>
> Gemeinsam mit der Wahrscheinlichkeitstheorie bildet die mathematische Statistik das als **Stochastik** bezeichnete Teilgebiet der Mathematik. Die mathematische Grundlage der mathematischen Statistik ist die Wahrscheinlichkeitstheorie."[3]

1.3 Notation der Statistik

Unter einem Parameter versteht man in der Statistik – wie bereits erläutert - einen Wert, der die Verteilung der Daten in einer Kennzahl beschreibt. Es existiert in der Statistik eine Vielzahl von Parametern. Diese können auch sehr ähnliche Phänomene beschreiben; z.B. kann der Mittelwert einer Stichprobe aber auch der Mittelwert einer Grundgesamtheit von Bedeutung sein. Beide Mittelwerte sind zwar ähnlich, aber nicht identisch, da sie sich auf unterschiedliche Menge beziehen. Um diesen Unterschied deutlich zu machen, werden die Parameter mit unterschiedlichen Alphabeten gekennzeichnet. Neben den lateinischen werden in der Statistik auch die griechischen Buchstaben, manchmal auch zusätzlich noch die deutschen Buchstaben, verwendet. Ferner ist die Groß- und Kleinschreibung in dem jeweiligen Alphabet für eine weitere Differenzierung notwendig. Eine letzte Komponente sind Zusätze wie $\bar{}$ oder $\hat{}$, die als Akzente bezeichnet werden. Nachstehend erfolgt eine systematische Darstellung der Schreibweisen in der Statistik.

Entsprechend der Differenzierung in Stichprobe, Wahrscheinlichkeitsrechnung und Grundgesamtheit werden ähnliche Parameter unterschiedlich bezeichnet. Vorerst erschließt sich für den statischen Anfänger allerdings nur die Differenzierung zwischen der Stichprobe und der Grundgesamtheit. Schätzer und Zufallsvariablen werden zu einem späteren Zeitpunkt erläutert.



- In einer **Stichprobe** werden die Parameter mit einem **kleinen lateinischen** Buchstaben $\bar{x};\ s;\ p$ bezeichnet.
- In der **Grundgesamtheit** werden die gleichen Parameter mit einem **kleinen griechischen** Buchstaben $\mu;\ \sigma;\ \theta$ bezeichnet.

Durch diese Differenzierung wird sofort deutlich, ob es sich bei dem betrachteten Wert um den Mittelwert einer Stichprobe $\bar{x}$ oder den Mittelwert einer Grundgesamtheit μ handelt.

Für das Verständnis der statistischen Literatur ist diese Gesamtschau der Notation wichtig. Grundsätzlich ist anzumerken, dass leider keine verbindliche Notation in der Statistik existiert; in machen Quellen werden andere Notationen verwendet. Allerdings ist die hier verwendete Notation die am häufigsten anzutreffende.

[3] Wikipedia „Mathematische Statistik"; 14.09.2021

1.4 Aufbau der Skripte

Entsprechend der obigen Erläuterungen gliedern sich die Materialien zur Statistik daher auch in

- Beschreibende Statistik,
- Wahrscheinlichkeitsrechnung,
- Schließende Statistik
- Multivariate Verfahren (weiterführende Verfahren)

Ergänzend existiert für die Konzeption von Erhebungen ein Skript zur

- empirischen Sozialforschung,

Des Weiteren finden Sie auch noch kurze Einführungen zu den technischen Werkzeugen wie

- SPSS,
- R und
- Excel,

auf die teilweise bei den statistischen Beispielen im Skript Bezug genommen wird.

1.5 Literatur

Literaturempfehlungen sind immer problematisch, da sie nur passgenau aus der Perspektive des Lernenden sinnvoll sind. Daher folgen hier zu der nachstehenden Liste einige ergänzenden Erläuterungen.

Die **mathematische Literatur** ist für Anwender der Statistik weniger geeignet. Sie liefert aber i.d.R. als einzige einen vertiefenden Einblick in die Struktur und den Aufbau der mathematischen Statistik.

Anwendungsorientiert, leider aber auch mit weniger mathematischem Hintergrund sind die Publikationen aus den **Wirtschafts- und Sozialwissenschaften**. Für Interessenten, die mit R arbeiten, sei auf das Buch Fahrmeir u.a. hingewiesen, dass bei den Übungen auch immer den entsprechenden R-Code aufführt. Die Bücher von Dürr und Mayer gehen nicht so tief in die Materie, sind aber durch die kompakte und übersichtliche Darstellung gut für die tägliche Anwendung. Die beiden Bänder von Backhaus u.a. beziehen sich nur auf die multivariaten Methoden sind aber in diesem Bereich sehr ausführlich und beschreiben die Verfahren auch rechnerisch sehr detailliert. Die Publikation von Bortz wird inzwischen von anderen Autoren aktualisiert; sie galt und gilt in der Psychologie als Standardwerk und beschreibt die statistischen Berechnungen, teilweise auch mit Herleitungen, sehr ausführlich.

Mathematische Literatur

Kreyszig, Erwin	Statistische Methoden und ihre Anwendungen	Vandenhoeck & Ruprecht ISBN 3-525-40717-3
Georgii; Hans-Otto	Stochastik; Einführung in die Wahrscheinlichkeitstheorie und Statistik	De Gruyter ISBN 978-3-11-035969-5
Mosler, Karl Schmid, Friedrich	Wahrscheinlichkeitsrechnung und schließende Statistik	Springer ISBN 978-3-642-15009-8

Statistik für Wirtschaft- und Sozialwissenschaften

Mayer, Horst	Beschreibende Statistik	Carl Hanser Verlag ISBN 3-446-18068-0
Dürr, Walter Mayer, Horst	Wahrscheinlichkeitsrechnung und Schließende Statistik	Carl Hanser Verlag ISBN 3-446-17050-2
Bleymüller, Josef Gehlert, Günther Gülicher, Herbert	Statistik für Wirtschaftswissenschaftler	Vahlen Verlag ISBN 3-8006-2081-2
Fahrmeir, Ludwig Heumann, Chrisitan Künstler, Rita	Statistik Der Weg zur Datenanalyse	Springer Verlag ISBN 978-3-662-50371-3

Pigeot, Iris Tutz, Gerhard		
Backhaus, Klaus Erichson, Bernd Plinke, Wulff Weiber, Rolf	Multivariate Analysemethoden	Springer-Verlag ISBN 3-540-27870-2
Bortz, Jürgen	Statistik für Sozialwissenschaftler	Springer-Verlag ISBN 3-540-56200-1

SPSS

Brosius, Felix	SPSS	Mitp Verlag

Formelsammlung

Bleymüller, Josef Gehlert, Günther	Statistische Formeln, Tabellen und Programme	Vahlen Verlag ISBN 978 3 8006 3371 X