

Veith Tiemann

Inferenzmethoden und Multivariate Statistik

Grundlagen mit SPSS verstehen





utb 5121



Eine Arbeitsgemeinschaft der Verlage

Böhlau Verlag • Wien • Köln • Weimar
Verlag Barbara Budrich • Opladen • Toronto
facultas • Wien
Wilhelm Fink • Paderborn
Narr Francke Attempto Verlag • Tübingen
Haupt Verlag • Bern
Verlag Julius Klinkhardt • Bad Heilbrunn
Mohr Siebeck • Tübingen
Ernst Reinhardt Verlag • München
Ferdinand Schöningh • Paderborn
Eugen Ulmer Verlag • Stuttgart
UVK Verlag • München
Vandenhoeck & Ruprecht • Göttingen
Waxmann • Münster • New York
wbv Publikation • Bielefeld



Veith Tiemann

Inferenzmethoden und Multivariate Statistik

Grundlagen mit SPSS verstehen

UVK Verlag · München

Der Autor

Prof. Dr. Veith Tiemann lehrt an der EBC Hochschule Hamburg.

Online-Angebote oder elektronische Ausgaben sind erhältlich
unter www.utb-shop.de

Bibliografische Information der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über [<http://dnb.ddb.de>](http://dnb.ddb.de) abrufbar.

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

© UVK Verlag 2019 – ein Unternehmen der Narr Francke Attempto Verlag GmbH & Co. KG

Lektorat: Rainer Berger, München
Einbandgestaltung: Atelier Reichert, Stuttgart
Einbandmotiv: © Stas_V – iStock
Druck und Bindung: CPI – Clausen & Bosse, Leck

UVK Verlag
Nymphenburger Str. 48
80335 München
Telefon: 089/452174-66
www.uvk.de

Narr Francke Attempto Verlag GmbH & Co. KG
Dischingerweg 5
72070 Tübingen
Telefon: 07071/9797-0
www.narr.de

UTB-Nr. 5121
ISBN 978-3-8252-5121-5

Vorwort

Das vorliegende Buch richtet sich an Studierende, die bereits eine Grundeinführung in die Statistik hinter sich haben. Datenanalyse, Wahrscheinlichkeiten, Verteilungen und das Konzept „der statistische Test“ sind keine völligen Fremdwörter.

Das bedeutet, dass zum Beispiel die Normalverteilung, die Binomialverteilung aber auch Konzepte wie bedingte Wahrscheinlichkeiten schon einmal studiert wurden. Im Wesentlichen werden diese Themen aufgegriffen und weitergeführt bzw. zusammengeführt.

Das Buch hat zum Ziel, ein grundsätzliches Handwerkszeug zusammenzustellen, welches Ihnen erlaubt, sich in Datensituationen zurechtzufinden und diese angemessen und furchtlos anzugehen.

Sie werden zunächst Inferenzmethoden kennenlernen, die dann zu multivariaten Methoden erweitert werden. SPSS wird Sie als Software stets begleiten. Statistik macht ohne eine vernünftige Software nur bedingt Sinn.

Ich möchte Ihnen noch einige Sätze zum Aufbau, zur Gestaltung und zu einigen Besonderheiten dieses Buches mit auf den Weg geben:

- Inferenzmethoden und multivariate Methoden werden in einem Lehrbuch behandelt. Multivariate Verfahren kommen ohne Testverfahren als Bewertungsinstrumente kaum aus. Die Testverfahren wiederum sind ohne Daten und Fragestellungen nutzlos. Diese entstehen meist in einem multivariaten Kontext. Man könnte also von einem kongenialen Team sprechen.
- Für jedes Verfahren, für jede statistische Methode, die im vorliegenden Buch behandelt wird, kann umgehend die Umsetzung und Anwendung in SPSS erfolgen. Das ist ein wichtiger Gleichschritt. Ohne eine gute Software kann man Statistik beobachten, aber niemals anwenden.
- Das Buch beginnt mit der ganz konkreten und umfangreichen Darstellung der Konstruktion zweier statistischer Tests, für stetige Daten und für diskrete Daten.
Jeder statistische Test läuft prinzipiell gleich ab, argumentiert gleich und hat die gleichen Schwächen. Die Konstruktion eines Testes zu beobachten, ist sehr lehr- und hilfreich, um das Konzept wirklich zu verstehen.

- Beginnen wird das Buch mit einer kleine Einführung in SPSS. Diese dient der Orientierung. Das vorliegende Buch ist nicht dazu gedacht, eine umfassende Einführung in die Software zu geben.
- Natürlich werden nicht alle Inferenz- und multivariaten Methoden behandelt. Aber Sie bekommen einen Überblick über die wichtigsten Werkzeuge. Und Sie werden sehen, dass alle Themen inhaltlich abgedeckt werden.
- Die Beispieldatensätze, die im Buch zum Einsatz kommen sowie weitere Aufgaben, finden Sie zum Download im Internet.

Inhaltsverzeichnis

I	Statistiksoftware	1
1	Wofür braucht man das?	3
2	Klassen von Software	5
3	SPSS – eine Statistiksoftware	7
3.1	Vorstellung	7
3.2	Daten	8
3.3	Transformationen	11
3.4	Graphiken	14
3.5	Analysen	20
4	Aufgaben Teil I	31
II	Inferenzstatistik	33
5	Einführung	35
5.1	Statistischer Test und multivariate Statistik	35
5.2	Der statistische Test – Die Konstruktion eines Testes mit stetigen Daten	35
5.2.1	Hypothesen werden benötigt	37
5.2.2	Vereinfachung durch Differenzenbildung	39
5.2.3	Die Prüfgröße	40
5.2.4	Die t-Verteilung kommt ins Spiel	42
5.2.5	Ein Ablaufplan für den t-Test	44

5.2.6	Die Entscheidungsmatrix mit den beiden Fehlern	45
5.2.7	Die Überschreitungswahrscheinlichkeit	45
5.2.8	Zusammenfassung	46
5.3	Der statistische Test – Die Konstruktion eines Tests mit diskreten Daten	47
5.3.1	Die Prüfgröße	48
5.3.2	Die Entscheidungsregel	48
5.3.3	Die Überschreitungswahrscheinlichkeit	50
5.3.4	Zweiseitig vs. Einseitig	52
5.4	Zusammenfassung	52
6	Der Einstichprobenfall	53
6.1	p-Test, Test auf Anteilswerte	53
6.1.1	Das Konzept	53
6.1.2	SPSS – Der Binomialtest	54
6.2	t-Test	56
6.2.1	Das Konzept	56
6.2.2	SPSS – Der t-Test	58
6.2.3	Die Power eines Test	61
6.3	Vorzeichentest	62
6.3.1	Das Konzept	62
6.3.2	SPSS – Der Vorzeichentest	64
6.4	Der χ^2 -Anpassungstest	65
6.4.1	Das Konzept	65
6.4.2	SPSS – Der χ^2 -Anpassungstest	69
6.5	Der Kolmogorov-Smirnov-Test	71
6.5.1	Das Konzept	71
6.5.2	SPSS- und graphische Ansätze	72
6.5.3	Anmerkungen zu Zufallszahlen	79
6.5.4	SPSS – KS-Test	80
7	Der Zweistichprobenfall	83
7.1	Ein paar Hinweise	83
7.2	Der χ^2 -Unabhängigkeitstest	85
7.2.1	Das Konzept	85
7.2.2	SPSS – χ^2 -Unabhängigkeitstest	87

7.3	Lagetest verbunden	89
7.3.1	Das Konzept	89
7.3.2	SPSS – Lagetest verbunden	90
7.4	Lagetest unverbunden	92
7.4.1	Das Konzept	92
7.4.2	t-Test	93
7.4.3	Varianztest	94
7.4.4	Welch-Test	96
7.4.5	SPSS – das unverbundene Zweistichprobenproblem	97
8	Der c-Stichprobenfall	101
8.1	Einfaktorielle univariate ANOVA	101
8.1.1	Das Konzept	101
8.1.2	SPSS – einfaktorielle ANOVA	105
8.2	Zweifaktorielle univariate ANOVA	108
8.2.1	Das Konzept	108
8.2.2	SPSS – zweifaktorielle ANOVA	108
9	Aufgaben Teil II	113
III	Abhängigkeitsstrukturen	119
10	Korrelation	121
10.1	Kontingenz	121
10.1.1	Das Konzept	121
10.1.2	SPSS – Kontingenz und Profile	123
10.2	Pearson	125
10.3	Spearman	127
10.4	Scheinkorrelation	129
10.5	SPSS – Korrelation	130
11	Regression	135
11.1	Einfache lineare Regression	135
11.1.1	Das Konzept	135
11.1.2	SPSS – einfache lineare Regression	136

11.2	Partielle Korrelation	140
11.2.1	Das Konzept	140
11.2.2	SPSS – partielle Korrelation	142
11.3	Autokorrelation	143
11.3.1	Das Konzept	143
11.3.2	SPSS – Autokorrelation	148
11.4	Erweiterung – multiple Regression	151
11.4.1	Das Konzept	151
11.4.2	SPSS – Regression	155
12	Hauptkomponentenanalyse	163
12.1	Hintergrund und worum es geht	163
12.2	Konkrete Durchführung einer HKA	166
12.3	Fahrplan einer HKA	172
12.4	Faktorenanalyse	173
12.5	SPSS – HKA	175
13	Aufgaben Teil III	183
IV	Gruppenstrukturen	191
14	Worum geht es?	193
15	Klassifizieren	195
15.1	Diskriminanzanalyse	195
15.2	SPSS – Diskriminanzanalyse	201
16	Segmentieren	211
16.1	Worum es geht	211
16.2	Hierarchische Verfahren – Clusteranalyse	212
16.3	Partitionierende Verfahren – K-Means	218
16.4	SPSS – Distanzen und Segmentierung	221
17	Aufgaben Teil IV	227

V	Anhänge	229
18	Literatur	231
19	Tabellenverzeichnis	233
20	Abbildungsverzeichnis	235
21	Stichwortverzeichnis	241

Teil I

Statistiksoftware

Kapitel 1

Wofür braucht man das?

Natürlich kann man Statistik erlernen, ohne jemals mit einer Software in Berührung zu kommen. Man kann sich mit Datenanalyse und Modellierung ebenfalls wunderbar ohne eine Software beschäftigen. Man kann sich in der Mathematik austoben und kann viele Tafeln füllen. Und das ist auch alles nicht unwichtig.

Das Verständnis der Statistik, so könnte man sagen, ist allerdings ggf. ein anderes. Statistik ist eine *Sprache*, die uns hilft Probleme zu formulieren und Lösungen zu finden. Diese Probleme, und das ist die Stärke der Statistik, dürfen der echten Welt, einem echten Kontext entspringen. Dieser kann betriebswirtschaftlich sein, aus der angewandten Psychologie kommen, aus der Biologie, wenn dort munter sequenziert wird oder sich aus Sicherheitsfragen bei Handygesprächen rekrutieren. Alle diese Themen (und vermutlich auch alle nicht genannten) haben einen gemeinsamen Nenner: **Daten, viele Daten**. Wir kommen hier mit Schwamm und Tafel nicht weiter.

Nicht dass wir uns falsch verstehen. Das statistische Arbeiten mit Schwamm und Tafel ist enorm wichtig, um die Disziplin voranzutreiben – Schwamm und Tafel soll stellvertretend gemeint sein für die mehr theoretische und abstrakte Herangehensweise, um der Methoden Willen. Hier soll es aber um die Anwendung gehen, um der Daten Willen. Und die braucht Daten und eine Software, um mit den Daten umgehen zu können.

Die Datenvolumen, die fast in jedem Kontext eine Rolle spielen, haben zum Teil Größenordnungen erreicht, mit denen auch SPSS nichts anfangen kann. Aber auch wenn wir von diesen Big-Data-Szenarien einmal absehen, werden uns z. B. in der Wirtschaft nicht selten Datensätze jenseits der 10000 Fälle mit 20 Feldern begegnen. Das sind viele Datenpunkte.

Datenbanken: Grundbegriffe

Daten werden typischerweise in einer Datenbank gehalten. Eine Datenbank besteht aus der physikalischen Einheit, z. B. eine Magnetplatte und dem Datenbankmanagementsystem. Die letztgenannte Software sorgt für Ordnung und Konsistenz.

Man findet in einer Datenbank **Datensätze** – das entspricht so in etwa einer Zeile bei SPSS oder auch Excel. Die Spalten bei SPSS werden in der Datenbankwelt **Felder** genannt. Die Zeilen sind also objektorientiert, während die Spalten merkmalsorientiert sind.

Um Daten aus einer Datenbank zu bekommen, verwendet man üblicherweise eine Abfragesprache, wahrscheinlich SQL (= Structured Query Language).

Datenbanken kommen in allen möglichen Formen daher. Darum soll es hier nicht gehen. Ohne Datenbanken kommt heutzutage kaum ein Unternehmen aus. Vermutlich liegt jeder statistischen Analyse zunächst eine Datenbankabfrage zugrunde.

Kapitel 2

Klassen von Software

In der Regel ist es hilfreich, im Rahmen von Klassen über neue Konzepte nachzudenken. Auch bei Software ist das nützlich. Folgende Gliederungskriterien sind vermutlich sinnvoll:

■ **Proprietär vs. Open Source**

Es gibt einerseits Software, bei der der Quellcode geschützt ist, nicht verändert oder weitergegeben werden darf, und die Geld kostet. SPSS gehört in diese Klasse.

Andererseits gibt es ganz bewusst Software, die (prinzipiell) nichts kostet, deren Quellcode man weitergeben und sogar verändern kann.

Auch hier gibt es Spielregeln und manchmal auch Beschränkungen. Viele Open-Source-Anwendungen sind unter dem GNU-Projekt geregelt.

■ **GUI vs. Kommando**

Graphical User Interface bedeutet, dass man z. B. per Maus menügesteuert die Software bedienen kann. Das hat Vor- und Nachteile.

Offensichtlich ist es einfacher, auf diese Art mit einer Software umzugehen. Es ist nicht notwendig, eine spezielle Steuerungssprache zu erlernen.

Andererseits sind der Anwendung Grenzen gesetzt. Nur das, was der Programmierer in den Menüs hinterlegt hat, kann auch realisiert werden. Darüber hinaus kann es problematisch sein, dass *jeder* eine statistische Analyse durchführen kann, ein Ergebnis bekommt, aber nicht wirklich weiß, was man damit anfangen kann.

Kommandoorientierte Software gibt einem die Freiheit, alles machen zu können. Allerdings muss man das Coding erlernen.

■ **Spreadsheet vs. Modellierung vs. wieder Kommando**

Hierbei geht es darum, wie eigentlich die Daten innerhalb der Software gehalten und visualisiert werden und wie diese mit den Methoden zusammengebracht werden.

Excel ist eine klassische Spreadsheet-Software, d. h. die Daten sind in einer Matrix angeordnet und stets sichtbar. Die Vorteile sind offensichtlich, man

sieht die Daten immer. Es ist wichtig, die Daten zu sehen. Die Nachteile sind deutliche Einbußen bei der Performance, wenn die Daten sehr umfangreich sind.

Viele Statistikprogramme bieten eine Modellierungssuite an. Das heißt, wie in einem Flussdiagramm können Daten mit Methoden verknüpft werden.

In der kommandoorientierten Welt sieht man die Daten und die Methoden nur, wenn man sie aufruft. Sie sind ansonsten im Hintergrund.

Kapitel 3

SPSS – eine Statistiksoftware

3.1 Vorstellung

SPSS war eine der ersten statistischen Softwareangebote. Bereits 1968 wurde es unter dem Namen *Statistical Package for the Social Sciences (SPSS)* veröffentlicht. Die Sozialwissenschaften waren auch zunächst die Hauptanwender.¹

SPSS wurde 2009 von IBM aufgekauft und wird nun unter dem Namen IBM SPSS Statistics vertrieben. SPSS fällt in die Klasse der proprietären Software, es gibt ein GUI und die Daten sind prinzipiell im Spreadsheet-Format dargestellt. Natürlich gibt es für SPSS die entsprechenden Ausbaumöglichkeiten, um z. B. auch per Coding Ergebnisse zu erzielen. Mit dem SPSS-Modeler hat man dann auch die vorhin erwähnte flussdiagrammartige sogenannte Workbench.

Es gibt sehr (!) viele Anbieter von statistischen Programmpaketen.²

Genannt seien an dieser Stelle nur wenige, die aber sehr bedeutend sind:

- Proprietär: SAS, S-Plus, STATISTICA
- Open Source: R, PSPP

R sei besonders hervorgehoben, da es sich mittlerweile zur wahrscheinlich bedeutendsten statistischen Programmiersprache entwickelt hat.³ Diese Aussage gilt sowohl für Hochschulen als auch für Unternehmungen. R hat eine ganze Reihe von Vorzügen, die die Software so attraktiv macht. Die hohe Verarbeitungsgeschwindigkeit sowie die phantastischen graphischen und statistischen Möglichkeiten seien hier erwähnt. Dazu kommt die Freiheit, die eine Programmiersprache bietet. Zudem ist sie kostenfrei nutzbar und wird durch das große und professionelle Netzwerk stets erweitert.

¹PSPP ist die open source Variante zu SPSS. PSPP wird unter der GNU Lizenz vertrieben.

²Diese Webseite gibt einen sehr guten Überblick: https://en.wikipedia.org/wiki/List_of_statistical_packages

³R finden Sie hier: <https://cran.r-project.org/>

Im Rahmen dieser Vorstellung und des Buches wird **IBM SPSS Statistics 24** unter Windows verwendet. Wenn Sie SPSS starten, werden Sie diese leere Arbeitsfläche sehen:

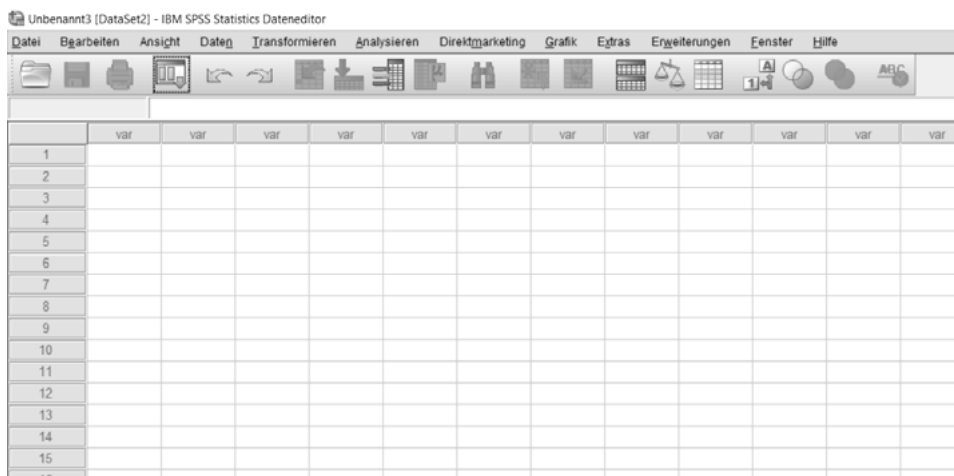


Abbildung 3.1: SPSS – Startscreen

Unschwer erkennbar ist das Windows-Look-and-Feel. Andererseits sehen Sie auch sofort, dass es durchaus andere Menüpunkte gibt, als man es beispielsweise von den Office-Anwendungen kennt. Die nächsten Kapitel stellen die Software vor. Die Reihenfolge orientiert sich dabei an dem typischen Ablauf einer Analyse. Der Punkt Analyse bezieht sich dabei aber ausschließlich auf rein deskriptive Vorgehensweisen.

3.2 Daten

Um mit SPSS arbeiten zu können, benötigen wir Daten. Die Oberfläche, das Spreadsheet, erlaubt natürlich eine direkte Dateneingabe. Es ist unwahrscheinlich, dass man seine Daten direkt in die Oberfläche eintippt. Vermutlich liegen die Daten in einer Datenbank oder stehen in einer Datei zur Verfügung.

SPSS erwartet die Daten stets als einen multivariaten Datensatz. Das bedeutet, die Spalten sind die Merkmale oder Variablen. Die Zeilen die Objekte. Unter dem Menüpunkt **Datei**-> **Öffnen**-> **Daten** können Sie genau das tun. Der Dialog bietet ihnen verschiedene Schnittstellen also Dateitypen an. Sie werden alle gängigen finden.

Es soll beispielhaft eine sogenannte CSV-Datei eingelesen werden – z. B. eine Befragung von Erstsemestern:

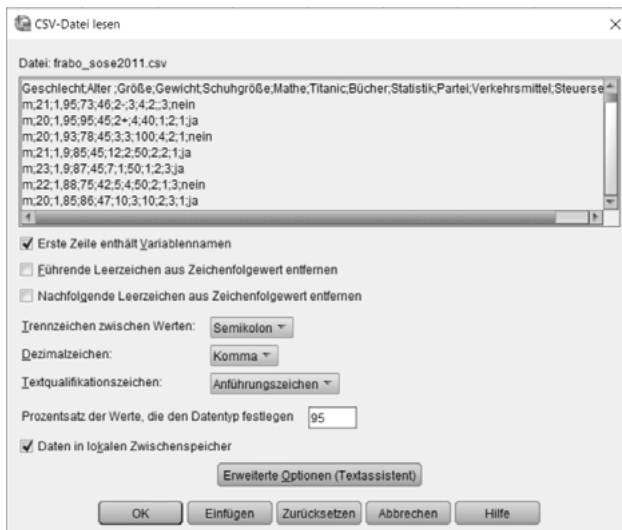


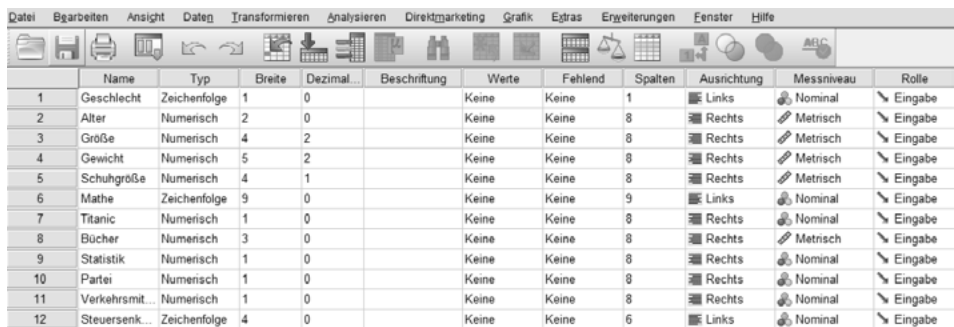
Abbildung 3.2: SPSS – Daten einlesen

Eine **CSV-Datei** ist eine Textdatei, die ein paar Metainformationen mitbekommen hat z. B. das Semikolon (;) als Trenner zwischen den verschiedenen Spalten. Wenn Sie den Dialog beenden, dann sieht das Datenfenster so aus:

	G s hle	Alter	Größe	Gewicht	Schuhgr öße	Mathe	Titanic	Bücher	Statistik	Partei	Verkehr smittel	Steuer senku ng
1	m	21	1,95	73,00	46,0 2-		3	4	2	.	3	nein
2	m	20	1,95	95,00	45,0 2+		4	40	1	2	1	ja
3	m	20	1,93	78,00	45,0 3		3	100	4	2	1	nein
4	m	21	1,90	85,00	45,0 12		2	50	2	2	1	ja
5	m	23	1,90	87,00	45,0 7		1	50	1	2	3	ja
6	m	22	1,88	75,00	42,0 5		4	50	2	1	3	nein
7	m	20	1,85	86,00	47,0 10		3	10	2	3	1	ja
8	m	19	1,84	82,00	45,0 2		2	10	2	3	1	nein
9	m	24	1,84	101,00	44,5 5		2	40	1	1	1	nein
10	w	21	1,83	72,00	38,5 10		1	30	2	3	3	ja
11	m	23	1,83	74,00	45,0 3		3	20	2	2	1	ja
12	m	20	1,82	73,00	45,0 9		2	30	1	.	3	nein
13	w	20	1,81	58,00	40,0 3		2	45	1	1	1	nein
14	w	19	1,81	75,00	41,0 2		3	50	2	2	1	ja
15	m	20	1,80	66,00	44,0 12		2	55	2	3	1	ja
16	m	20	1,80	83,00	43,0 1-		2	300	2	1	3	ja
17	w	18	1,80	70,00	41,0 2-		3	100	4	4	3	nein
18	m	19	1,80	80,00	45,0 2		2	30	2	3	1	nein
19	m	23	1,80	78,00	44,0 9		3	400	2	2	1	ja

Abbildung 3.3: SPSS – Datenansicht

Dies ist die sogenannte **Datenansicht**. Sie haben unten rechts die Möglichkeit, zur **Variablenansicht** zu wechseln:



	Name	Typ	Breite	Dezimal	Beschriftung	Werte	Fehlend	Spalten	Ausrichtung	Messniveau	Rolle
1	Geschlecht	Zeichenfolge	1	0		Keine	Keine	1	Links	Nominal	Eingabe
2	Alter	Numerisch	2	0		Keine	Keine	8	Rechts	Metrisch	Eingabe
3	Größe	Numerisch	4	2		Keine	Keine	8	Rechts	Metrisch	Eingabe
4	Gewicht	Numerisch	5	2		Keine	Keine	8	Rechts	Metrisch	Eingabe
5	Schuhgröße	Numerisch	4	1		Keine	Keine	8	Rechts	Metrisch	Eingabe
6	Mathe	Zeichenfolge	9	0		Keine	Keine	9	Links	Nominal	Eingabe
7	Titanic	Numerisch	1	0		Keine	Keine	8	Rechts	Nominal	Eingabe
8	Bücher	Numerisch	3	0		Keine	Keine	8	Rechts	Metrisch	Eingabe
9	Statistik	Numerisch	1	0		Keine	Keine	8	Rechts	Nominal	Eingabe
10	Partei	Numerisch	1	0		Keine	Keine	8	Rechts	Nominal	Eingabe
11	Verkehrsmit.	Numerisch	1	0		Keine	Keine	8	Rechts	Nominal	Eingabe
12	Steuersenk...	Zeichenfolge	4	0		Keine	Keine	6	Links	Nominal	Eingabe

Abbildung 3.4: SPSS – Variablenansicht

Die **Variablenansicht** erlaubt es, die notwendigen Einstellungen bei den Merkmalen zu prüfen und ggf. anzupassen z.B. das Skalenniveau oder den Typ. Es können bei kodierten Merkmalen auch entsprechende Label vergeben werden.

Ihnen ist natürlich aufgefallen, dass sich ein weiteres Fenster geöffnet hat, der **Statistics Viewer**.

Alles, was Sie im **Dateneditor**, also dem eigentlichen Fenster machen, wird im Viewer für Sie protokolliert, nichts geht verloren.

Vielleicht soll ein stetiges Merkmal auf Klassen transformiert werden – Gewicht soll als Beispiel dienen: **Transformieren-> Visuelle Klassierung**

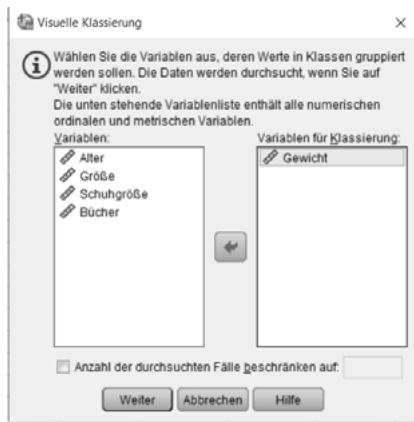


Abbildung 3.7: SPSS – Klassen bilden

Unter **Trennwerte erstellen** können Sie gerade dies tun und z. B. angeben, dass Sie 10 Klassen haben möchten usw:

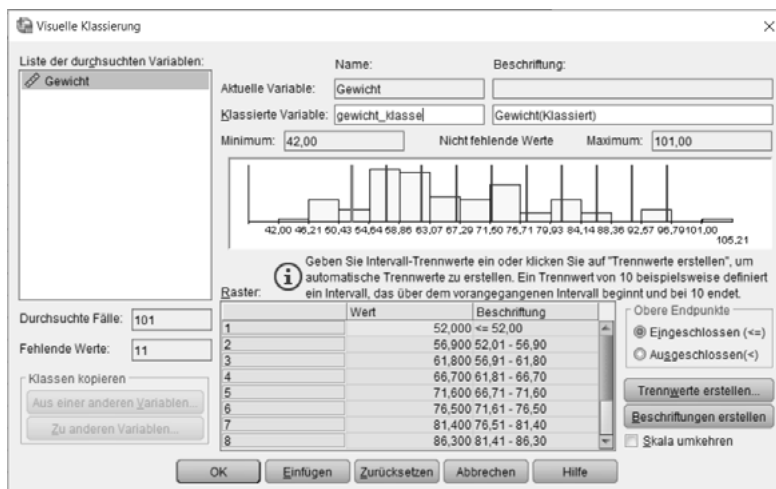


Abbildung 3.8: SPSS – Klassen bilden