

Studienbrief



Psychologie

Empirisches Forschungsprojekt

Datenanalyse mit SPSS (II)

Yvonne Ziert
Daniela Wenzel

2 EMF

Verfasser

Dipl. Sozialwiss. Yvonne Ziert, MPH

Studium der Sozialwissenschaften an der Leibniz Universität Hannover sowie an der John Moores University in Liverpool. Berufsbegleitender Master of Public Health an der Medizinischen Hochschule Hannover (MHH). Aktuell arbeitet sie als wissenschaftliche Mitarbeiterin am Institut für Biometrie der MHH. Im Rahmen ihrer beruflichen Tätigkeit hat sie viel Erfahrung in der Planung und Umsetzung von empirischen Forschungsprojekten im Gesundheitswesen erworben. Darüber hinaus hält sie seit mehreren Jahren Lehrveranstaltungen im Bereich der empirischen Forschungsmethoden (v. a. Biometrie und SPSS-Kurse).

Daniela Wenzel, MSc

Daniela Wenzel hat Mathematik mit den Schwerpunkten Biomathematik und Statistik studiert. Sie war wissenschaftliche Mitarbeiterin am Institut für Biometrie an der Medizinischen Hochschule Hannover. Tätigkeit bei Volkswagen Financial Services. Sie hat Erfahrung bei der Planung klinischer Studien, der statistischen Auswertung von Daten mit verschiedenen Statistikprogrammen (z. B. SPSS, R und SAS) sowie bei der Vermittlung von Lehrinhalten in den Bereichen Statistik und Biometrie.

Lektorat

Wissenschaftlicher Mitarbeiterinnen und Mitarbeiter der Hamburger Fern-Hochschule

Satz/Repro

Haussatz

Redaktionsschluss

Dezember 2019

1. Auflage 2019

© HFH · Hamburger Fern-Hochschule, Alter Teichweg 19, 22081 Hamburg

Das Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere das Recht der Vervielfältigung und der Verbreitung sowie der Übersetzung und des Nachdrucks, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Kein Teil des Werkes darf in irgendeiner Form ohne schriftliche Genehmigung der Hamburger Fern-Hochschule reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden.

Gedruckt auf 100 % chlorfrei gebleichtem Papier.

Inhaltsverzeichnis

Abkürzungsverzeichnis	4
Einleitung	5
1 Theoretische Grundlagen der deskriptiven Statistik	6
1.1 Ziel der deskriptiven Statistik	6
1.2 Messniveaus	7
1.3 Lage- und Streuungsmaße	8
1.4 Tabellarische und grafische Darstellungen der univariaten Statistik	9
2 Erstellen von Häufigkeitstabellen	11
2.1 Einfache Häufigkeitsauszählung	11
2.2 Bedingte Häufigkeitsauszählungen	13
2.3 Formate für Häufigkeitstabellen	14
Übungsaufgaben	15
3 Berechnung von statistischen Kennwerten	16
3.1 Einfache Berechnung von Lage- und Streuungsmaßen	16
3.2 Bedingte Berechnung von Lage- und Streuungsmaßen	19
Übungsaufgaben	19
4 Erstellen von Grafiken	20
4.1 Balkendiagramm	20
4.2 Kreisdiagramm	21
4.3 Histogramm	23
4.4 Boxplot	25
Übungsaufgaben	27
5 Editieren von Tabellen und Grafiken	28
5.1 Editieren von Tabellen	28
5.2 Editieren von Grafiken	29
Übungsaufgaben	31
6 Vergleich dichotomer Variablen	32
6.1 Erstellen von Kreuztabellen	32
6.2 Durchführung des Chi-Quadrat-Tests	34
Übungsaufgaben	36
7 Schätz- und Testverfahren zum Vergleich stetiger Variablen	37
7.1 Der Einstichproben t-Test	37
7.2 t-Test für zwei unabhängige Stichproben	40
7.3 t-Test für zwei abhängige Stichproben	44
Übungsaufgabe	47
8 Korrelationsmaße zur Analyse von Zusammenhängen	48
8.1 Korrelationskoeffizient	48
8.2 Korrelationskoeffizient nach Pearson	49
8.3 Korrelationskoeffizient nach Spearman	52
8.4 Bestimmtheitsmaß	54
Übungsaufgabe	55
Anhang 1: Kodeplan	56
Glossar	57
Lösungen zu den Übungsaufgaben	58
Literaturverzeichnis	66

Abkürzungsverzeichnis

BMI	Body Mass Index
ICD	International Statistical Classification of Diseases and Related Health Problems
IQR	Interquartilsabstand
kg	Kilogramm
m	Meter
ZUF-8	Fragebogen zur Patientenzufriedenheit

Einleitung

Jede statistische Auswertung sollte mit einer umfassenden Beschreibung der zu Grunde liegenden Daten beginnen. Die dazu benötigten methodischen Verfahren stellt die deskriptive Statistik zur Verfügung. Ihr Ziel ist es, durch die Berechnung spezifischer Kenngrößen und das Erstellen von Tabellen und grafischen Darstellungen den Auswertenden dabei zu unterstützen, die Daten zu verstehen und angemessene Verfahren für weiterführende bi- und multivariate Verfahren auszuwählen.

Für weitere Analysen, die über deskriptive Merkmale hinausgehen, z. B. die Untersuchung, ob zwei Gruppen sich hinsichtlich eines Merkmals zufallsbedingt oder systematisch voneinander unterscheiden, nutzt man Methoden der induktiven Statistik (auch Inferenzstatistik oder schließende Statistik genannt). In diesem Studienbrief wird vermittelt, die gängigen Verfahren für den Vergleich von zwei Gruppen hinsichtlich eines Merkmals mit dem Statistikprogramm SPSS anzuwenden und zu interpretieren. Darüber hinaus wird vermittelt, wie der Zusammenhang zwischen genau zwei Merkmalen praktisch in SPSS umgesetzt werden kann (bivariate Analysen).

Die theoretischen Grundlagen der deskriptiven Statistik sowie der induktiven Statistik wurden bereits in den Modulen Statistik I und II vermittelt. In diesem Studienbrief steht primär ihre praktische Umsetzung mit dem Statistikprogramm SPSS im Fokus. So bietet SPSS einfache, aber zugleich auch tiefgehende und vielfältige Möglichkeiten, die Methoden der deskriptiven und induktiven Statistik anzuwenden. Das Ziel dieses Studienbriefs ist es entsprechend, die diversen Möglichkeiten, die SPSS bietet, anwendungsorientiert zu vermitteln.

Die Bearbeitung der Kapitel dieses Studienbriefs setzt die in Studienbrief 1 vermittelten Kenntnisse voraus. Um auch in diesem Studienbrief den praktischen Anwendungsbezug beizubehalten, werden die unterschiedlichen Prozeduren anhand des Anwendungsbeispiels – Patientenfragebogen und Routinedaten des Klinikverbunds „Nordsterne“ – präsentiert, welches bereits in Studienbrief 1 dieses Moduls ausführlich beschrieben wurde.

Alle im Studienbrief verwendeten Abbildungen (*Screenshots*) wurden mit dem Programm *IBM SPSS Statistics 25* erzeugt.

Nach der Bearbeitung dieses Studienbriefs sind die Studierenden in der Lage mit Hilfe von SPSS:

- ⇒ einfache und bedingte Häufigkeitsauszählungen durchzuführen,
- ⇒ einfache und bedingte Lage- und Streuungsmaße zu berechnen und zu interpretieren,
- ⇒ adäquate Grafiken zur prägnanten Darstellung von deskriptiven Ergebnissen auszuwählen und zu erstellen,
- ⇒ ihre theoretisch erworbenen Kenntnisse über statistische Testverfahren praktisch anzuwenden;
- ⇒ den t-Test für verschiedene Szenarien durchzuführen und zu interpretieren sowie
- ⇒ verschiedene Korrelationsmaße zu berechnen und zu interpretieren.

Ziel des Studienbriefs

Wichtige Vorbemerkungen

Studienziele

1 Theoretische Grundlagen der deskriptiven Statistik

In der psychologischen Forschung möchte man häufig Fragen beantworten, die sich auf eine große Anzahl von Individuen beziehen. Die Gesamtheit an Individuen, auf die sich eine bestimmte Hypothese bezieht, wird als **Grundgesamtheit** bezeichnet (siehe *Abb. 1.1*). Aus logistischen, finanziellen oder sonstigen Gründen (z. B. mangelnde Teilnahmebereitschaft) ist es jedoch in der Regel unmöglich, alle Individuen einer Grundgesamtheit in seine Untersuchung einzubeziehen. Vor diesem Hintergrund ist es notwendig, per Zufallsauswahl eine **Stichprobe** aus der Grundgesamtheit zu ziehen, die diese möglichst gut repräsentiert.

Beispiel

Im Anwendungsbeispiel lassen sich **Grundgesamtheit** und **Stichprobe** folgendermaßen gegeneinander abgrenzen:

Die Grundgesamtheit stellen alle Patienten und Patientinnen dar, die im Befragungszeitraum an der vierwöchigen Reha-Maßnahme in einer der Kliniken der Reha-Zentren „Nordsterne“ teilgenommen haben.

In der Stichprobe hingegen sind alle Patienten und Patientinnen, die an der Befragung teilgenommen haben.

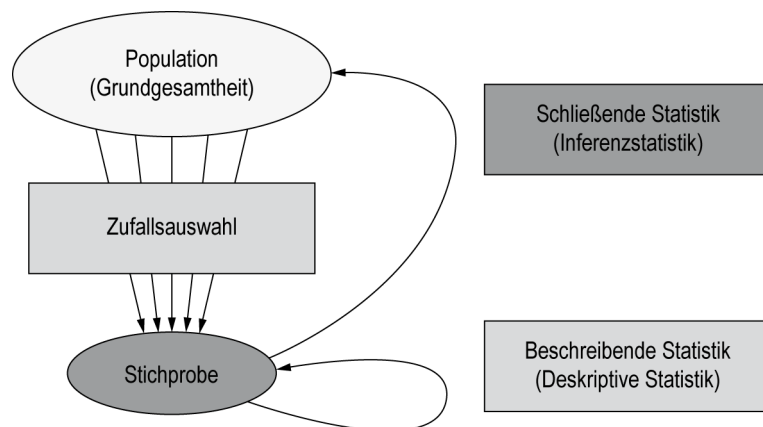


Abb. 1.1: Graphische Darstellung zur Verknüpfung von deskriptiver und induktiver Statistik (eigene Darstellung)

1.1 Ziel der deskriptiven Statistik

Das Ziel der **deskriptiven Statistik** ist es, die Stichprobe in Bezug auf die erfassten Merkmale (auch: Variablen genannt) zu beschreiben. Dazu zählen im Detail die folgenden Aspekte (vgl. Weiß, 2002, S. 17):

- Zusammenfassen und Ordnen der Daten in Tabellen
- Erstellen von Diagrammen und Grafiken
- Berechnen von Kenngrößen (z. B. Mittelwert, Standardabweichung).

Univariate vs. induktive Statistik

Die **univariate Statistik** beschränkt sich darauf, jeweils ein Merkmal zu beschreiben (vgl. Weiß 2002: 31). Diese Art der Aufbereitung der Daten unterstützt den Auswertenden dabei, die Daten zu verstehen und angemessene Verfahren für weiterführende bi- und multivariate Verfahren auszuwählen. Darüber hinaus ermöglichen sie es, Ergebnisse im Rahmen von Präsentationen oder Publikationen angemessen zu kommunizieren.

Mithilfe der Verfahren der deskriptiven Statistik können jedoch „nur“ Aussagen über die Verhältnisse in der Stichprobe getroffen werden. Um die Frage beantworten zu können, ob die Ergebnisse in der Stichprobe rein zufällig oder systematisch (signifikant) sind – d. h. dass man sie auch auf die Grundgesamtheit übertragen kann – sind Verfahren der **schließenden Statistik** notwendig.

1.2 Messniveaus

Die Merkmale, die im Rahmen einer empirischen Untersuchung erhoben werden, weisen in der Regel unterschiedliche Messniveaus (auch Skalen- oder Datenniveaus genannt) auf. Diese sind von zentraler Bedeutung, da das Messniveau die Wahl der korrekten statistischen Analyseverfahren bestimmt. Üblicherweise unterscheidet man zwischen (vgl. Zöfel, 2002):

- Nominalniveau
- Ordinalniveau
- Intervallniveau
- Verhältnisniveau.

Skalenniveau

Je höher das Skalenniveau ist, desto aussagekräftiger ist ein Merkmal und desto mehr statistische Verfahren sind anwendbar (vgl. Rasch et al. 2010: 8). Das Nominalniveau weist das geringste Skalenniveau auf, das Verhältnisniveau das höchste. Im Folgenden werden die unterschiedlichen Messniveaus vorgestellt, wobei mit dem Niedrigsten begonnen wird, um dann auf der Hierarchieleiter der Messniveaus bis nach oben durchzugehen.

Eine Variable weist dann Nominalniveau auf, wenn sich ihre Merkmalsausprägungen **keiner Ordnung** unterziehen lassen (z. B. *v10_geschlecht* – Geschlecht). Das heißt, die Ziffern, die man im Rahmen des Kodeplans den Ausprägungen zuordnet, können willkürlich gewählt werden. Variablen, die genau zwei Ausprägungen aufweisen, werden als **dichotom** oder auch **binär** bezeichnet.

Nominalniveau

Man spricht dann von Ordinalniveau, wenn sich die Ausprägungen einer Variablen einer **Ordnung** unterziehen lassen (z. B. *v14_1_rauch_vor* – Raucherstatus vor der Reha). Das heißt, die Ziffern, die im Kodeplan vergeben werden, spiegeln die tatsächliche Ordnungsrelation der Merkmalsausprägungen wider (vgl. Rasch et al. 2010: 13). In Bezug auf die Variable „Raucherstatus“ ist somit die Aussage möglich, dass eine höhere Ziffer für einen stärkeren Konsum von Zigaretten steht. Allerdings sind die numerisch gleichen Abstände zwischen den einzelnen Ausprägungen nicht als tatsächlich gleich zu betrachten.

Ordinalniveau

Merkmalsausprägungen einer Variablen, die nicht nur eine Rangordnung aufweisen, sondern zudem **gleiche Abstände zwischen den numerischen Ausprägungen** haben, weisen Intervallniveau auf (Rasch et al., 2010). Typische Beispiele für Merkmale mit Intervallskalenniveau sind – neben physikalischen Skalen wie der Celsius-Skala – psychologische Messskalen, die das Ergebnis eines langen Entwicklungsprozesses darstellen, in dem es u. a. darum geht, sicherzustellen, dass die Abstände zwischen Skalenstufen gleich sind oder als gleich wahrgenommen werden. Die Annahme gleicher Abstände ist nämlich die Voraussetzung dafür, die einzelnen Ausprägungen von Items zu einer Summenskala zu addieren, wie es bei der Skala ZUF-8 der Fall ist.

Intervallniveau

Eine Variable weist dann Verhältnisniveau auf, wenn sich nicht nur ihre Ausprägungen in eine Rangreihe bringen lassen und gleiche Abstände aufweisen, sondern sie

Verhältnisniveau

zusätzlich einen sog. **natürlichen Nullpunkt** aufweist (siehe Rasch et al. 2010: 12). Dies hat zur Folge, dass auch eine Verhältnissbildung zwischen den Ausprägungen möglich ist.

Ein Beispiel für eine Variable mit Verhältnissniveau ist die Variable *v15_sport*. Wenn eine Person dort eine 0 angibt, bedeutet es, dass sie sich gar nicht sportlich betätigt hat. Und eine Person, die dort 60 einträgt, hat sich doppelt so viele Minuten sportlich betätigt wie eine Person, die bei dem Item 30 angegeben hat.

Diskrete Merkmale

Darüber hinaus gibt es noch eine alternative Klassifizierungsmöglichkeit von Variablen – nämlich in diskrete und stetige Merkmale. Diese Klassifizierung liegt quer zu der bisherigen Einteilung. Ein Merkmal wird dann als diskret bezeichnet, wenn die **Ausprägungen**, die es annehmen kann, **abzählbar** sind. Trivialerweise sind alle qualitativen Merkmale diskret (vgl. Weiß 2002: 22). Am häufigsten wird der Begriff „diskret“ jedoch zur Beschreibung von quantitativen Merkmalen verwendet, deren Merkmalsausprägungen durch einen Zählvorgang ermittelt werden (z. B. Anzahl an Unterzuckerungen bei Diabetikern) (vgl. Weiß 2002: 22 ff.).

Stetige Merkmale

Stetige Merkmale sind hingegen dadurch gekennzeichnet, dass sie innerhalb eines bestimmten Intervalls theoretisch alle reellen Zahlen annehmen können. In der Regel werden diese **Werte durch einen Messvorgang ermittelt**. Ein Beispiel für eine stetige Variable ist z. B. die Variable „Gewicht“. Aufgrund der mangelnden Genauigkeit des Messverfahrens ist es jedoch häufig nur möglich, bei der Bestimmung eines stetigen Merkmals abzählbare Werte anzugeben, z. B. Gewicht in Kilogramm (vgl. Weiß 2002: 23). Demnach sind bei der praktischen Untersuchung fast alle Merkmale diskret. Für die Anwendung einiger statistischer Verfahren ist es jedoch erforderlich, dass ein Merkmal stetig ist, d. h. dass zwischen den gemessenen Werten theoretisch unendlich viele Werte liegen können (vgl. Weiß 2002: 24).

1.3 Lage- und Streuungsmaße

Um die empirisch erfassten Merkmale einer Stichprobe adäquat zu beschreiben, stehen in der deskriptiven Statistik eine Reihe von Lage- und Streuungsmaßen zur Verfügung. Dabei geben Lagemaße Auskunft darüber, wo sich die Stichprobenwerte konzentrieren (vgl. Weiß 2002: 43). Mit Hilfe von Streuungsmaßen kann die Variabilität der Messwerte beschrieben werden (vgl. Weiß 2002: 53). Die wichtigsten Lage- und Streuungsmaße mit ihren Anforderungen an das vorliegende Messniveau werden im Folgenden vorgestellt.

Lagemaße

Der **Modus** (auch: Modalwert genannt) ist der am häufigsten auftretende Wert einer Variablen. Dieses Maß lässt sich für jede Variable unabhängig vom Messniveau bestimmen.

Der **Median** liegt in der Mitte aller beobachteten Werte einer Variablen. Er ist als Lagemaß insbesondere bei Variablen mit Ordinalniveau und bei Variablen mit Intervall- und Verhältnissniveau, die nicht normalverteilt sind, sinnvoll, da er durch einzelne Extremwerte nicht beeinflusst wird.

Der arithmetische **Mittelwert** ist die Summe der beobachteten Werte geteilt durch die Anzahl der beobachteten Werte einer Variablen. Es ist ein passendes Maß zur Beschreibung der zentralen Lage der Daten für normalverteilte Variablen mit Intervall- und Verhältnissniveau.

Weitere wichtige Lagemaße sind die sog. **Quartile**. Das untere Quartil (Q_1) besagt, dass 25 % der Stichprobenwerte kleiner oder gleich Q_1 sind. Analog dazu besagt das

obere Quartil (Q_3), dass 25 % der Werte größer oder gleich Q_3 sind (Weiß, 2002). Der Median entspricht dem zweiten Quartil, d. h. 50 % der Stichprobenwerte sind kleiner oder gleich Q_2 . Häufig werden auch bestimmte Perzentile zur Beschreibung einer Stichprobe angegeben. Dabei wird die Verteilung in 100 umfanggleiche Teile aufgeteilt, so dass 1 %-Segmente entstehen. Das heißt, es kann z. B. ein 2 %-Perzentil oder auch ein 95 %-Perzentil bestimmt werden.

Die **Spannweite** ist das einfachste aller Streuungsmaße und wird berechnet, indem man die Differenz zwischen dem größten (Maximum) und kleinsten Wert (Minimum) einer Variablen bildet. Der Nachteil dieses Streuungsmaßes besteht darin, dass es nur die Extremwerte einer Variablen berücksichtigt. Der **Interquartilsabstand (IQR)** wird berechnet, indem die Differenz aus dem dritten und ersten Quartil gebildet wird ($Q_3 - Q_1$). Der Interquartilsabstand enthält 50 % aller Stichprobenwerte. Sowohl die Spannweite als auch der IQR lassen sich für Variablen mit Ordinal-, Intervall- und Verhältnissniveau berechnen.

Die **Standardabweichung** ist das gebräuchlichste Streuungsmaß und eignet sich für normalverteilte Variablen mit Intervall- und Verhältnissniveau. Die empirische Standardabweichung wird berechnet, indem man die Summe der quadratischen Abweichung aller Messwerte vom Mittelwert bildet, diese durch die Fallzahl teilt und hieraus die Wurzel zieht (wie Sie aus Statistik II wissen, muss man allerdings, wenn man einen erwartungstreuen Schätzer der Standardabweichung der Population erhalten möchte, die Fallzahl bei der Berechnung um 1 vermindern). Indem man auf das Ziehen der Wurzel verzichtet, erhält man die **Varianz**.

Streuungsmaße

1.4 Tabellarische und grafische Darstellungen der univariaten Statistik

In Abhängigkeit vom Messniveau, Art und Anzahl der Ausprägungen der erfassten Variablen bieten sich unterschiedliche Darstellungen zur prägnanten Beschreibung eines erfassten Merkmals an. Im Folgenden wird eine Auswahl der gängigsten Darstellungsmöglichkeiten der univariaten Statistik präsentiert.

Häufigkeitstabellen (siehe Kapitel 2) eignen sich insbesondere für die Darstellung von Variablen, bei denen die Ausprägungen überschaubar sind, wie es in der Regel bei Merkmalen mit Nominalniveau oder Ordinalniveau der Fall ist. Auch für quantitative Variablen mit einer überschaubaren Anzahl an diskreten Merkmalausprägungen, kann die Darstellung einer Häufigkeitstabelle zielführend sein. In der Regel beinhalten Häufigkeitstabellen Informationen darüber, wie viele Angaben pro Variable vorliegen (**absolute Häufigkeiten**) und welchem prozentualen Anteil dies in Bezug auf die gesamte Stichprobe entspricht (**relative Häufigkeiten**).

Häufigkeitstabellen

Eine alternative Darstellung von Häufigkeitstabellen für Variablen mit Nominal- oder Ordinalniveau ist die Erstellung eines Balkendiagramms (siehe Abschnitt 4.1). Dies gilt ebenfalls für quantitative Variablen, bei denen die Anzahl an diskreten Ausprägungen begrenzt ist. Die Länge der Balken eines Balkendiagramms korrespondiert entweder mit der Anzahl der absoluten Häufigkeiten pro Ausprägung der Variablen oder mit den relativen Häufigkeiten pro Ausprägung der Variablen.

Balkendiagramme

Kreisdiagramme sind v. a. für Variablen mit Nominalniveau geeignet, da dort nicht zum Ausdruck gebracht werden kann, welches die kleinste und welches die größte Ausprägung ist (siehe Abschnitt 4.2). In dem Kreisdiagramm entspricht jedes Segment entweder den absoluten oder den relativen Häufigkeiten der Ausprägung einer Variablen.

Kreisdiagramme

- Histogramme** Um sich grafisch einen Überblick über (quantitative) Daten, die in sehr vielen unterschiedlichen Ausprägungen vorliegen (z.B. die durchschnittliche Reaktionszeit jeder Person), und ihre Verteilung zu verschaffen, werden sog. Histogramme erstellt (siehe Abschnitt 4.3). Dazu ist es zunächst erforderlich, auf Basis der erhobenen Daten, **Klassen zu bilden**. Auf Basis dieser Klassen wird dann das Histogramm erstellt. Dies ähnelt auf den ersten Blick einem Balkendiagramm. Ein Unterschied besteht jedoch darin, dass zwischen den Balken keine Lücken vorhanden sind. Der Grund dafür ist, dass bei diesen Merkmalen theoretisch unendlich viele Werte zwischen den Klassengrenzen liegen können. Weiterhin ist es für die Darstellung von Histogrammen üblich, dass nicht die Höhe der Balken pro Klasse der absoluten Häufigkeit entspricht, sondern die Häufigkeit in einer Klasse der Fläche des Balkens entspricht. Sofern jedoch die Klassen alle gleich groß sind, sind die Balkenhöhen auch im Sinne der absoluten Häufigkeit interpretierbar und somit vergleichbar. Diese Art der Darstellung kommt der Dichtefunktion der Normalverteilung sehr nahe. Da es häufig bei stetigen Merkmalen für die Anwendung weiterführender Verfahren wichtig ist, diese auf **Normalverteilung** zu prüfen, ist die Erstellung eines Histogramms ein wichtiger Schritt bei der Deskription der Daten.
- Boxplots** Boxplots (auch: *Box-and-Whisker-Plots* genannt) eignen sich ausschließlich für Variablen mit Intervall- und Verhältnisniveau mit einer ausreichend großen Anzahl an unterschiedlichen Variablenausprägungen als grafische Darstellungsmöglichkeit (siehe Abschnitt 4.4). Für die Darstellung der Werte in der Stichprobe wird eine rechteckige Box gezeichnet, die oben und unten vom 1. und 3. Quartil begrenzt wird und somit 50% der Werte in der Stichprobe umfasst. Der Strich innerhalb der Box zeigt die Lage des Medians. Die Striche, die von der Box ausgehen (*Whisker*), zeigen die Lage des Minimums bzw. des Maximums der Variablen an.

2 Erstellen von Häufigkeitstabellen

In SPSS ist es grundsätzlich möglich, für jede Variable eine Häufigkeitstabelle zu erstellen. Besonders gut eignet sich die Darstellung jedoch für Variablen mit Nominal- oder Ordinalniveau. Als Datengrundlage können dabei alle Fälle in der Daten-datei verwendet werden (Abschnitt 2.1) oder nur eine definierte Subgruppe von Fäl-len (Abschnitt 2.2). Zudem besteht die Möglichkeit, unterschiedliche Formate für die Häufigkeitstabellen auszuwählen (Abschnitt 2.3).

Alle in diesem Kapitel durchgeführten Berechnungen basieren auf der Datendatei *Zentrum1_Spieldaten.sav*.

Beispieldatensatz

2.1 Einfache Häufigkeitsauszählung

Um für die Variable „Raucherstatus nach Reha“ eine Häufigkeitstabelle zu erstellen ist zunächst in der allgemeinen Menüleiste die Prozedur *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten* auszuwählen. Daraufhin öffnet sich die Dialogbox *Häufigkeiten* (siehe **Abb. 2.1**).

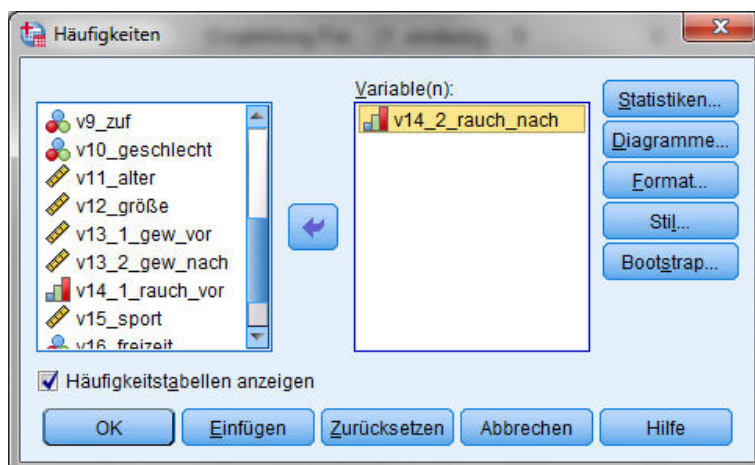


Abb. 2.1: Dialogbox – Häufigkeiten

Aus der Quellvariablenliste ist die Variable *v14_2_rauch_nach* mit der Transport-schaltfläche in das Feld *Variable(n)* zu bringen. Zudem ist darauf zu achten, dass die Checkbox *Häufigkeitstabellen anzeigen* aktiviert ist. Abschließend sind die ge-troffenen Einstellungen mit dem Schalter *OK* zu bestätigen, so dass im Viewer die entsprechenden Tabellen erscheinen. Der Tabelle „Statistiken“ (siehe **Abb. 2.2**) ist zu entnehmen, wie viele Patienten gültige Angaben zu der Variablen *v14_2_rauch_nach* gemacht haben. Demnach haben 27 Patienten gültige Angaben bei der Variablen gemacht und bei 3 Fällen gibt es fehlende Werte.

Statistiken

Raucherstatus nach Reha		
N	Gültig	27
	Fehlend	3

Abb. 2.2: Statistiken für die Variable „Raucherstatus nach Reha“ (*v14_2_rauch_nach*)

Bezug zum Kodeplan

Die eigentliche Häufigkeitstabelle ist die Tabelle mit der Überschrift „v14_2_rauch_nach“ (siehe **Abb. 2.3**), die folgendermaßen aufgebaut ist: Die ersten Zeilen der Tabelle entsprechen jeweils den gültigen Werten der betreffenden Variablen. Laut Kodeplan zum Patientenfragebogen (siehe Anhang 1) gibt es die drei Merkmalsausprägungen mit ihren entsprechenden Kodierungen „Nichtraucher“ (0), „Gelegenheitsraucher“ (1) und „Raucher“ (2). In der Häufigkeitstabelle werden aufgrund der Voreinstellung in SPSS ausschließlich die Wertelabels ohne Kodierung angegeben. Da zu allen Ausprägungen auch jeweils Angaben gemacht worden sind, tauchen diese alle in der Häufigkeitstabelle auf. Darauf folgt eine Zeile „Gesamt“, die sich auf alle gültigen Angaben der Variablen bezieht. Die darauffolgenden Zeilen listen die Werte auf, die als „Fehlend“ definiert sind. Hier unterscheidet SPSS benutzerdefinierte fehlende Werte – in diesem Fall wurde die 9 als fehlend definiert und mit dem Label „Angabe fehlend“ versehen. Sofern zu einer Person, gar keine Angabe vorliegt, definiert SPSS dies als „systemdefiniert fehlend“ (in der Tabelle mit „System“ bezeichnet). Darauf folgt wieder eine Zeile „Gesamt“, die sich auf die Summe aller fehlenden Werte bezieht. Die letzte Zeile ist noch einmal mit „Gesamt“ betitelt und bezieht sich auf alle Fälle der Datendatei (gültige plus fehlende Angaben).

Raucherstatus nach Reha					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Nichtraucher	21	70,0	77,8	77,8
	Gelegenheitsraucher	3	10,0	11,1	88,9
	Raucher	3	10,0	11,1	100,0
	Gesamt	27	90,0	100,0	
Fehlend	Angabe fehlend	2	6,7		
	System	1	3,3		
	Gesamt	3	10,0		
Gesamt		30	100,0		

Abb. 2.3: Häufigkeitstabelle für die Variable „Raucherstatus nach Reha“ (v14_2_rauch_nach)

Beschreibung Häufigkeitstabelle

Für die unterschiedlichen Zeilen der Tabelle werden zunächst in der Spalte „Häufigkeit“ die **absoluten Häufigkeiten** angegeben. So haben 21 Personen angegeben, Nichtraucher zu sein, 3 haben angegeben, Gelegenheitsraucher zu sein und 3 gaben an, Raucher zu sein. Das heißt, in der Summe haben 27 Personen gültige Angaben gemacht (siehe Zeile „Gesamt“). Insgesamt haben 3 Personen keine Angaben gemacht, wobei für 2 Personen die 9 kodiert wurde, der das Label „Angabe fehlend“ zugeordnet wurde. Bei einer Person, gibt es gar keine Angabe zu der Variablen, so dass sie von SPSS als „systemdefiniert fehlend“ gewertet wird. Insgesamt sind 30 Patienten und Patientinnen in der Datendatei enthalten (siehe letzte Zeile „Gesamt“).

Die Spalte „Prozent“ gibt die **relativen Häufigkeiten** zu den einzelnen Kategorien an. So haben 70,0 % der Patienten angegeben, Nichtraucher zu sein. insgesamt haben 90,0% der befragten Patienten gültige Angaben gemacht. Für 6,7% wurde die 9 (= Angabe fehlend) kodiert und für 3,3 % der Fälle gibt es gar keine Angabe. Die prozentualen Häufigkeiten aller Kategorien (gültig und fehlend) addieren sich zu 100 % auf.

Sofern es keine fehlenden Werte in den Daten gibt, entsprechen die Werte in der Spalte „Gültige Prozente“ den Angaben in der Spalte „Prozent“. Wenn es fehlende Werte gibt, werden die gültigen Prozente ausschließlich auf Grundlage der Personen mit gültigen Angaben berechnet. Das heißt, für die Berechnung des prozentualen

Anteils an Nichtrauchern auf Grundlage der gültigen Werte ist das folgende Verhältnis zu bilden $21/27 = 77,8\%$.

Der Spalte „Kumulierte Prozente“ ist zu entnehmen, wie viel Prozent der Personen mit gültigen Werten, Angaben zu einer betroffenen Kategorie und alle darunter liegenden Kategorien gemacht haben. So haben 88,9% angegeben, Gelegenheitsraucher oder Nichtraucher zu sein. Das heißt, um die kumulierten Prozente für eine bestimmte Kategorie zu erhalten, werden die gültigen Prozente dieser Kategorie und aller kleineren Kategorien aufaddiert.

Sofern für mehr als eine Variable eine Häufigkeitstabelle erstellt werden soll, kann dies dadurch umgesetzt werden, dass mit Hilfe der Transportschaltfläche einfach alle gewünschten Variablen aus der Quellvariablenliste in das Fenster *Variable(n)* übertragen werden. Anschließend ist die getroffene Auswahl mit dem Schalter *OK* zu bestätigen, so dass die Häufigkeitstabellen im Viewer erscheinen.

Häufigkeitstabellen für mehrere Variablen

2.2 Bedingte Häufigkeitsauszählungen

Wenn eine Häufigkeitstabelle nicht für alle Fälle in einer Datendatei erstellt werden soll, sondern nur für eine Subgruppe, dann ist diese zunächst über die **temporäre Fallauswahl** auszuwählen. Angenommen, man möchte die Häufigkeitstabelle nur für männliche Patienten erstellen: Für die temporäre Auswahl der männlichen Patienten ist im allgemeinen Menü die Auswahl *Daten* → *Fälle auswählen* vorzunehmen, so dass sich die Dialogbox *Fälle auswählen* öffnet (siehe **Abb. 2.4**).

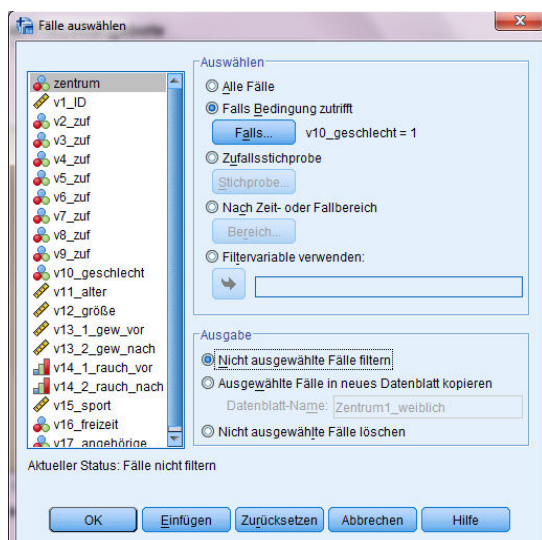


Abb. 2.4: Dialogbox – Fälle auswählen

In der Dialogbox *Fälle auswählen* ist die Auswahl *Alle Fälle* voreingestellt. Um die Fallauswahl auf die männlichen Patienten zu beschränken, ist die Einstellung *Falls Bedingung zutrifft* zu aktivieren und die Schaltfläche *Falls* zu betätigen. Daraufhin öffnet sich die Dialogbox *Fälle auswählen: Falls*. Die Variable *v10_geschlecht* ist mit Hilfe der Transportschaltfläche aus der Quellvariablenliste in den Konditional-Editor zu bringen. Anschließend ist auf der Rechentastatur das „=“ und die „1“ zu aktivieren, so dass der folgende Ausdruck im Konditional-Editor steht: *v10_geschlecht = 1*. Abschließend ist in der Dialogbox *Fälle auswählen: Falls* auf *Weiter* zu klicken, so dass sich die Dialogbox *Fälle auswählen* öffnet, in der die Einstellungen mit *OK* zu bestätigen sind.

Anschließend ist über die Menüfunktion wie in Abschnitt 2.1 beschrieben, eine Häufigkeitstabelle für die Variable „Raucherstatus nach Reha“ zu erstellen, so dass im Viewer die entsprechenden Tabellen (siehe **Abb. 2.5** und **Abb. 2.6**) erscheinen.

Statistiken		
Raucherstatus nach Reha		
N	Gültig	11
	Fehlend	0

Abb. 2.5: Statistiken für die Variable „Raucherstatus nach Reha“ (v14_2_rauch_nach) (Männer)

Es fällt auf, dass alle 11 Männer gültige Angaben zu ihrem Raucherstatus nach der Reha gemacht haben (siehe **Abb. 2.6**). Dies hat zur Folge, dass sich die Werte in den Spalten „Prozent“ und „Gültige Prozente“ entsprechen (siehe **Abb. 2.7**).

Raucherstatus nach Reha					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Nichtraucher	9	81,8	81,8	81,8
	Gelegenheitsraucher	1	9,1	9,1	90,9
	Raucher	1	9,1	9,1	100,0
	Gesamt	11	100,0	100,0	

Abb. 2.6: Häufigkeitstabelle für die Variable „Raucherstatus nach Reha“ (v14_2_rauch_nach) (Männer)

2.3 Formate für Häufigkeitstabellen

SPSS bietet die Möglichkeit, bereits bei der Erstellung von Häufigkeitstabellen ein bestimmtes Format zu berücksichtigen. Um die unterschiedlichen Möglichkeiten aufzuzeigen, ist zunächst dieselbe Menüauswahl wie bei der Erstellung einer Häufigkeitstabelle zu treffen: *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten*.

Indem auf den Schalter *Format* geklickt wird, öffnet sich die Dialogbox *Häufigkeiten: Format* (siehe **Abb. 2.7**). In dem Auswahlkasten *Sortieren nach* wird die Reihenfolge festgelegt, in der die Datenwerte der Häufigkeitstabelle im Viewer angezeigt werden. Hier kann zwischen den folgenden Optionen gewählt werden:

- *Aufsteigende Werte:* Hier werden die Merkmalsausprägungen nach steigender Ordnung sortiert. Diese Option ist voreingestellt.
- *Absteigende Werte:* Hier werden die Merkmalsausprägungen nach absteigender Ordnung sortiert.
- *Aufsteigende Häufigkeiten:* Hier werden die Merkmalsausprägungen nach steigender Ordnung der Häufigkeiten sortiert.
- *Absteigende Häufigkeiten:* Hier werden die Merkmalsausprägungen nach absteigender Ordnung der Häufigkeiten sortiert.

Häufigkeitstabellen für mehrere Variablen

Wenn für mehrere Variablen gleichzeitig Häufigkeitstabellen erstellt werden, dann kann der Benutzer über den Auswahlkasten *Mehrere Variablen* auswählen, wie die Tabelle „Statistiken“ im Viewer angezeigt werden soll (siehe **Abb. 2.7**). Voreingestellt ist die Option *Variablen vergleichen*. Dies hat zur Folge, dass alle Variablen in der Tabelle „Statistiken“ nebeneinander angezeigt werden, so dass die Anzahl an

gültigen bzw. fehlenden Werten direkt verglichen werden kann. Durch die Aktivierung der Option *Ausgabe nach Variablen ordnen*, wird für jede Variable eine separate Tabelle „Statistiken“ erstellt.

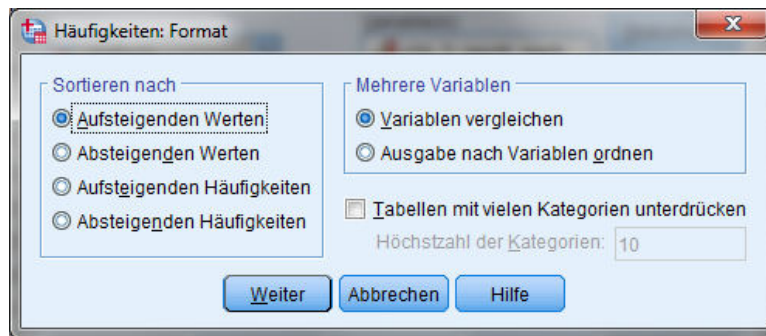


Abb. 2.7: Dialogbox – Häufigkeiten: Format

Darüber hinaus besteht die Möglichkeit, durch die Aktivierung der Checkbox *Tabellen mit vielen Kategorien unterdrücken*, das Ausgeben langer Häufigkeitstabellen zu unterdrücken. Als Höchstzahl an Kategorien (= Merkmalsausprägungen) ist 10 voreingestellt. Dies kann jedoch beliebig angepasst werden.

Übungsaufgaben

- 2.1) Erstellen Sie unter Rückgriff auf die Datei *Zentrum1_Spieldaten.sav* für die Variable *v3_zuf* eine Häufigkeitstabelle und beschreiben Sie, welche Informationen Sie dieser Tabelle entnehmen können.
- 2.2) Erstellen Sie unter Rückgriff auf die Datei *Zentrum1_Spieldaten.sav* eine Häufigkeitstabelle nur für Frauen für die Variable *v17_angehörige* und beschreiben Sie, welche Informationen Sie dieser Tabelle entnehmen können.

3 Berechnung von statistischen Kennwerten

In diesem Kapitel wird demonstriert, wie in SPSS Lage- und Streuungsmaße entweder auf Basis der gesamten Fälle in einer Datendatei (Abschnitt 3.1) oder für eine vorab definierte Subgruppe (Abschnitt 3.2) berechnet werden.

Beispieldatensatz

Alle Berechnungen in diesem Kapitel basieren auf der Datendatei *Zentrum1_Spieldaten.sav*.

3.1 Einfache Berechnung von Lage- und Streuungsmaßen

Für die Berechnung von Lage- und Streuungsmaßen stehen in SPSS diverse Möglichkeiten zur Verfügung, von denen zwei Vorgehensweisen im Folgenden dargestellt werden.

Um die erste Möglichkeit zur Berechnung von Lage- und Streuungsmaßen in SPSS zu zeigen, sollen diverse Lage- und Streuungsmaße für die Variable „Körpergewicht nach Reha“ berechnet werden. Dazu ist zunächst die folgende Menüauswahl zu treffen: *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten*. Nachdem sich die Dialogbox *Häufigkeiten* geöffnet hat, ist mit Hilfe der Transportschaltfläche die Variable *v13_2_gew_nach* aus der Quellvariablenliste in das Feld *Variable(n)* zu bringen. Damit für die Variable keine Häufigkeitstabelle angezeigt wird, ist die Checkbox *Häufigkeitstabellen anzeigen* zu deaktivieren. Jetzt ist durch Klicken auf den Schalter *Statistiken* (siehe *Abb. 2.1*) die Dialogbox *Häufigkeiten: Statistiken* zu öffnen (siehe *Abb. 3.1*).

Deaktivieren der Checkbox Häufigkeitstabellen anzeigen

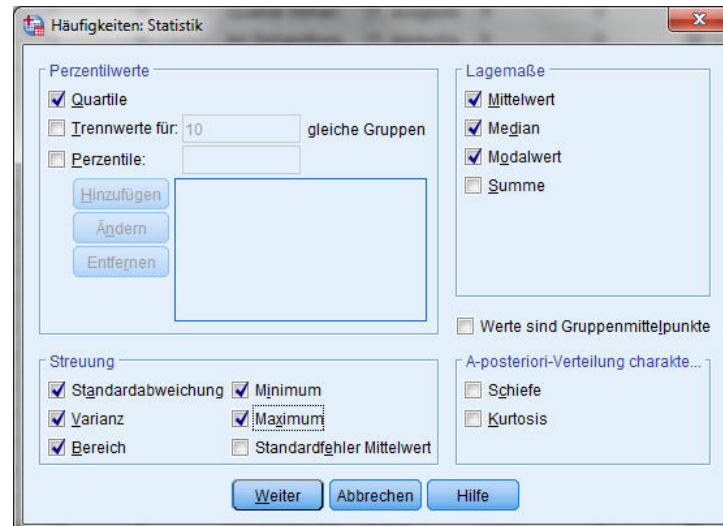


Abb. 3.1: Dialogbox – Häufigkeiten Statistiken

In den jeweiligen Auswahlkästen (Perzentilwerte, Lagemaße und Streuung) der Dialogbox *Häufigkeiten: Statistiken* können die gewünschten Lage- und Streuungsmaße durch Aktivierung der entsprechenden Checkbox ausgewählt werden. Da es sich bei der Variablen „Körpergewicht nach Reha“ um ein quantitatives Merkmal mit Verhältnissniveau handelt, können hier alle bekannten Lage- und Streuungsmaße berechnet werden (siehe Abschnitt 1.3). Nachdem die entsprechenden Maße ausgewählt wurden, sind die Einstellungen mit der Taste *Weiter* zu bestätigen, so dass man wieder in die Dialogbox *Häufigkeiten: Statistiken* zurückkommt. Um die Prozedur zur Ausführung zu bringen, ist hier der Schalter *OK* zu betätigen. Daraufhin erscheint im Viewer eine Tabelle mit den entsprechenden Kennzahlen (siehe **Abb. 3.2**). Diese Tabelle ist so aufgebaut, dass in der ersten Zeile angezeigt wird, dass auf Basis von 30 Fällen die unterschiedlichen Kennwerte berechnet wurden und dass es keine fehlenden Werte gibt. Darauf folgen die Werte zu den ausgewählten Lage- und Streuungsmaßen. So beträgt der Mittelwert für das Gewicht nach der Reha 81,53 kg mit einer Standardabweichung von 11,322 kg. Der Median liegt bei 80,50 kg (entspricht dem 50. Perzentil) und der Modus bei 79 kg. Die leichteste Person in der Stichprobe wiegt 58 kg (Minimum) und die schwerste Person wiegt 107 kg. Dies führt zu einer Spannweite von 49 kg. Der Wert für das erste Quartil liegt bei 75 kg (25. Perzentil) und der Wert für das dritte Quartil liegt bei 89,75 kg (75. Perzentil).

Beschreibung der Lage- und Streuungsmaße

Statistiken		
Körpergewicht nach Reha		
N	Gültig	30
	Fehlend	0
Mittelwert		81,53
Median		80,50
Modus		79
Std.-Abweichung		11,322
Varianz		128,189
Spannweite		49
Minimum		58
Maximum		107
Perzentile	25	75,00
	50	80,50
	75	89,75

Abb. 3.2: Lage- und Streuungsmaße für die Variable „Körpergewicht nach Reha“ (v13_2_gew_nach)

Die zweite Möglichkeit zur Berechnung von Lage- und Streuungsmaßen in SPSS wird anhand der Variablen „Sportliche Betätigung“ präsentiert. Dazu ist zunächst die folgende Menüauswahl *Analysieren → Deskriptive Statistiken → Deskriptive Statistik* zu treffen, so dass sich die Dialogbox *Deskriptive Statistik* öffnet (siehe **Abb. 3.3**). Anschließend ist die Variable v15_sport aus der Quellvariablenliste mit Hilfe der Transportschaltfläche in das Fenster *Variable(n)* zu bringen. Für die Auswahl der gewünschten Lage- und Streuungsmaße ist auf den Schalter *Optionen* zu klicken, so dass sich die Dialogbox *Deskriptive Statistik: Optionen* öffnet. Hier können wieder mit Hilfe der Checkboxes die gewünschten Lage- und Streuungsmaße ausgewählt werden. Die Auswahl ist jedoch weitaus begrenzter. So können als Lage- und Streuungsmaße nur Mittelwert, Standardabweichung, Varianz, Minimum, Maximum und Spannweite ausgewählt werden. Da es sich bei der Variablen „Sportliche Betätigung“ um eine quantitative Variable mit Verhältnissniveau handelt, dürfen alle angebotenen Streuungsmaße berechnet werden. Nachdem die Auswahl getroffen

wurde, ist diese mit dem Schalter *Weiter* zu bestätigen, so dass man in die Dialogbox *Deskriptive Statistik* zurückgelangt. Um die Prozedur zur Ausführung zu bringen, ist der Schalter *OK* zu aktivieren.

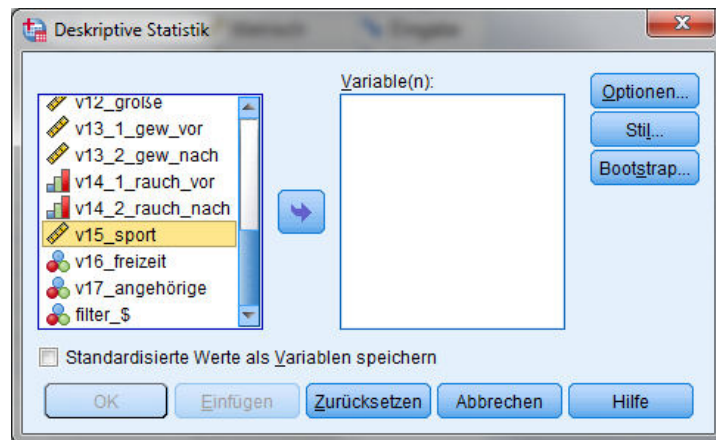


Abb. 3.3: Dialogbox – Deskriptive Statistik

In der Ausgabe erscheint daraufhin die Tabelle mit den Kennwerten (siehe **Abb. 3.4**). Im Gegensatz zu der ersten Tabelle mit den Kennwerten (siehe **Abb. 3.2**) weist diese Tabelle nun Querformat auf. Zunächst wird in der Tabelle wieder dargestellt, auf welcher Basis die Kennwerte berechnet wurden ($N = 30$). Darauf folgen dann die Werte für die ausgewählten Lage- und Streuungsmaße. So liegt der Mittelwert für die sportliche Betätigung bei 45,47 Minuten mit einer Standardabweichung von 9,737 Minuten. Die maximale Zeit, die sich ein Patient im Durchschnitt am Tag bewegt hat, liegt bei 62 Minuten. Die minimale Zeit, die für sportliche Betätigung im Durchschnitt am Tag aufgebracht wurde, liegt bei 28 Minuten. Daraus ergibt sich eine Spannweite von 34 Minuten.

Deskriptive Statistik					
	N	Minimum	Maximum	Mittelwert	Std.-Abweichung
Sportliche Betätigung	30	28	62	45,47	9,737
Gültige Werte (Listenweise)	30				

Abb. 3.4: Lage- und Streuungsmaße für die Variable „Sportliche Betätigung“ (v15_sport)

Auch die Berechnung von Lage- und Streuungsmaßen kann für mehrere Variablen gleichzeitig erfolgen, indem in der Dialogbox *Häufigkeiten* die gewünschten Merkmale mit Hilfe der Transportschaltfläche aus der Quellvariablenliste in das Feld *Variable(n)* befördert und die getroffenen Einstellungen mit dem Schalter *OK* bestätigt werden.

3.2 Bedingte Berechnung von Lage- und Streuungsmaßen

Wenn Lage- und Streuungsmaße nicht für alle Fälle in einer Datendatei berechnet werden sollen, sondern nur für eine definierte Auswahl an Fällen, dann sind diese zunächst über die temporäre Fallauswahl auszuwählen. Angenommen, man möchte die Lage- und Streuungsmaße für die Variable „Sportliche Betätigung“ (*v15_sport*) nur für Patienten über 60 Jahre berechnen. Für die temporäre Auswahl der Patienten über 60 Jahren ist im allgemeinen Menü die Auswahl *Daten* → *Fälle auswählen* vorzunehmen, so dass sich die Dialogbox *Fälle auswählen* öffnet (siehe **Abb. 2.4**).

In der Dialogbox *Fälle auswählen* ist die Auswahl *Alle Fälle* voreingestellt. Um die Fallauswahl auf die Patienten über 60 Jahre zu beschränken, ist die Einstellung *Falls Bedingung zutrifft* zu aktivieren und die Schaltfläche *Falls* zu betätigen. Daraufhin öffnet sich die Dialogbox *Fälle auswählen: Falls*. Die Variable *v11_alter* ist mit Hilfe der Transportschaltfläche aus der Quellvariablenliste in den Konditional-Editor zu bringen. Anschließend ist auf der Rechentastatur das „>“ und die „60“ zu aktivieren, so dass der folgende Ausdruck im Konditional-Editor steht: *v11_alter* > 60. Abschließend ist in der Dialogbox *Fälle auswählen: Falls* auf *Weiter* zu klicken, so dass sich die Dialogbox *Fälle auswählen* öffnet, in der die Einstellungen mit *OK* zu bestätigen sind.

Um nun die Lage- und Streuungsmaße für die sportliche Betätigung (*v15_sport*) der Subgruppe der über 60-Jährigen zu berechnen, ist nach einer der in Abschnitt 3.1 aufgezeigten Möglichkeiten zu verfahren. Sofern die Lage- und Streuungsmaße über die Menüauswahl *Analysieren* → *Deskriptive Statistiken* → *Deskriptive Statistik* erfolgt und alle Lage- und Streuungsmaße ausgewählt werden, erscheint im Viewer die folgende Abbildung:

Deskriptive Statistik					
	N	Minimum	Maximum	Mittelwert	Std.-Abweichung
Sportliche Betätigung	5	34	60	43,00	10,416
Gültige Werte (Listenweise)	5				

Abb. 3.5: Lage- und Streuungsmaße für die Variable „sportliche Betätigung“ (*v15_Sport*) der Subgruppe der über 60-Jährigen (> 60 Jahre)

Übungsaufgaben

- 3.1) Berechnen Sie unter Rückgriff auf die Datei *Zentrum1_Spieldaten.sav* für die Variable *v11_alter* die geeigneten Lage- und Streuungsmaße. Bitte führen Sie die Berechnung einmal für alle Patienten in der Datei durch und anschließend nur für die Personen, die angegeben haben, vor der Reha Raucher gewesen zu sein.
- 3.2) Berechnen Sie unter Rückgriff auf die Datei *Zentrum1_Spieldaten.sav* für die Variable *v14_2_rauch_nach* die geeigneten Lage- und Streuungsmaße.

Temporäre Fallauswahl

4 Erstellen von Grafiken

SPSS bietet dem Benutzer eine Reihe von Möglichkeiten, diverse Grafiken zu erstellen. Die Wahl der adäquaten Grafik für die prägnante Beschreibung eines Merkmals wird von seinem Messniveau sowie der Art und Anzahl der Merkmalsausprägungen bestimmt (siehe Abschnitt 1.4).

Beispieldatensatz

In diesem Kapitel wird die Erstellung der gängigsten Grafiken auf Basis der Daten-datei *Zentren_gesamt.sav* vorgenommen.

4.1 Balkendiagramm

Die Erstellung eines Balkendiagramms wird anhand der Variablen „Raucherstatus vor Reha“ (*v14_1_rauch_vor*) demonstriert. Die einfachste Möglichkeit dies in SPSS umsetzen, besteht darin, über den Menüpunkt *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten* die Dialogbox *Häufigkeiten* aufzurufen (siehe *Abb. 2.1*). In der Dialogbox *Häufigkeiten* ist dann mit der Transportschaltfläche die Variable *v14_1_rauch_vor* aus der Quellvariablenliste in das Feld *Variable(n)* zu bringen. Anschließend ist der Schalter *Diagramme* zu betätigen, so dass sich die Dialogbox *Häufigkeiten: Diagramme* öffnet (siehe *Abb. 4.1*). In dem Auswahl-feld *Diagrammtyp* hat der Benutzer nun die Wahl zwischen den gängigsten Typen an Diagrammen für die univariate Darstellung von Merkmalen einen Diagrammtyp auszuwählen. Um ein Balkendiagramm zu erstellen, ist der Punkt *Balkendiagramme* zu aktivieren. In dem Auswahl-feld *Diagrammwerte* kann gewählt werden, ob die Balken die Häufigkeiten der Merkmalsausprägungen oder die Prozentwerte der Merkmalsausprägungen anzeigen sollen. Für das Anwendungsbeispiel werden die *Häufigkeiten* ausgewählt. Um die Einstellungen zu bestätigen, ist der Schalter *Weiter* zu klicken, so dass man wieder in die Dialogbox *Häufigkeiten* zurückgelangt. Um die Höhe der Balken präziser interpretieren zu können, wird empfohlen, die Check-box *Häufigkeitstabellen anzeigen* bei Variablen mit Nominal- oder Ordinalniveau aktiviert zu lassen. Um die Prozedur zur Ausführung zu bringen, ist der Schalter *OK* zu betätigen, so dass im Viewer die entsprechende Ausgabe erscheint. Im Viewer erscheinen zunächst die in *Abbildung 4.2* dargestellten Tabellen.

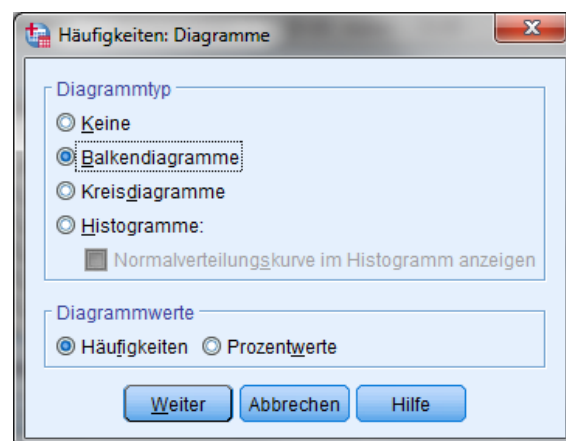


Abb. 4.1: Dialogbox – Häufigkeiten: Diagramme

Der Tabelle „Statistiken“ (siehe *Abb. 4.2*) ist zu entnehmen, auf Basis welcher Fälle die Grafik erstellt wurde. In dem Beispiel sind alle Fälle in die Erstellung der Grafik eingeflossen (Gültig=90).

Statistiken

Raucherstatus vor Reha

N	Gültig	90
	Fehlend	0

Raucherstatus vor Reha

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig Nichtraucher	64	71,1	71,1	71,1
Gelegenheitsraucher	9	10,0	10,0	81,1
Raucher	17	18,9	18,9	100,0
Gesamt	90	100,0	100,0	

Abb. 4.2: Statistiken und Häufigkeitstabelle für die Variable „Raucherstatus vor Reha“ (v14_1_rauch_vor)

Da die Checkbox Häufigkeitstabellen anzeigen aktiviert war, folgt auf die Tabelle „Statistiken“ eine Häufigkeitstabelle für die Variable „Raucherstatus vor Reha“ (siehe Abb. 4.2). Daran schließt sich dann das Balkendiagramm an (siehe Abb. 4.3). Die Höhe der Balken entspricht den absoluten Häufigkeiten der Merkmalsausprägung. Die einzelnen Balken tragen die Bezeichnung der jeweiligen Ausprägungen, wobei diese aufsteigend nach der Kodierung sortiert sind. Man sieht, dass die meisten Patienten und Patientinnen vor der Reha angegeben haben, dass sie Nichtraucher sind. Möchte man nun die genaue Anzahl an Personen wissen, ist ein Blick in die zugehörige Häufigkeitstabelle erforderlich. Durch das Editieren der Grafik (siehe Abschnitt 5.1) ist es z. B. auch möglich, die Häufigkeit direkt im Balkendiagramm anzeigen zu lassen.

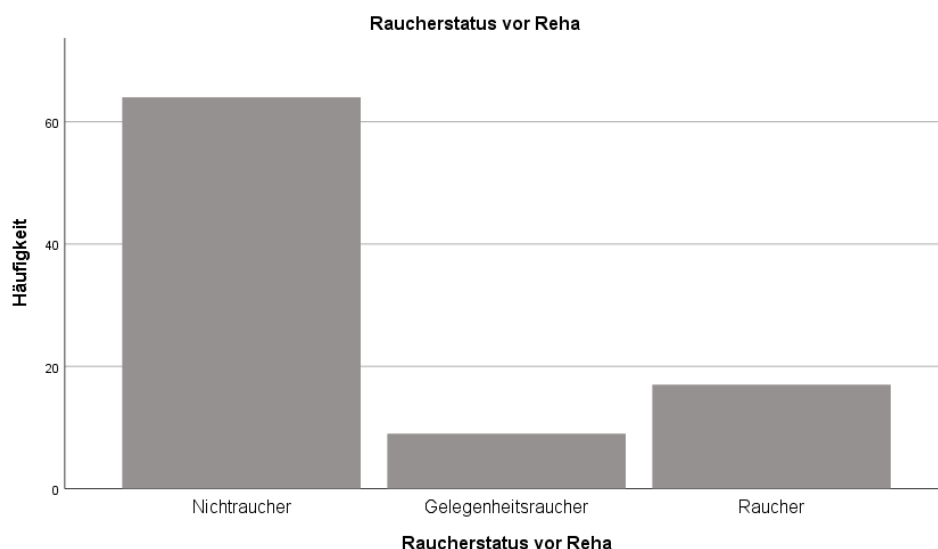
Interpretation Balkendiagramm

Abb. 4.3: Balkendiagramm für die Variable „Raucherstatus vor Reha“ (v14_1_rauch_vor)

4.2 Kreisdiagramm

Um zu demonstrieren, wie ein Kreisdiagramm in SPSS erstellt wird, wird auf die Variable „Anzahl wahrgenommener Freizeitangebote“ (v16_freizeit) zurückgegriffen. Die einfachste Möglichkeit in SPSS ein Kreisdiagramm zu erstellen, ist, genauso

vorzugehen wie bei der Erstellung eines Balkendiagramms. Das heißt, über den Menüpunkt *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten* ist die Dialogbox *Häufigkeiten* aufzurufen (siehe **Abb. 2.1**). Anschließend ist in der Dialogbox *Häufigkeiten* mit der Transportschaltfläche die Variable *v16_freizeit* aus der Quellvariablenliste in das Feld *Variable(n)* zu bringen und der Schalter *Diagramme* zu betätigen, so dass sich die Dialogbox *Häufigkeiten: Diagramme* öffnet (siehe **Abb. 4.1**).

In dem Auswahlfeld *Diagrammtyp* ist der Punkt *Kreisdiagramm* zu aktivieren. In dem Auswahlfeld *Diagrammwerte* soll für das Anwendungsbeispiel *Prozentwerte* ausgewählt werden. Um die Einstellungen zu bestätigen, ist der Schalter *Weiter* zu klicken, so dass man wieder in die Dialogbox *Häufigkeiten* zurückgelangt. Um die Breite der „Kuchenstücke“ präziser interpretieren zu können, wird empfohlen, die Checkbox *Häufigkeitstabellen anzeigen* aktiviert zu lassen. Um die Prozedur zur Ausführung zu bringen, ist der Schalter *OK* zu betätigen, so dass im Viewer die entsprechende Ausgabe erscheint.

Statistiken				
Anzahl wahrgenommener Freizeitangebote				
N	Gültig		90	
	Fehlend		0	

Anzahl wahrgenommener Freizeitangebote					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	bis 2	34	37,8	37,8	37,8
	3-6	42	46,7	46,7	84,4
	>6	14	15,6	15,6	100,0
	Gesamt	90	100,0	100,0	

Abb. 4.4: Statistiken und Häufigkeitstabelle für die Variable „Anzahl wahrgenommener Freizeitangebote“ (*v16_freizeit*)

Interpretation Kreisdiagramm

Im Viewer erscheinen zunächst die Tabelle „Statistiken“, der zu entnehmen ist, auf Basis welcher Fälle die Grafik erstellt wurde (siehe **Abb. 4.4**). In dem Beispiel sind alle Fälle in die Erstellung der Grafik eingeflossen (Gültig = 90). Danach ist die Häufigkeitstabelle für die entsprechende Variable aufgeführt. Darauf folgt das Kreisdiagramm (siehe **Abb. 4.5**). Es ist so aufgebaut, dass jede Merkmalsausprägung der Variablen, ein farbiges „Kuchenstück“ darstellt. Rechts oben neben der Grafik ist eine Legende abgedruckt, in der die Farben der einzelnen „Kuchenstücke“ erläutert werden. So kann man mit einem Blick erkennen, dass die meisten Patienten und Patientinnen angeben, 3–6 Freizeitangebote während der Reha in Anspruch genommen haben.

Ein Nachteil des so formatierten Kuchendiagramms ist, dass man dem Diagramm nicht ansehen kann, ob die Kuchenstücke den absoluten oder den prozentualen Häufigkeiten der Merkmalsausprägungen entsprechen. Dies kann aber im Nachhinein durch Editieren des Diagramms ergänzt werden.

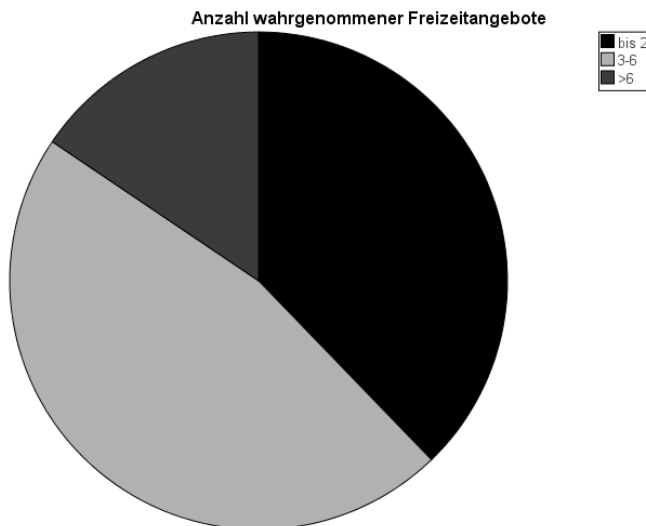


Abb. 4.5: Kreisdiagramm für die Variable „Anzahl wahrgenommener Freizeitangebote“ (v16_freizeit)

4.3 Histogramm

Die Erstellung eines Histogramms wird anhand der stetigen Variablen „Körpergewicht vor Reha“ (v13_1_gew_vor) präsentiert. Die einfachste Möglichkeit dies in SPSS umsetzen, besteht darin, über den Menüpunkt *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten* die Dialogbox *Häufigkeiten* aufzurufen (siehe Abb. 2.1).

In der Dialogbox *Häufigkeiten* ist dann mit der Transportschaltfläche die Variable v13_1_gew_vor aus der Quellvariablenliste in das Feld *Variable(n)* zu bringen. Anschließend ist der Schalter *Diagramme* zu betätigen, so dass sich die Dialogbox *Häufigkeiten: Diagramme* öffnet (siehe Abb. 4.1). In dem Auswahlfeld *Diagrammtyp* ist der Punkt *Histogramm* zu aktivieren. Da man für stetige Variablen in der Regel gern anhand des Histogramms beurteilen möchte, ob diese einer Normalverteilung folgen, gibt es zusätzlich die Möglichkeit, die Checkbox *Normalverteilungskurve im Histogramm anzeigen* zu aktivieren. Indem der Punkt *Histogramm* aktiviert wird, werden die Einstellungsmöglichkeiten in dem Auswahlfeld *Diagrammwerte* deaktiviert. Um die Einstellungen zu bestätigen, ist der Schalter *Weiter* zu klicken, so dass man wieder in die Dialogbox *Häufigkeiten* zurückgelangt. Um die Prozedur zur Ausführung zu bringen, ist der Schalter *OK* zu betätigen, so dass im Viewer die entsprechende Ausgabe erscheint.

Da bei stetigen Variablen in der Regel sehr viele unterschiedliche Werte in der Datendatei vorhanden sind, wird empfohlen, die Checkbox *Häufigkeitstabellen anzeigen* zu deaktivieren, da ansonsten immer die, u. U. sehr umfangreiche, Häufigkeitstabelle mit ausgegeben wird.

Hinweis

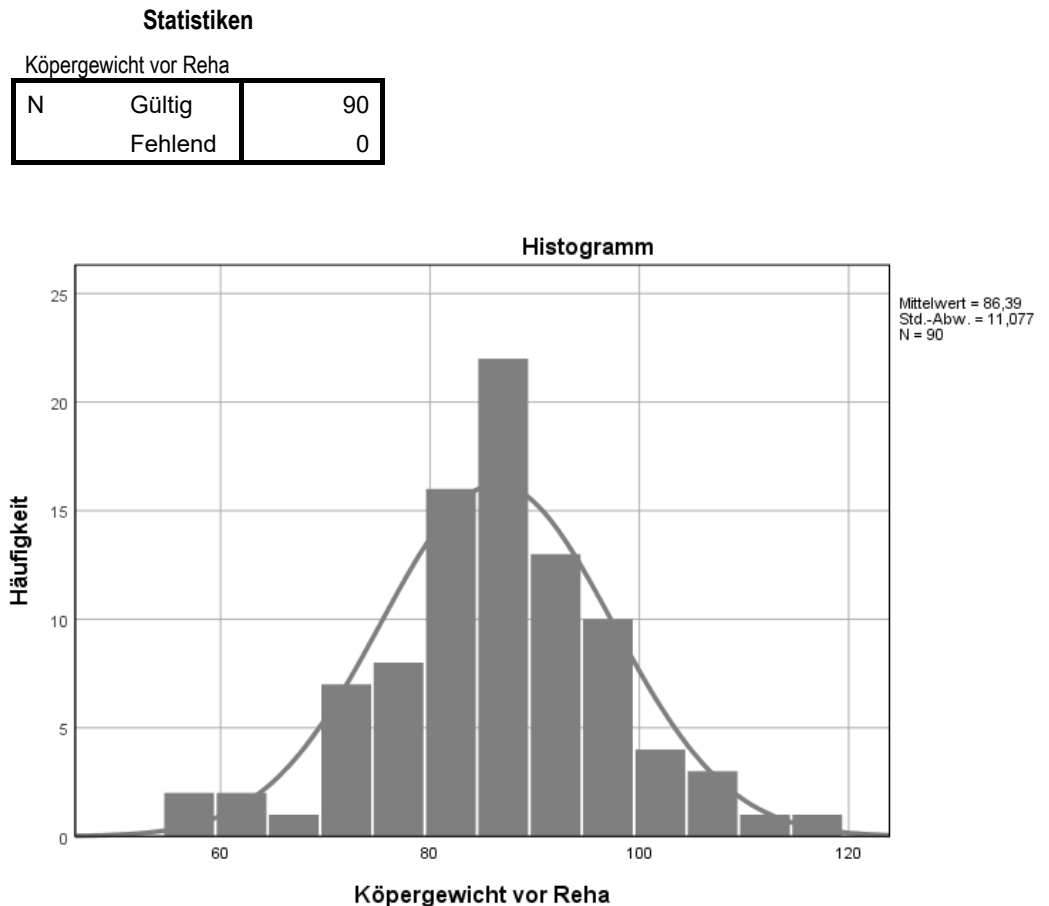


Abb. 4.6: Statistiken und Histogramm für die Variable „Körpergewicht vor Reha“ (v13_1_gew_vor)

Interpretation Histogramm

Im Viewer erscheint zunächst wieder die Tabelle „Statistiken“ für die Variable „Körpergewicht vor Reha“ (v13_1_gew_vor) um anzuzeigen, auf Basis welcher Fälle das Histogramm erstellt wurde (siehe **Abb. 4.6**). Der Tabelle „Statistiken“ ist zu entnehmen, dass das Histogramm auf Basis aller Fälle in der Datendatei erstellt worden ist (Gültig=90). Darauf folgt das Histogramm, das folgendermaßen aufgebaut ist: Auf der x-Achse sind die gleichgroßen Gewichtsklassen (hier: 5 kg) abgebildet, die SPSS automatisch für die Variable erstellt hat. Auf der y-Achse sind die Häufigkeiten abgetragen. Da die Gewichtsklassen gleich breit sind, kann anhand der Höhe der Balken pro Gewichtsklasse abgelesen werden, wie viele Patienten und Patientinnen in die jeweilige Gewichtsklasse fallen. Rechts neben dem Histogramm finden sich zudem für die Variable zusätzliche Informationen, wie der Mittelwert (= 86,39), die Standardabweichung (= 11,077) und die Datenbasis (N = 90). An dem Histogramm ist zu erkennen, dass die Variable „Körpergewicht vor Reha“ annähernd einer Normalverteilung folgt. Die meisten Patienten und Patientinnen befinden in der mittleren Gewichtskategorie (85–90 Kg). Das Histogramm fällt jeweils rechts und links ab und ist symmetrisch.

4.4 Boxplot

Grafische Darstellung von stetigen Merkmalen

Für die grafische Darstellung von stetigen Merkmalen sind neben Histogrammen auch Boxplots besonders gut geeignet. Um zu demonstrieren, wie ein Boxplot in SPSS erstellt wird, wird auf die Variable „Sportliche Betätigung“ (*v15_sport*) zurückgegriffen. Dazu ist zunächst über den Menüpunkt *Analysieren* → *Deskriptive Statistiken* → *Explorative Datenanalyse* die Dialogbox *Explorative Datenanalyse* aufzurufen (siehe **Abb. 4.7**). In der Dialogbox ist dann mit der Transportschaltfläche die Variable *v15_sport* aus der Quellvariablenliste in das Feld *Abhängige Variablen* zu bringen. Die Felder *Faktorenliste* und *Fallbeschriftung* bleiben leer. Sofern ausschließlich ein Boxplot für die ausgewählte Variable angelegt werden soll, ist in dem Auswahlfeld *Anzeige* der Punkt *Diagramme* zu aktivieren. Anschließend ist der Schalter *Diagramme* zu betätigen, so dass sich die Dialogbox *Explorative Datenanalyse: Diagramme* öffnet. Dort ist im Auswahlfeld *Deskriptiv* die Checkbox *Stengel-Blatt* zu deaktivieren. Diese Einstellung ist mit dem Schalter *Weiter* zu bestätigen. Um nun die Prozedur zur Ausführung zu bringen, ist in der Dialogbox *Explorative Datenanalyse* die Taste *OK* zu klicken, so dass im Viewer die entsprechende Ausgabe erscheint.

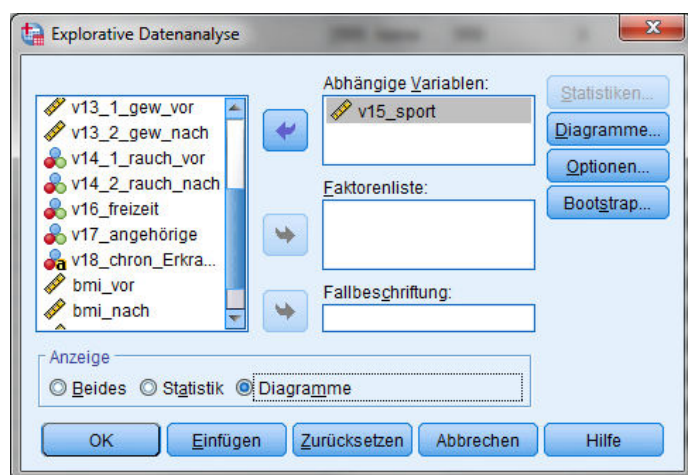


Abb. 4.7: Dialogbox – Explorative Datenanalyse

Im Viewer erscheint zunächst die Tabelle „Verarbeitete Fälle“ (siehe **Abb. 4.8**). Dieser Tabelle ist zu entnehmen, wie viele Fälle in die Erstellung des Boxplots eingeflossen sind. Da alle Patienten und Patientinnen Angaben bei der Variablen „Sportliche Betätigung“ gemacht haben, wurde der Boxplot auf Basis aller Fälle in der Datei erstellt (Gültig: N=90).

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
Sportliche Betätigung	90	100,0%	0	0,0%	90	100,0%

Abb. 4.8: Verarbeitete Fälle für die Variable „sportliche Betätigung“ (*v15_sport*)

Interpretation Boxplot

Auf die Tabelle „Verarbeitete Fälle“ folgt die Abbildung des Boxplots (siehe **Abb. 4.9**). Die Grafik zu dem Boxplot ist in SPSS so aufgebaut, dass auf der x-Achse angegeben ist, auf welche Variable sich der Boxplot bezieht (hier: Sportliche Betätigung). Auf der y-Achse ist die Skala für die Merkmalsausprägungen der Variablen abgetragen (hier: Minuten). Der Boxplot selbst ist so aufgebaut, dass die dicke schwarze Linie in der gelben Box den Median repräsentiert. Der obere Rand der Box ist das obere Quartil und der untere Rand ist das untere Quartil. Das heißt, innerhalb der Box liegen die mittleren 50 % der Werte. Das Ende der oberen „Schnurrhaare“ (= *Whisker*) gibt das Maximum an, sofern es keine **Ausreißer** gibt. Als Ausreißer werden in SPSS die Werte definiert, die mindestens 1,5-mal von dem Interquartilsabstand vom Median entfernt liegen. Bei der Variable *v15_sport* gibt es einen Ausreißer, der als Kreis gekennzeichnet ist und rechts neben sich eine 36 stehen hat. Diese Zahl gibt die Zeilennummer des Falls im Viewer an, der den Ausreißer darstellt. Das heißt, der Ausreißer entspricht bei der Variablen „Sportliche Betätigung“ dem Maximum. Das Ende der unteren „Schnurrhaare“ entspricht in dem Beispiel dem Minimum.

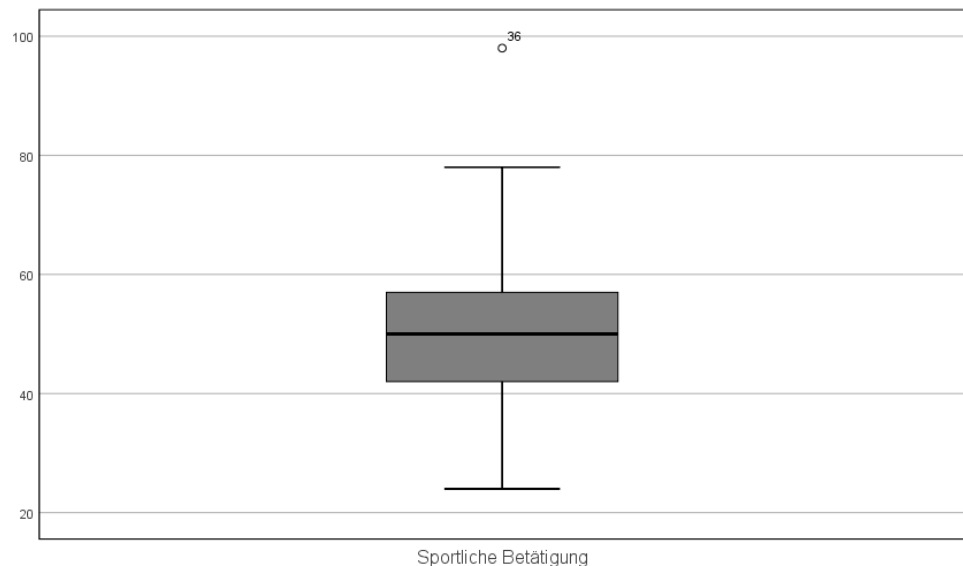


Abb. 4.9: Boxplot für die Variable „Sportliche Betätigung“ (*v15_sport*)

Anhand des Boxplots ist es ebenfalls möglich, eine Variable auf **Normalverteilung** zu untersuchen. Diese ist dann annähernd gegeben, wenn der Median ungefähr mittig in der gelben Box liegt, die oberen und unteren „Schnurrhaare“ ungefähr gleich lang sind und sofern Ausreißer vorhanden sind, diese sich gleichmäßig auf beide Seiten verteilen. Da dies auf den vorliegenden Boxplot ziemlich gut zutrifft, kann die Variable „Sportliche Betätigung“ als normalverteilt betrachtet werden.

Da es schwierig ist, direkt aus dem Boxplot die tatsächlichen Werte für den Median, die Quartile und das Minimum bzw. Maximum abzulesen, wird empfohlen, zusätzlich zu dem Boxplot die deskriptiven Statistiken ausgeben zu lassen. Dazu ist in der Dialogbox *Explorative Datenanalyse* bei dem Auswahlfeld *Anzeige* der Punkt *Beides* zu aktivieren und die Einstellung mit der Taste *OK* zu bestätigen. Daraufhin erscheint im Viewer zwischen der Tabelle „Verarbeitete Fälle“ und dem „Boxplot“ die Tabelle „Deskriptive Statistik“ (siehe **Abb. 4.10**). In der zweiten Spalte der Tabelle sind u. a. die bekannten Lage- und Streuungsmaße für die Variable „Sportliche Betätigung“ aufgelistet. Das heißt, damit wurde eine dritte Möglichkeit aufgezeigt, wie sich Lage- und Streuungsmaße in SPSS berechnen lassen.

Deskriptive Statistik

		Statistik	Std.-Fehler
Sportliche Betätigung	Mittelwert	50,14	1,259
	95% Konfidenzintervall des Mittelwerts	Untergrenze Obergrenze	47,64 52,65
	5% getrimmtes Mittel	49,71	
	Median	50,00	
	Varianz	142,732	
	Std.-Abweichung	11,947	
	Minimum	24	
	Maximum	98	
	Spannweite	74	
	Interquartilbereich	15	
	Schiefe	,756	,254
	Kurtosis	2,102	,503

Abb. 4.10: Deskriptive Statistik für die Variable „Sportliche Betätigung“

In diesem Kapitel wurden die einfachsten Möglichkeiten demonstriert, unterschiedliche Diagramme in SPSS zu erstellen. Der Großteil lässt sich über die Dialogbox *Häufigkeiten* realisieren. Dies hat den Vorteil, dass man sich über diese Dialogbox gleichzeitig Häufigkeitstabellen sowie Lage- und Streuungsmaße ausgeben lassen kann.

Eine weitere intuitive Möglichkeit, Diagramme in SPSS zu erstellen, besteht über den Menüpunkt *Diagramme* → *veraltete Dialogfenster*. Dort werden dem Benutzer neben Kreisdiagrammen, Balkendiagrammen, Histogrammen und Boxplots auch weitere Diagramme zur grafischen Darstellung von mehr als einer Variablen angeboten.

Genau wie für die Erstellung von Häufigkeitstabellen und die Berechnung von Lage- und Streuungsmaßen, ist es auch für die Erstellung von Diagrammen in SPSS möglich, dies auf eine bestimmte Auswahl an Fällen zu begrenzen. Dazu ist die temporäre Fallauswahl zu nutzen (siehe Abschnitt 2.2 und 3.2).

Weitere Optionen zur Diagrammerstellung

Übungsaufgaben

- 4.1) Erstellen Sie auf Basis der Datendatei *Zentren_gesamt.sav* ein Kreisdiagramm auf Basis der relativen Häufigkeiten für die Variable *v17_angehörige*.
- 4.2) Erstellen Sie auf Basis der Datendatei *Zentren_gesamt.sav* ein Histogramm für die Variable *v12_körpergröße*. Liegt eine Normalverteilung vor?
- 4.3) Erstellen Sie auf Basis der Datendatei *Zentren_gesamt.sav* einen Boxplot für die Variable *v13_2_gew_nach*. Beschreiben Sie die Verteilung der Variablen *v13_2_gew_nach*.

5 Editieren von Tabellen und Grafiken

SPSS bietet die Möglichkeit, Tabellen und Grafiken nach ihrer Erstellung entsprechend den individuellen Wünschen des Benutzers anzupassen. Aufgrund des begrenzten Umfangs des vorliegenden Studienbriefs kann dies nicht für jede einzelne Tabelle und Grafik in SPSS demonstriert werden. Stattdessen wird sich darauf beschränkt, die Grundfunktionen in SPSS aufzuzeigen, die zum Editieren von Tabellen (Abschnitt 5.1) und Grafiken (Abschnitt 5.2) genutzt werden können.

Beispieldatensatz

Alle Berechnungen in diesem Kapitel basieren auf der Datendatei *Zentren_gesamt.sav*.

5.1 Editieren von Tabellen

Um eine in SPSS erstellte Tabelle zu editieren ist die entsprechende Tabelle im Viewer mit der linken Maustaste zu aktivieren. Die erfolgreiche Aktivierung der Tabelle wird dem Benutzer durch einen roten Pfeil links neben der Tabelle angezeigt. Anschließend ist die rechte Maustaste zu betätigen, so dass sich das Menüfenster öffnet (siehe *Abb. 5.1*). Dort ist dann der Menüpunkt *Inhalt bearbeiten* auszuwählen. SPSS bietet dem Benutzer die Möglichkeit, die Tabelle im Viewer oder in einem separaten Fenster zu bearbeiten.

Anzahl wahrgenommener Freizeitangebote

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig bis 2	34	37,8	37,8	37,8
3-6	42	46,7	46,7	84,4
>6	14	15,6	15,6	100,0
Gesamt	90	100,0	100,0	

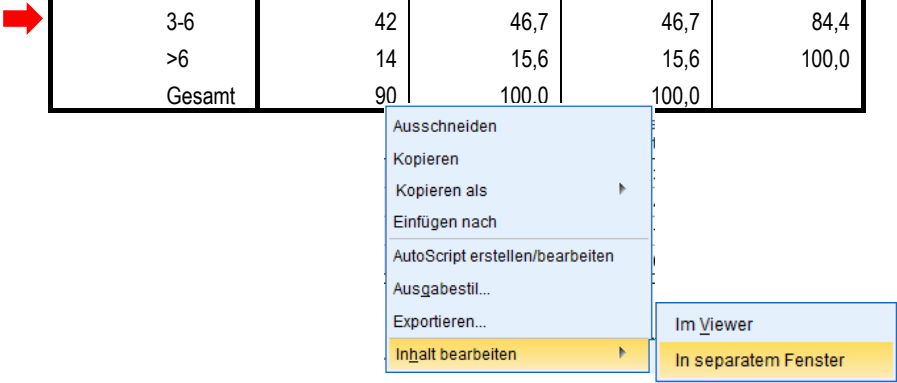


Abb. 5.1: Editieren von Häufigkeitstabellen

Editieren in separatem Fenster

Sofern die Variante *In separatem Fenster* gewählt wurde, öffnet sich ein separates Fenster zum Editieren (siehe *Abb. 5.2*). Hier stehen dem Benutzer diverse Möglichkeiten zum Anpassen der Tabelle zur Verfügung. Über die folgende Menüauswahl *Format* → *Tabelleneigenschaften*, ist es möglich, das Tabellenformat zu ändern (z. B. Breite und Höhe der Zeilen und Spalten, Breite und Art des Rahmens). Wenn man zunächst Zellen markiert und dann auf die Menüauswahl *Format* → *Zelleneigenschaften* geht, können Farbe etc. von Zellen verändert werden. Wichtig ist, jede durchgeführte Änderung mit *OK* zu bestätigen, damit die Einstellungen in das Tabellenlayout übernommen werden.

Pivot-Tabelle Anzahl wahrgenommener Freizeitangebote					
Anzahl wahrgenommener Freizeitangebote					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	bis 2	34	37,8	37,8	37,8
	3-6	42	46,7	46,7	84,4
	>6	14	15,6	15,6	100,0
	Gesamt	90	100,0	100,0	

Abb. 5.2: Pivot-Tabelle

5.2 Editieren von Grafiken

Um in SPSS erstellte Grafiken zu editieren, muss der sog. Diagramm-Editor geöffnet werden. Dazu ist entweder ein Doppelklick mit der rechten Maustaste auf die zu editierende Grafik erforderlich oder die Grafik muss mit der linken Maustaste aktiviert werden, so dass links neben ihr ein roter Pfeil erscheint. Bei der ersten Variante öffnet sich der Diagramm-Editor direkt. Bei der zweiten Variante muss erst noch die rechte Maustaste geklickt werden, so dass sich das Kontextmenü öffnet (siehe **Abb. 5.3**). Dort ist die folgende Menüauswahl zu treffen: *Inhalt bearbeiten* → *In separatem Fenster*, so dass sich der Diagramm-Editor öffnet.

Diagramm-Editor

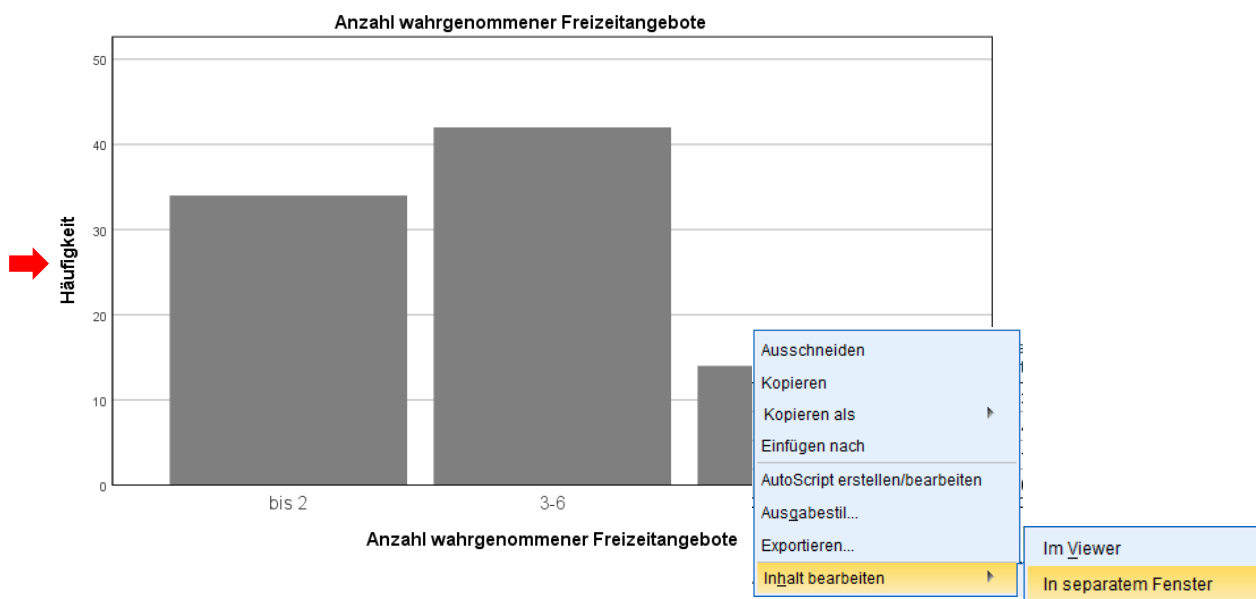


Abb. 5.3: Editieren von Grafiken

In dem Diagramm-Editor stehen dem Benutzer nun eine Reihe von Optionen zum Editieren der Grafik zur Verfügung. Um z. B. die **Häufigkeiten pro Balken** anzeigen zu lassen, ist ein beliebiger Balken mit der linken Maustaste so anzuklicken, dass alle Balken automatisch gelb umrandet werden. Anschließend ist die rechte Maustaste zu klicken, so dass sich das Kontextmenü öffnet. Im Kontextmenü ist der vorletzte Menüpunkt *Elemente* auszuwählen.

Um die **Farben der Balken** zu verändern, ist im Viewer ebenfalls ein beliebiger Balken anzuklicken, so dass sich alle Balken automatisch gelb umranden. Anschließend kann die rechte Maustaste geklickt werden, so dass das Kontextmenü angezeigt wird (siehe *Abb. 5.4*). Dort ist mit dem Mauszeiger auf den zweiten Menüpunkt von oben zu gehen: *Auswählen*, so dass sich die drei Unterpunkte *Dies Balken*, *Diese Gruppe von Balken*, *Alle Balken* öffnen. Hier ist der Punkt *Alle Balken* auszuwählen.¹

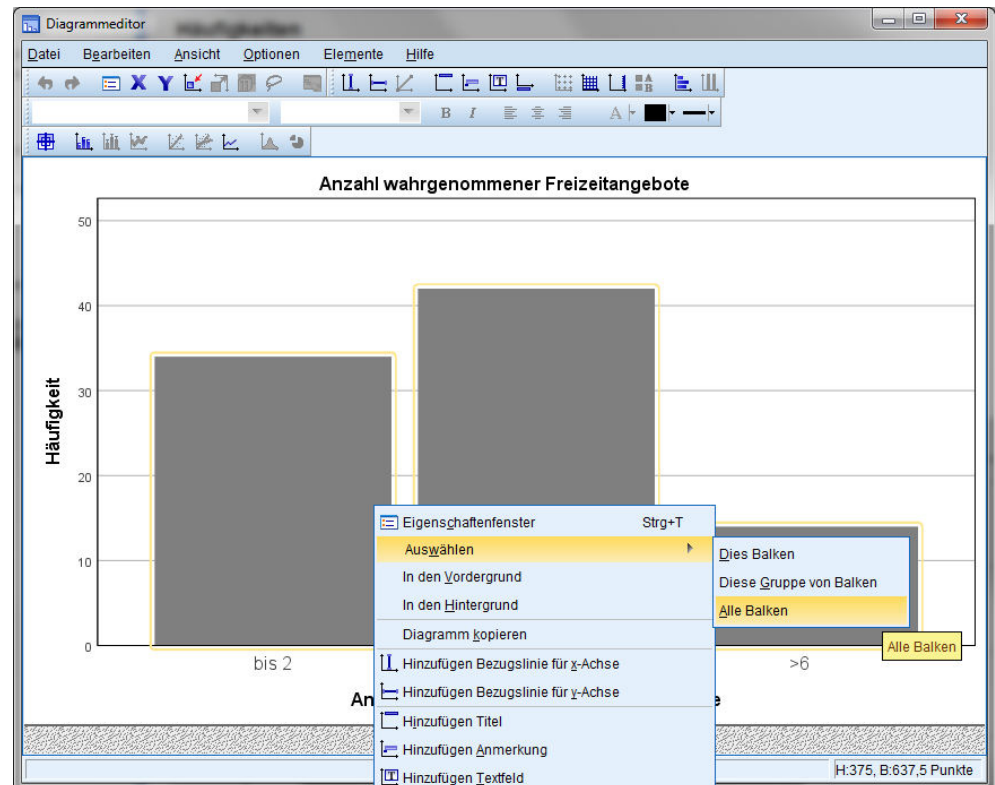


Abb. 5.4: Änderung der Farben in Diagrammen

Daraufhin öffnet sich das Fenster *Eigenschaften*, das über verschiedene Registerkarten die Optionen zur Veränderung der Balken anzeigt. Um nun die Farbe anzupassen, ist auf die Registerkarte *Füllung und Rahmen* zu gehen. Dort wird dem Benutzer eine Farbpalette angeboten, so dass die gewünschte Farbe ausgewählt werden kann. Möchte man die Balken zusätzlich noch mit einem **3-D-Effekt** versehen, dann ist auf die Registerkarte *Tiefe und Winkel* zu gehen. Dort ist bei dem Auswahlkasten *Effekt* auf *3D* zu klicken. Jede Einstellung, die man bei einer Registerkarte getroffen hat, ist mit dem Schalter *Zuweisen* zu bestätigen, damit die Anpassung übernommen wird (siehe *Abb. 5.5*).

¹ **Technischer Hinweis:** Die falsche Rechtschreibung findet sich tatsächlich so in SPSS und ist vermutlich der Übersetzung aus dem Englischen geschuldet.

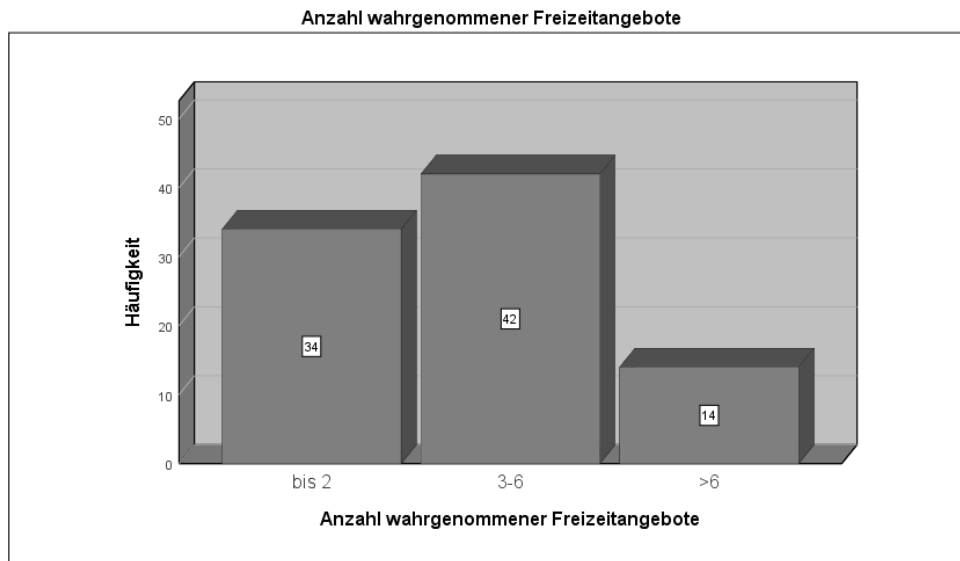


Abb. 5.5: Diagramm-Editor (v14_1_rauch_vor)

Übungsaufgaben

- 5.1) Editieren Sie das Kreisdiagramm zu der Variablen *v17_angehörige* (siehe Übungsaufgabe 4.1) so, dass die relativen Häufigkeiten direkt im Kreisdiagramm eingefügt werden.
- 5.2) Editieren Sie das Kreisdiagramm zu der Variablen *v17_angehörige* (siehe Übungsaufgabe 4.1) so, dass links unten im Kreisdiagramm „Datenbasis: n = 90“ eingefügt wird.
- 5.3) Vertauschen Sie auf Basis der Datei *Zentrum1_Spieldaten.sav* in der Häufigkeitstabelle zu der Frage *v3_zuf* die Zeilen und Spalten mit Hilfe der Editierfunktion von SPSS.

6 Vergleich dichotomer Variablen

In diesem Kapitel werden Schätz- und Testverfahren vorgestellt, die bei dem Vergleich von zwei Gruppen (z. B. Männer vs. Frauen) in Bezug auf eine dichotome Variable zur Anwendung kommen. Dichotome Variablen sind dadurch gekennzeichnet, dass sie nur zwei mögliche Ausprägungen haben und sich Raten bzw. Anteile als Kenngrößen zur Beschreibung dieser Variablen eignen. Raten und ihre Bedeutung lassen sich auch im alltäglichen Sprachgebrauch finden: z. B. wenn die Rede davon ist, dass 64 von insgesamt 90 Teilnehmer(inne)n der Reha-Maßnahme Frauen sind (zur Berechnung von Raten siehe Formelverzeichnis 1.2). Multipliziert man eine Rate (hier: $64/90$) mit 100 %, erhält man einen Anteil (hier: 71,1 %). Andere Berechnungsmöglichkeiten bei dichotomen oder auch nominalen Variablen wurden bereits im Modul „Empirische Methoden I“ vorgestellt. Diese Methoden sollen in diesem Kapitel in SPSS angewandt und interpretiert werden.

6.1 Erstellen von Kreuztabellen

Eine übersichtliche und anschauliche Darstellung für den Vergleich von zwei Gruppen hinsichtlich der Verteilung einer dichotomen Variablen, stellen Kreuztabellen dar. Der Vergleich von zwei Gruppen bezüglich einer dichotomen Variablen mit Hilfe einer Kreuztabelle ist eine Möglichkeit der deskriptiven Beschreibung des Unterschieds. Um sich einen Überblick zu verschaffen, sollte eine deskriptive Betrachtung stets der erste Schritt sein, bevor Verfahren der induktiven Statistik zur Anwendung kommen. Im Folgenden wird das Aufstellen einer Kreuztabelle erläutert, die den Vergleich von Frauen und Männern hinsichtlich des Auftretens von chronischen Erkrankungen (ja/nein) zum Ziel hat. Dies stellt den ersten Analyseschritt dar, um die Forschungsfrage 4 „Unterscheiden sich Männer und Frauen bezüglich der Rate an chronischen Erkrankungen?“ zu beantworten.

Dichotomisierung von Variablen

Da die Variable chronische Erkrankungen (*v18_chron_Erkrankung*) in der Datendatei *Zentren_gesamt.sav* bisher nur als String-Variable vorliegt, muss diese zunächst umkodiert (d. h. dichotomisiert) werden. Es bietet sich an, immer wenn eine chronische Erkrankung vorliegt, der neuen Variable (hier genannt: *chron_Erkr_dich*) eine 1 zuzuordnen. Wenn in *v18_chron_Erkrankung* nichts eingetragen ist, ist eine 2 zu kodieren (siehe Kapitel 4 in Studienbrief 1 zum Umkodieren von Variablen).

Für das Erstellen einer Kreuztabelle in SPSS ist zunächst die Menüauswahl *Analisieren* → *deskriptive Statistik* → *Kreuztabellen* zu treffen. Anschließend öffnet sich das Dialogfenster, wie es in **Abb. 6.1** dargestellt ist. Um eine zielgerichtete Interpretation der Kreuztabelle für dichotome Variablen zu ermöglichen, wird empfohlen, diese stets nach dem folgenden Schema aufzubauen: In den Zeilen sollte immer die unabhängige Variable stehen. In dem gewählten Beispiel ist dies die Gruppierungsvariable *v10_geschlecht*. In den Spalten sollte entsprechend immer die abhängige Variable stehen, was in dem Beispiel die neu erstellte Variable *chron_Erkr_dich* ist. Dann ist auf *Zellen* zu klicken und alle Voreinstellungen sind so zu belassen und zusätzlich sind in dem Abschnitt *Prozentwerte* die Zeilenprozente (*Zeilenweise*) zu aktivieren. Die Einstellungen sind zunächst mit *Weiter* zu bestätigen, bevor die Prozedur dann abschließend mit *OK* ausgeführt werden muss.

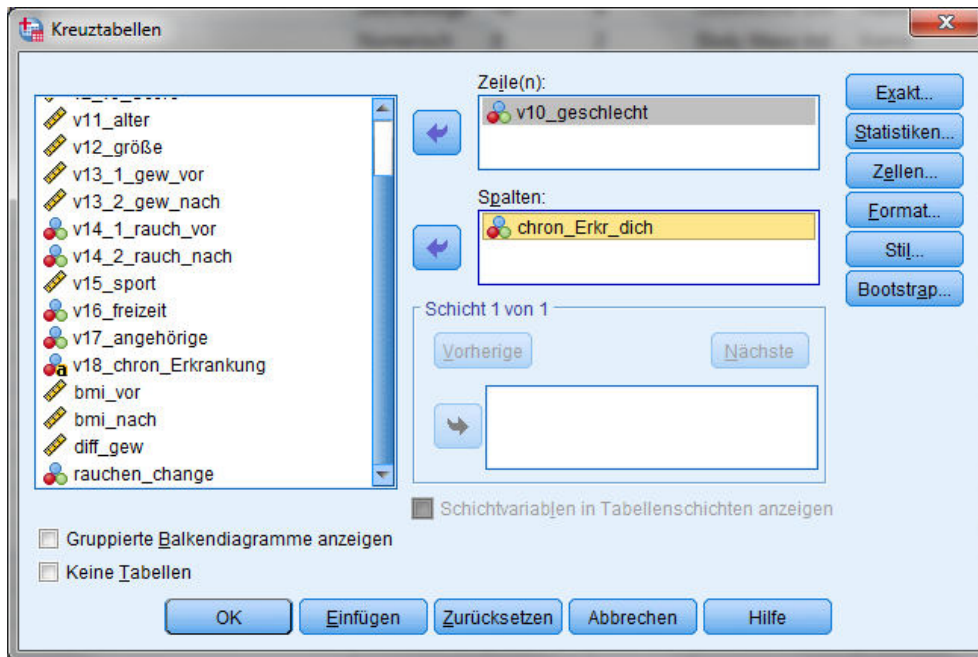


Abb. 6.1: Auswahl der Variablen für eine Kreuztabelle

In dem Ausgabefenster erscheinen dann die Tabellen in **Abb. 6.2**

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
Geschlecht * chron_Erkr_dich	90	100,0%	0	0,0%	90	100,0%

Geschlecht * chron_Erkr_dich Kreuztabelle

			chron_Erkr_dich		Gesamt
			nein	ja	
Geschlecht	weiblich	Anzahl	59	5	64
		% innerhalb von Geschlecht	92,2%	7,8%	100,0%
	männlich	Anzahl	23	3	26
		% innerhalb von Geschlecht	88,5%	11,5%	100,0%
Gesamt		Anzahl	82	8	90
		% innerhalb von Geschlecht	91,1%	8,9%	100,0%

Abb. 6.2: Ausgabefenster für Kreuztabelle

Die erste Tabelle „Verarbeitete Fälle“ bietet einen Überblick über die Anzahl gültiger und fehlender Werte. Aus der Tabelle geht hervor, dass es bei den beiden zu analysierenden Variablen für alle Fälle gültige Werte gibt (Spalte Fehlend – N = 0). Die Tabelle „Geschlecht * chron. Erkrankung dichotomisiert Kreuztabelle“ zeigt die Kreuztabelle für die beiden ausgewählten Variablen. Der gewählte Aufbau der Tabelle ermöglicht es, den Anteil der Frauen mit chronischen Erkrankungen (7,8 %) mit dem Anteil der Männer mit chronischen Erkrankungen (11,8 %) direkt zu vergleichen. In der Stichprobe zeigt sich also, dass der Anteil an Männern mit

chronischen Erkrankungen im Vergleich zu den Frauen leicht höher ist. Ob sich dieses Ergebnis auch auf die Grundgesamtheit übertragen lässt, soll nun mit Hilfe von Schätz- und Testverfahren überprüft werden.

6.2 Durchführung des Chi-Quadrat-Tests

Eine Möglichkeit, um zu prüfen, ob sich der in der Stichprobe gefundene Unterschied zwischen Männern und Frauen bezüglich des Anteils an Patient(inn)en mit chronischen Erkrankungen gegen den Zufall absichern und auf die Grundgesamtheit übertragen lässt, stellt die Anwendung eines statistischen Tests dar. Für den Vergleich von Anteilen bzw. Raten zwischen zwei Gruppen wird als Test der Chi-Quadrat-Test gewählt. Dieser testet, ob sich zwei oder mehr Gruppen hinsichtlich eines Merkmals signifikant voneinander unterscheiden (vgl. *Rasch et al. 2010b: 185*).

Anwendungsbereich

Der Chi-Quadrat-Test kann auf 2x2-Kreuztabellen oder Variablen mit mehr als zwei Ausprägungen angewendet werden. An dieser Stelle wird er jedoch nur für das bereits eingeführte Beispiel präsentiert, das zwei Variablen (`v10_geschlecht` und `chron_Erkr_dich`) mit je zwei Ausprägungen (`v10_geschlecht`: weiblich = 1 und männlich = 0; `chron_Erkr_dich`: 1 = ja und 0 = nein) enthält. Die Berechnung der Teststatistik zum χ^2 -Test erfolgt in SPSS wie im Formelverzeichnis dargestellt (siehe dort Formel 2.2).

Das Hypothesenpaar zu der Fragestellung „Unterscheiden sich Männer und Frauen bezüglich der Rate an chronischen Erkrankungen?“ lautet:

Formulieren des Hypothesenpaars

H_0 : Es gibt keinen Unterschied bezüglich der Rate an chronischen Erkrankungen zwischen Männern und Frauen.

H_1 : Es gibt einen Unterschied bezüglich der Rate an chronischen Erkrankungen zwischen Männern und Frauen.

Formal lassen sich die Hypothesen folgendermaßen beschreiben:

$$H_0: p_{\text{chronF}} = p_{\text{chronM}}$$

$$H_1: p_{\text{chronF}} \neq p_{\text{chronM}}$$

Das Ziel des χ^2 -Tests besteht nun darin, zu prüfen, ob die H_0 zu einem vorher festgesetzten Signifikanzniveau (üblicherweise von 5 %) zu Gunsten der H_1 abgelehnt werden kann. Für das Ausführen des χ^2 -Tests ist die folgende Prozedur zu wählen: *Analysieren* → *Deskriptive Statistiken* → *Kreuztabellen*. Für das ausgewählte Beispiel sind zunächst alle Schritte und Einstellungen exakt so zu wählen, wie in Kapitel 6.1 erläutert. Zusätzlich ist die Option *Statistiken* zu wählen, und mit Hilfe der Klickbox *Chi-Quadrat* zu aktivieren (vgl. **Abb. 6.3**).

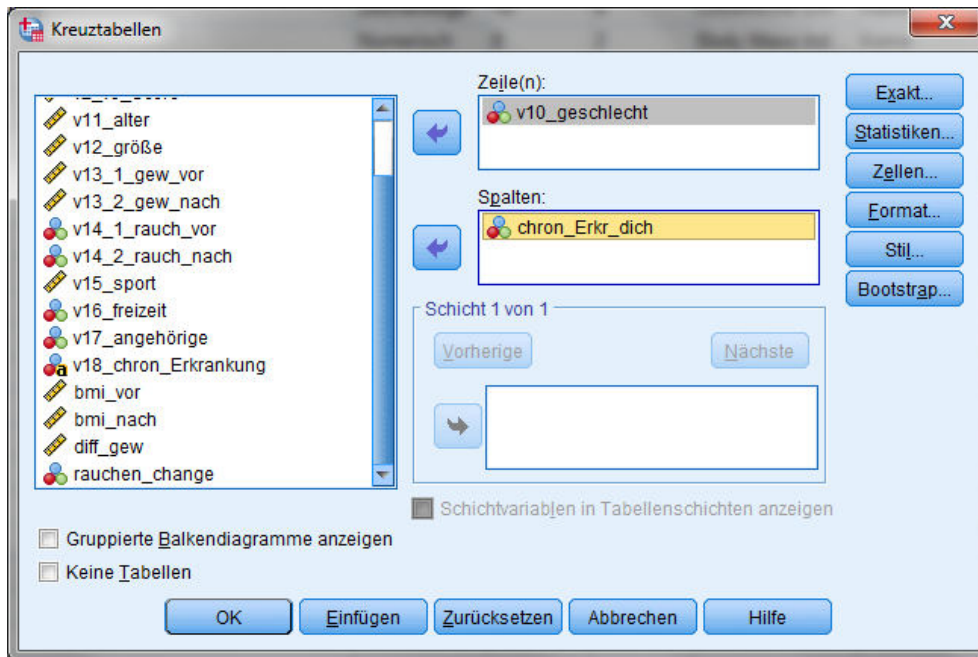


Abb. 6.3: SPSS-Fenster – Auswahl der Variablen

Im Fenster „Kreuztabellen: Statistiken“ muss nun der Haken bei Chi-Quadrat gesetzt werden. Mit dem Klick auf *Weiter* und *OK* wird die Berechnung der Teststatistik ausgeführt und die Ergebnisse ausgegeben (siehe Abb. 6.4).

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	,317 ^a	1	,573		
Kontinuitätskorrektur ^b	,024	1	,877		
Likelihood-Quotient	,303	1	,582		
Exakter Test nach Fisher				,686	,420
Zusammenhang linear-mit-linear	,313	1	,576		
Anzahl der gültigen Fälle	90				

a. 1 Zellen (25,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 2,31.

b. Wird nur für eine 2x2-Tabelle berechnet

Abb. 6.4: Ergebnisse des Chi-Quadrat-Tests

Den Ergebnissen des „Chi-Quadrat-Tests“ kann man verschiedene Berechnungsmöglichkeiten für die Teststatistik entnehmen. Am gebräuchlichsten ist die Teststatistik nach Pearson (Bühl, 2010).

Der Wert der Teststatistik nach Pearson beträgt in dem gewählten Beispiel 0,317. Es gibt einen Freiheitsgrad (da es auch nur zwei Ausprägungen pro Variable gibt) und der p-Wert (hier mit Asymptotische Signifikanz (2-seitig) bezeichnet) beträgt 0,573. Das heißt, das Ergebnis des χ^2 -Tests ergibt kein signifikantes Ergebnis, da der p-Wert mit 0,573 größer als 0,05 ist. Strenggenommen, darf der χ^2 -Test in dem Beispiel gar nicht angewendet werden, da die Durchführungsbedingungen für diesen Test nicht erfüllt sind. So ist die Durchführung des Tests an die Bedingung geknüpft, dass keine der Zellen, eine erwartete Häufigkeit kleiner 5 haben darf. Ob gegen diese Voraussetzung verstoßen wird, erfährt man in der SPSS-Ausgabe in Fußnote a. Dort

Chi-Quadrat-Test nach Pearson

steht für das Anwendungsbeispiel, dass „1 Zellen (25,0 %) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 2,31“. Da damit die Voraussetzung für die Testdurchführung verletzt ist, muss ein alternatives Testverfahren zum Einsatz kommen. Immer dann wenn der χ^2 -Test nicht durchgeführt werden darf und eine 2x2 Tabelle (Vierfeldertafel) vorliegt, darf der Exakte Test nach Fisher zur Anwendung kommen. Die Ergebnisse dieses Tests werden von SPSS automatisch bei 2x2 Tabellen mit ausgegeben. Der p-Wert des Exakten Tests nach Fisher beträgt in dem Beispiel 0,686. Das heißt, auch dieser Test zeigt ein nicht signifikantes Ergebnis für die betrachtete Fragestellung. Folglich darf die H_0 Hypothese nicht zu Gunsten der H_1 abgelehnt werden.

Übungsaufgaben

- 6.1) Es soll untersucht werden, ob sich Personen, bei denen Angehörige in die Reha-Maßnahme mit einbezogen wurden, bezüglich ihrer Zufriedenheit unterscheiden. Der Einbezug von Angehörigen soll über die Variable *v17_angehörige* (1=ja/0=nein) operationalisiert werden. die Zufriedenheit soll über eine Dichotomisierung der Variablen *v2_v9_Score* anhand des Medians erfolgen. Bitte führen Sie deshalb zunächst eine Dichotomisierung der Variablen *v2_v9_Score* anhand des Medians durch, indem Sie die neue Variable *score_dich* bilden. Dabei sollen hohe Zufriedenheitswerte im Score (d. h. > Median) mit „1 = zufrieden“ kodiert und geringe Wert im Score (d. h. <= Median) mit „2 = unzufrieden“ kodiert werden.
- 6.2) Stellen Sie die Forschungshypothesen für die Fragestellung „Unterscheiden sich Personen, bei denen Angehörige in die Reha-Maßnahme mit einbezogen wurden bezüglich ihrer Zufriedenheit?“ auf.
- 6.3) Erstellen Sie eine Kreuztabelle zur deskriptiven Beschreibung des Gruppenunterschieds (Einbezug von Angehörigen ja/nein) in Bezug auf die Zufriedenheit und interpretieren Sie diese.
- 6.4) Führen Sie einen χ^2 -Test durch, um die Gültigkeit der H_0 -Hypothese zu prüfen und interpretieren Sie das Testergebnis.

7 Schätz- und Testverfahren zum Vergleich stetiger Variablen

Wenn stetige Variablen vorliegen, kommt als statistisches Testverfahren der t -Test in Betracht. In diesem Kapitel werden die drei Varianten des t -Tests vorgestellt, die sich aus der Anzahl der miteinander zu vergleichenden Stichproben oder Variablen erklären (t -Test bei einer Stichprobe, bei zwei unabhängigen Stichproben sowie bei zwei abhängigen Stichproben). In SPSS wird der t -Test mit T -Test bezeichnet, es handelt sich trotz verschiedener Schreibweise um denselben Test.

Die wichtigste Voraussetzung für die Durchführung des t -Tests ist die Annahme über die **Normalverteilung** der Variablen in der Grundgesamtheit. Da wir in der Regel die wahre Verteilung der Variablen in der Grundgesamtheit nicht kennen, ist es erforderlich, mit Hilfe der Verteilung der Variablen in der Stichprobe Rückschlüsse auf ihre Verteilung in der Grundgesamtheit zu ziehen. Diese Abschätzung kann z. B. dadurch erfolgen, dass per Augenschein geprüft wird, ob sich Median und Mittelwert der Variablen ungefähr entsprechen. Für die Überprüfung des Vorliegens einer Normalverteilung können auch statistische Tests genutzt werden, auf die im Exkurs näher eingegangen wird. Der Exkurs zeigt auch Möglichkeiten des statistischen Testens auf, wenn nicht davon ausgegangen werden kann, dass die zu analysierende Variable in der Grundgesamtheit normalverteilt ist.

Voraussetzung für t -Test

7.1 Der Einstichproben t -Test

Wie bereits aus dem Namen des Einstichproben t -Tests erkennbar ist, soll genau eine Stichprobe analysiert werden. Dabei wird gegen einen festen Wert (Testwert) getestet, ob der Mittelwert der Stichprobe von diesem Wert signifikant verschieden ist. Die allgemeine Berechnung der Teststatistik des Einstichproben t -Testes kann im Formelverzeichnis nachgelesen werden (siehe dort 3.1).

Als Beispiel für die Durchführung des Einstichproben t -Tests soll folgende Frage untersucht werden: Der Therapieplan für die Reha sieht vor, dass die Patienten ca. eine Stunde Sport pro Tag machen. Haben die Patienten in der Studie während der Reha täglich 60 Minuten Sport getrieben?

Untersuchungsfrage

Im Datensatz *Zentren_gesamt.sav* ist in der Variablen *v15_sport* die Sportdauer pro Tag in Minuten für jeden Patienten angegeben.

Beispieldatensatz

Die Hypothesen lassen sich wie folgt definieren:

H_0 : Die Patienten und Patientinnen treiben während der Reha pro Tag im Durchschnitt 60 Minuten Sport.

H_1 : Die Patienten und Patientinnen treiben während der Reha pro Tag im Durchschnitt entweder mehr als 60 Minuten oder weniger als 60 Minuten Sport.

Nimmt man für den theoretischen Mittelwert für die Sportdauer pro Tag μ_{sport} an, lassen sich die Hypothesen folgendermaßen formalisieren:

$H_0: \mu_{\text{sport}} = 60 \text{ Minuten}$

$H_1: \mu_{\text{sport}} \neq 60 \text{ Minuten}$

Formulieren des Hypothesenpaars

Für die Überprüfung der Nullhypothese wird in SPSS im Menü die folgende Funktion gewählt: *Analysieren* → *Mittelwerte vergleichen* → *t-Test bei einer Stichprobe*.

Auswahl der Variablen

Als *Testvariable* wird in der Dialogbox *v15_sport* und als *Testwert* 60 gewählt, da auf einen Unterschied zu 60 Minuten getestet werden soll (siehe **Abb. 7.1**).

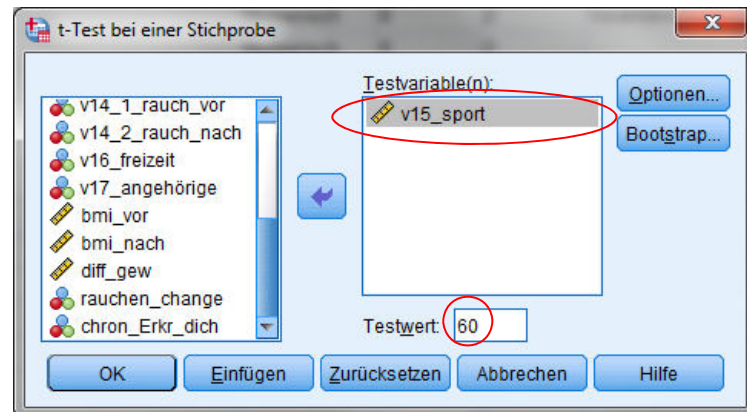


Abb. 7.1: Auswahl der Variablen beim Einstichproben t-Test

Mit Klick auf *OK* folgt das Ergebnisfenster im Viewer (siehe **Abb. 7.2**).

Statistik bei einer Stichprobe

	N	Mittelwert	Std.-Abweichung	Standardfehler des Mittelwertes
Sportliche Betätigung	90	50,14	11,947	1,259

Test bei einer Stichprobe

	Testwert = 60					
	T	df	Sig. (2-seitig)	Mittlere Differenz	95% Konfidenzintervall der Differenz	
					Untere	Obere
Sportliche Betätigung	-7,826	89	,000	-9,856	-12,36	-7,35

Abb. 7.2: Ergebnisse des t-Tests bei einer Stichprobe

Interpretation der Ergebnisse

Das Ergebnisfenster enthält zwei Tabellen. In der oberen Tabelle („Statistik bei einer Stichprobe“) sind die deskriptiven Werte der Variablen *v15_sport* aufgelistet. Man kann bereits hier erkennen, dass der Mittelwert der Variablen der Stichprobe mit 50,14 unter dem Testwert von 60 liegt. Ein konfirmatorischer Beweis ist dies jedoch noch nicht, dafür muss das Ergebnis des statistischen Tests herangezogen bzw. ein 95%-Konfidenzintervall berechnet werden.

Die Tabelle „Test bei einer Stichprobe“ zeigt die Ergebnisse der Teststatistik:

- Teststatistik (T)
- Freiheitsgrade (df = *degrees of freedom*)
- p-Wert (Sig. (2-seitig))
- Mittlere Differenz
- Konfidenzintervall

Für die Entscheidung, ob die Nullhypothese beibehalten werden kann, reicht es aus, den p-Wert zu betrachten (Testergebnis). Dieser ist hier signifikant, da er kleiner als 0,05 ist, was bedeutet, dass die Patienten und Patientinnen während der Reha entweder mehr oder weniger als 60 Minuten Sport am Tag im Mittel treiben. Die

Nullhypothese wird also verworfen. Mit dem p-Wert lässt sich eine Aussage darüber treffen, dass ein signifikanter Unterschied zu 60 Minuten vorliegt.

Für die Interpretation, ob die Patienten nun mehr oder weniger als 60 Minuten Sport im Mittel am Tag getrieben haben, ist die mittlere Differenz interessant (siehe die Tabelle „Test bei einer Stichprobe“ in **Abb. 7.2**). Die mittlere Differenz berechnet sich aus dem Mittelwert der Variablen *v15_sport* (50,14) und dem Testwert (60). Hierfür wird zusätzlich ein 95 %-Konfidenzintervall angegeben (allgemeine Berechnung siehe Formelverzeichnis 3.2 und 3.3). Wenn das Konfidenzintervall die Null enthält, kann die Nullhypothese nicht abgelehnt werden, da die Differenz aus Stichprobenmittelwert und Testwert mit einer Wahrscheinlichkeit von 95 % auch Null sein kann und damit eine mittlere Sportdauer von 60 Minuten pro Tag möglich wäre. Die mittlere Differenz mit seinem Konfidenzintervall liegt im negativen Bereich, d. h., die Patienten und Patientinnen in dieser Stichprobe haben mit einer Sicherheit von 95 % im Durchschnitt weniger als 60 Minuten Sport pro Tag getrieben.

Für die Entscheidung, ob ein signifikanter Unterschied zu 60 Minuten Sport pro Tag im Durchschnitt vorliegt, kann also der p-Wert betrachtet werden (Entscheidung der Signifikanz). Um eine Aussage darüber zu treffen, in welche Richtung der Unterschied zu 60 Minuten Sport pro Tag vorliegt, muss das Konfidenzintervall für die mittlere Differenz betrachtet werden.

Berechnung des Konfidenzintervalls für einen Mittelwert

Das Konfidenzintervall wird beim *t*-Test standardmäßig mit ausgegeben. Im Fall, dass man für einen Mittelwert an dem zugehörigen Konfidenzintervall interessiert ist, muss die folgende Option gewählt werden: *Analysieren* → *Deskriptive Statistiken* → *Explorative Datenanalyse*.

Als Variable wird *v15_sport* in das Kästchen *Abhängige Variablen* gezogen. Da man nur an dem Mittelwert mit dem Konfidenzintervall interessiert ist, reicht es aus, bei *Statistiken* die Voreinstellung *Deskriptive Statistik* zu belassen (siehe **Abb. 7.3**).

Als Ausgabe erscheint die in **Abbildung 7.4** dargestellte Tabelle. Im Durchschnitt haben die Patienten und Patientinnen in dieser Stichprobe pro Tag 50,14 Minuten Sport getrieben. Das 95 %-Konfidenzintervall reicht von 47,64 bis 52,65. Man kann also sehen, dass das Konfidenzintervall mit der oberen Grenze unterhalb der 60 Minuten (Testwert) liegt, weshalb man auch hiermit auf einen signifikanten Unterschied zu 60 Minuten schließen kann.

Exkurs

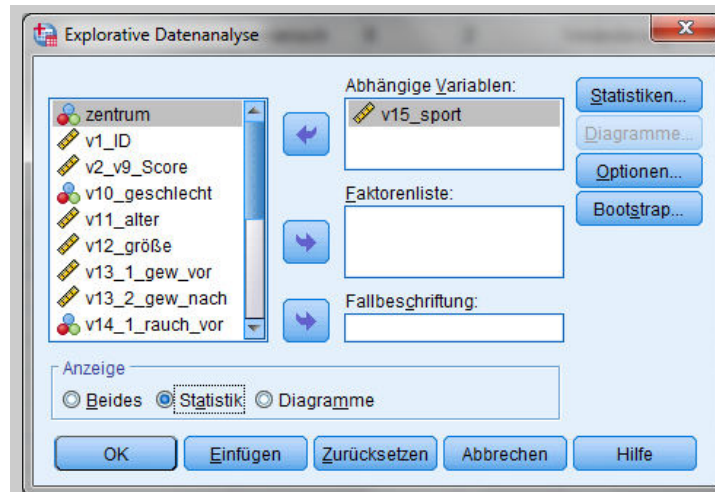


Abb. 7.3: Berechnung des Konfidenzintervalls für einen Stichprobenmittelwert

Deskriptive Statistik			
		Statistik	Std.-Fehler
Sportliche Betätigung	Mittelwert	50,14	1,259
	95% Konfidenzintervall des Mittelwerts		
	Untergrenze	47,64	
	Obergrenze	52,65	
	5% getrimmtes Mittel	49,71	
	Median	50,00	
	Varianz	142,732	
	Std.-Abweichung	11,947	
	Minimum	24	
	Maximum	98	
	Spannweite	74	
	Interquartilbereich	15	
	Schiefte	,756	,254
	Kurtosis	2,102	,503

Abb. 7.4: Ergebnisfenster für die Variable v15_sport

7.2 *t*-Test für zwei unabhängige Stichproben

Anwendung

Der *t*-Test für zwei unabhängige Stichproben wird dafür genutzt, zwei Stichproben, die unabhängig voneinander erhoben wurden, hinsichtlich eines stetigen Merkmals miteinander zu vergleichen. Zwei Stichproben sind unabhängig, wenn ein Merkmal an zwei Gruppen erhoben wurde, die getrennt voneinander betrachtet werden können. Ein typisches Beispiel für unabhängige Stichproben sind Männer und Frauen.

Im Gegensatz dazu werden bei abhängigen Stichproben Merkmale mehrmals an ein und derselben Population erhoben. Beispielhaft ist hier die wiederholte Messung eines Merkmals an einer Population.

Die allgemeine Berechnung der Teststatistik für den Zwei-Stichproben *t*-Test befindet sich im Formelverzeichnis (siehe dort 3.4).

Als Beispiel für die Durchführung dieses Tests soll der durchschnittliche Zufriedenheitsscore zwischen Reha-Zentrum 1 und Reha-Zentrum 2 verglichen werden, was der Forschungsfrage 1 aus Kapitel 1 entspricht. Dieser Summenscore kann Werte zwischen 8 und 32 annehmen und wird als stetiges Merkmal betrachtet.

Das Hypothesenpaar bezüglich des Zufriedenheitsscores (Mittelwertvergleich) zwischen Reha-Zentrum 1 und Reha-Zentrum 2 kann wie folgt aufgestellt werden:

H_0 : Es gibt keinen Unterschied in der mittleren Zufriedenheit der Patientinnen und Patienten (Zufriedenheitsscore) zwischen Zentrum 1 und Zentrum 2.

H_1 : Es gibt einen Unterschied in der mittleren Zufriedenheit der Patientinnen und Patienten (Zufriedenheitsscore) zwischen Zentrum 1 und Zentrum 2.

Nimmt man für den theoretischen Mittelwert für die Patientenzufriedenheit in Zentrum 1 μ_1 und für die Patientenzufriedenheit in Zentrum 2 μ_2 an, lassen sich die Hypothesen folgendermaßen formalisieren:

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Beispiel

Der Datensatz *Zentren_gesamt.sav* enthält die Variable *v2_v9_Score*, die den aufsummierten Zufriedenheitsscore aller Patienten und Patientinnen darstellt. Außerdem gibt es die Variable *zentrum*, in der die Reha-Zentren mit 1, 2 und 3 kodiert sind.

Für die Durchführung des *t*-Tests für zwei unabhängige Stichproben benötigt man eine Merkmalsvariable und eine Variable, die die Gruppenzugehörigkeit widerspiegelt. Die Überprüfung der Nullhypothese erfolgt in SPSS mit der Auswahl der folgenden Funktion: *Analysieren* → *Mittelwerte vergleichen* → *T-Test bei unabhängigen Stichproben*. In der Dialogbox wird als *Testvariable(n)* *v2_v9_Score* und als *Gruppierungsvariable* *zentrum* gewählt (siehe Abb. 7.5). Anschließend muss durch Klick auf *Gruppen def. ...* angegeben werden, welche Gruppen miteinander verglichen werden sollen. In diesem Fall soll Gruppe 1 das Reha-Zentrum 1 und Gruppe 2 soll das Reha-Zentrum 2 sein.

Auswahl der Variablen

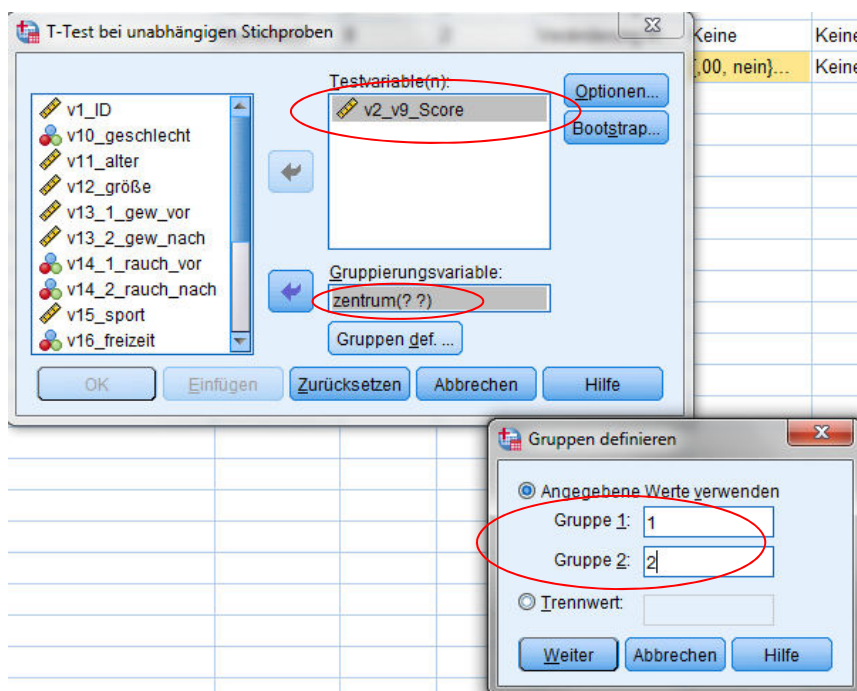


Abb. 7.5: Auswahl der Variablen für den *t*-Test bei unabhängigen Stichproben für die Variable „Zufriedenheitsscore“ (*v2_v9_Score*)

Mit Klick auf *Weiter* und *OK* erhält man das Ausgabefenster mit den Ergebnissen (siehe **Abb. 7.6**).

Gruppenstatistiken										
Nummer Zentrum		N	Mittelwert	Std.-Abweichung	Standardfehler des Mittelwertes					
ZUF 8 Score	1	30	16,87	4,652	,849					
	2	30	22,47	4,696	,857					

Test bei unabhängigen Stichproben										
		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	95% Konfidenzintervall der Differenz	
ZUF 8 Score	Varianzen sind gleich	,002	,964	-4,641	58	,000	-5,600	1,207	-8,016	-3,184
	Varianzen sind nicht gleich			-4,641	57,995	,000	-5,600	1,207	-8,016	-3,184

Abb. 7.6: Ergebnisfenster für den t-Test bei unabhängigen Stichproben für die Variable „Zufriedenheitsscore“ (v2_v9_Score)

Das Ausgabefenster besteht aus zwei Teilen, wobei der obere Teil („Gruppenstatistiken“) deskriptive Kenngrößen der Variable v2_v9_Score für die beiden Gruppen ausgibt. Für beide Zentren sind dies jeweils Mittelwert, Standardabweichung und Standardfehler des Mittelwertes.

Interpretation der Ergebnisse

Im unteren Teil („Test bei unabhängigen Stichproben“) befindet sich zunächst das Testresultat des **Levene-Tests der Varianzgleichheit**, der standardmäßig vor der Durchführung des t-Tests für zwei unverbundene, d. h. unabhängige, Stichproben in SPSS durchgeführt wird. Dieser Test dient dazu, zu testen, ob die Varianzen in beiden Gruppen gleich sind. Diese Annahme wird beim nachfolgenden Testen auf Unterschiede zwischen den Gruppen meist stillschweigend vorausgesetzt.

Für die Nullhypothese, dass die Varianzen gleich sind, wird eine Teststatistik (F) sowie der zugehörige p-Wert berechnet. Der p-Wert beträgt 0,964. Damit kann davon ausgegangen werden, dass die Varianzen in beiden Stichproben gleich sind. Die Ergebnisse des t-Tests können deshalb aus der oberen der beiden Zeilen abgelesen werden. Wenn der Levene-Test ergeben hätte, dass die Varianzen nicht gleich sind, hätten die Werte aus der unteren Zeile abgelesen werden müssen.

In der Ausgabe „T-Test für die Mittelwertgleichheit“ wird der Wert der Teststatistik (T=-4,641), die Freiheitsgrade (df=58) und die Signifikanz (p-Wert=0,000) angegeben. Der p-Wert ist signifikant, da er kleiner als das Signifikanzniveau von 5 % ist.

Die mittlere Differenz zwischen Zentrum 1 und Zentrum 2 beim Zufriedenheitsscore beträgt -5,6 (Zufriedenheitsscore im Zentrum 1 minus Zufriedenheitsscore im Zentrum 2), d. h. im Reha-Zentrum 2 ist der Score im Mittel um 5 Punkte höher. Das zugehörige 95 %-Konfidenzintervall für diese Differenz (allgemeine Berechnung siehe Formelverzeichnis 3.4 und 3.5) reicht von -8,016 bis -3,184 und liegt damit also vollständig im negativen Bereich.

Schlussfolgerung

Anhand des p-Werts kann demzufolge geschlussfolgert werden, dass die Nullhypothese, wonach sich die Mittelwerte der Patientenzufriedenheit mit der Reha zwischen den Reha-Zentren 1 und 2 nicht unterscheiden, zu verwerfen ist. Die Patienten und

Patientinnen in den beiden Zentren sind vielmehr unterschiedlich zufrieden mit der Reha.

Die mittlere Differenz weist einen negativen Wert auf. Daraus kann man schließen, dass Zentrum 2 im Durchschnitt höhere Werte im Zufriedenheitsscore aufwies als Zentrum 1. Das 95 %-Konfidenzintervall enthält nicht die Null, was erneut das signifikante Ergebnis unterstreicht. Die Patientinnen und Patienten im Zentrum 2 waren also deutlich zufriedener mit der Reha als die Patientinnen und Patienten in Zentrum 1.

ANOVA

Für den **Vergleich von mehr als 2 unabhängigen Gruppen** ist in SPSS die ANOVA (kurz für: *Analysis of Variance*) implementiert (siehe **Abb. 7.7**). Mit Hilfe von *Analysieren* → *Mittelwerte vergleichen* → *Einfaktorielle ANOVA* können die Mittelwerte von mehreren unabhängigen Stichproben miteinander verglichen werden.

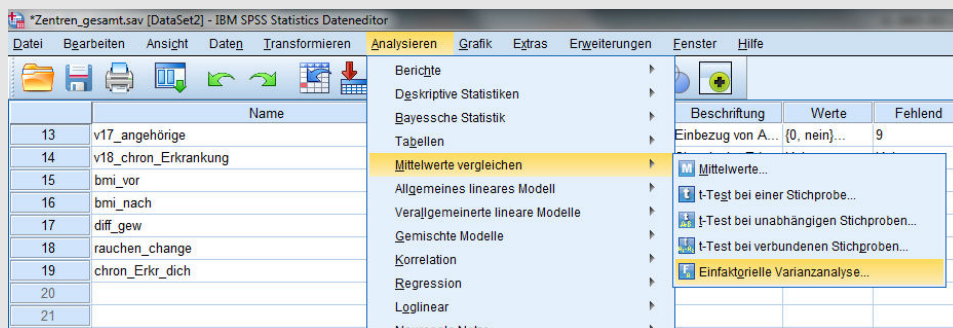


Abb. 7.7: Auswahl der ANOVA in SPSS

Mit Hilfe dieser Funktion lässt sich ein globaler Vergleich zwischen allen Gruppen durchführen. Das bedeutet, die *Einfaktorielle ANOVA* testet, ob ein Unterschied zwischen den Gruppen vorliegt. Um herauszufinden, zwischen welchen Gruppen genau ein Unterschied vorhanden ist, kann über das Feld *Post-Hoc-Mehrfachvergleiche* die Adjustierung für den Fehler 1. Art (Multiplizitätsproblem – siehe Studienbrief 4 des Moduls „Empirische Methoden I“) festgelegt werden, woraufhin alle paarweisen Vergleiche erscheinen.

Mit einer ANOVA kann z. B. die Frage untersucht werden, ob sich die drei Zentren hinsichtlich des Zufriedenheitsscores unterscheiden. Die Nullhypothese lautet dann: Es gibt keinen Unterschied beim Zufriedenheitsscore zwischen den drei Gruppen ($H_0: \mu_1 = \mu_2 = \mu_3$).

Das Ergebnisfenster enthält die in **Abbildung 7.8** dargestellten Ergebnisse und zeigt im oberen Bereich („Einfaktorielle ANOVA“) den globalen Vergleich zwischen den drei Gruppen und im unteren Bereich („Mehrfachvergleiche“) die paarweisen Vergleiche, wobei als α -Korrektur *Bonferroni* verwendet wurde.

Der globale Vergleich zeigt mit $p=0,000$ einen signifikanten Unterschied zwischen den drei Zentren hinsichtlich des Zufriedenheitsscores. Das bedeutet, die Patienten waren in den drei Zentren verschieden zufrieden mit ihrer Behandlung. Die Ergebnisse des Post-Hoc-Tests zeigen, dass sich Zentrum 1 und Zentrum 2 stark signifikant ($p = 0,000$) voneinander unterscheiden, Zentrum 2 und Zentrum 3 unterscheiden sich ebenfalls signifikant ($p = 0,007$). Zentrum 1 und Zentrum 3 unterscheiden sich nicht voneinander ($p = 0,105$).

Exkurs

Einfaktorielle ANOVA

ZUF 8 Score

	Quadrat-summe	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	476,089	2	238,044	14,162	,000
Innerhalb der Gruppen	1462,400	87	16,809		
Gesamt	1938,489	89			

Mehrfachvergleiche

Abhängige Variable: ZUF 8 Score

Bonferroni

(I) Nummer Zentrum	(J) Nummer Zentrum	Mittlere Differenz (I-J)	Std.-Fehler	Signifikanz	95%-Konfidenzintervall	
					Untergrenze	Obergrenze
1	2	-5,600*	1,059	,000	-8,18	-3,02
	3	-2,267	1,059	,105	-4,85	,32
2	1	5,600*	1,059	,000	3,02	8,18
	3	3,333*	1,059	,007	,75	5,92
3	1	2,267	1,059	,105	-,32	4,85
	2	-3,333*	1,059	,007	-5,92	-,75

*. Die Differenz der Mittelwerte ist auf dem Niveau 0.05 signifikant.

Abb. 7.8: Ergebnisse der ANOVA für den Vergleich des Zufriedenheitsscore über alle Reha-Zentren

7.3 *t*-Test für zwei abhängige Stichproben

Anwendung

Der *t*-Test für zwei abhängige oder verbundene Stichproben wird dann durchgeführt, wenn für jedes Subjekt (z. B. Patient, Proband) zwei Messungen vorliegen. Diese zwei Messungen können daher resultieren, weil sie zu zwei verschiedenen Zeitpunkten erhoben wurden (vorher-nachher-Vergleich) oder an einem Patienten zwei Messungen (z. B. an verschiedenen Körperregionen) miteinander verglichen werden sollen. Bei dem *t*-Test für zwei abhängige Stichproben gibt es zwei Stichproben, die in SPSS ausgewählt werden müssen, um diese miteinander vergleichen zu können (im Gegensatz zu dem unabhängigen Fall, bei dem voneinander unabhängige Gruppen miteinander verglichen werden).

Als Beispiel für die Durchführung soll die Frage untersucht werden, ob eine Veränderung des Körpergewichts bei Patienten und Patientinnen mit Adipositas und Typ-II-Diabetes nach einem 4-wöchigen Reha-Aufenthalt zu beobachten ist (Vorher-Nachher-Vergleich), was der Forschungsfrage 3 aus Kapitel 1 entspricht. Im Datensatz *Zentren_gesamt.sav* ist in den Variablen *v13_1_gew_vor* und *v13_2_gew_nach* das Körpergewicht vor und nach der Reha angegeben. Die Berechnung der Teststatistik basiert auf den Differenzen zwischen den Angaben der Patient(inn)en zu ihrem Gewicht vor und nach der Reha. Dabei erfolgt die Differenzbildung auf Patientenebene folgendermaßen: *v13_1_gew_vor* - *v13_2_gew_nach*. Das heißt, auf Patientenebene erhält man Werte, die einem Auskunft darüber geben, ob ein Patient während der Reha zu- oder abgenommen hat oder ob sich sein Gewicht nicht verändert hat. Wie die spezifischen Werte(bereiche) zu interpretieren sind, wird im Folgenden erläutert:

Ist *v13_1_gew_vor* > *v13_2_gew_nach* (Differenz > 0) hat der Patient während des Reha-Aufenthalts abgenommen.

Ist *v13_1_gew_vor* < *v13_2_gew_nach* (Differenz < 0) hat der Patient während des Reha-Aufenthalts zugenommen.

Ist *v13_1_gew_vor* = *v13_2_gew_nach* (Differenz = 0) hat der Patient während des Reha-Aufenthalts sein Gewicht nicht verändert.

Beispiel

Das Hypothesenpaar für die Forschungsfrage lässt sich wie folgt formulieren:

H_0 : Es gibt keinen Unterschied im mittleren Gewicht vor und nach der Reha.

H_1 : Es gibt einen Unterschied im mittleren Gewicht vor und nach der Reha.

Wie bereits erörtert wurde, wird im Rahmen der Teststatistik zunächst die Differenz zwischen den beiden Variablen *v13_1_gew_vor* - *v13_2_gew_nach* auf Patientenebene gebildet. Basierend auf den auf Individualebene gebildeten Differenzen wird dann der Mittelwert gebildet, der gleichzeitig den primären Effektschätzer des Gruppenvergleichs darstellt. Die Hypothesen zum Gruppenvergleich lassen sich entsprechend so formalisieren, dass μ_{diff} die gemittelte Differenz in der Grundgesamtheit darstellt. H_0 beschreibt, dass die Differenz im mittleren Gewicht vor und nach der Reha 0 beträgt. H_1 beschreibt, dass es einen Unterschied im mittleren Gewicht gibt. Die Hypothesen lassen sich folgendermaßen formalisieren:

$H_0: \mu_{\text{diff}} = 0$

$H_1: \mu_{\text{diff}} \neq 0$

Formulieren des Hypothesenpaars

Für die Überprüfung der Nullhypothese wird in SPSS im Menü die folgende Funktion gewählt: *Analysieren* → *Mittelwerte vergleichen* → *T-Test bei verbundenen Stichproben*. In der Dialogbox *T-Test bei gepaarten Stichproben* wird als Paar 1 die Variable 1 *v13_1_gew_vor* und für Variable 2 *v13_2_gew_nach* gewählt (siehe Abb. 7.9).

Auswahl der Variablen

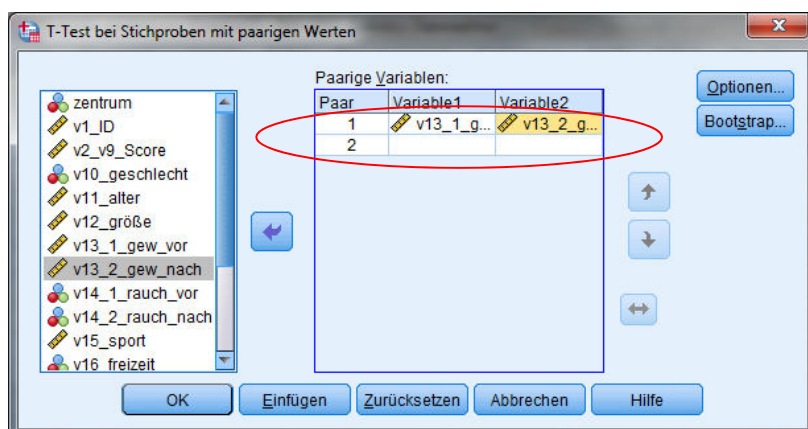


Abb. 7.9: Auswahl der Variablen für den verbundenen t-Test (für abhängige Stichproben)

Mit Klick auf *OK* folgt im Viewer die Ausgabe des Ergebnisses des *t*-Tests für zwei abhängige Stichproben (siehe *Abb. 7.10*).

Statistik bei gepaarten Stichproben

		Mittelwert	N	Std.-Abweichung	Standardfehler des Mittelwertes
Paaren 1	Körpergewicht vor Reha	86,39	90	11,077	1,168
	Körpergewicht nach Reha	81,79	90	10,341	1,090

Korrelationen bei gepaarten Stichproben

		N	Korrelation	Signifikanz
Paaren 1	Körpergewicht vor Reha & Körpergewicht nach Reha	90	,974	,000

Test bei gepaarten Stichproben

		Gepaarte Differenzen				T	df	Sig. (2-seitig)
		Mittelwert	Std.-Abweichung	Standardfehler des Mittelwertes	95% Konfidenzintervall der Differenz Untere Obere			
Paaren 1	Körpergewicht vor Reha - Körpergewicht nach Reha	4,600	2,565	,270	4,063 5,137	17,013	89	,000

Abb. 7.10: Ergebnisse des verbundenen t-Tests (für abhängige Stichproben) für die Entwicklung des Körpergewichts während der Reha (v13_1_gew_vor, v13_2_gew_nach)

Interpretation der Ergebnisse

Das Ausgabefenster im Viewer besteht aus drei Teilen:

- In der oberen Tabelle („Statistik bei gepaarten Stichproben“) sind deskriptive Kenngrößen für beide Stichproben berechnet worden. Diese umfassen jeweils den Mittelwert, die Fallzahl (N), die Standardabweichung und den Standardfehler des Mittelwertes für beide Stichproben.
- In der mittleren Tabelle („Korrelationen bei gepaarten Stichproben“) ist die Korrelation zwischen den beiden Variablen bestimmt worden. Das Thema Korrelation wird im folgenden Kapitel behandelt, daher wird auf die Interpretation des Ergebnisses hier nicht näher eingegangen.
- Die untere Tabelle („Test bei gepaarten Stichproben“) enthält die wichtigen Ergebnisse des verbundenen *t*-Tests. Für die auf Individualebene gebildeten Differenzen (v13_1_gew_vor - v13_2_gew_nach) ist der Mittelwert mit Standardabweichung und Standardfehler angegeben. Außerdem ist das 95%-Konfidenzintervall für die Differenz genannt, das von 4,063 bis 5,137 reicht. Aufgrund des Konfidenzintervalls kann die Entscheidung getroffen werden, dass mit 95 % Wahrscheinlichkeit die Differenz des Gewichts im positiven Bereich liegt und damit im Mittel die Patient(inn)en während des Reha-Aufenthalts abgenommen haben. Die Null ist im Konfidenzintervall nicht enthalten, weshalb die Nullhypothese verworfen werden kann. Es folgt die Angabe der Teststatistik ($T = 17,013$), der Freiheitsgrade (df) und des *p*-Werts (Sig. (2-seitig)). Der *p*-Wert ist kleiner als 0,05 und damit signifikant. Die Nullhypothese kann verworfen werden und es lässt sich schlussfolgern, dass ein signifikanter Unterschied zwischen dem Gewicht vor und nach der Reha vorliegt.

Übungsaufgabe

- 7.1) Es soll überprüft werden, ob sich Männer und Frauen hinsichtlich des Zufriedenheitsscores im Mittel unterscheiden.

Führen Sie dazu einen 2-Stichproben t -Test durch und interpretieren Sie das Ergebnis für eine vorher definierte Nullhypothese.

8 Korrelationsmaße zur Analyse von Zusammenhängen

In den vorherigen Kapiteln wurde der Unterschied zwischen zwei Variablen mit Schätz- und Testverfahren untersucht. In diesem Kapitel soll der Zusammenhang zwischen zwei Variablen analysiert werden.

8.1 Korrelationskoeffizient

Anwendung Ein typisches Beispiel für die Untersuchung des Zusammenhangs zwischen zwei Variablen stellen die beiden Variablen Körpergröße und Gewicht dar. Es lässt sich leicht vorstellen, dass wenn Personen größer sind, sie auch ein höheres Gewicht aufweisen. Genau dieser **Zusammenhang**, wenn eine Variable größere Werte annimmt, dann gilt dies auch für die andere Variable, ist eine Beschreibung für den Zusammenhang zwischen zwei Variablen. Für die Beschreibung dieses Zusammenhangs berechnet man den Korrelationskoeffizienten **zwischen zwei Variablen**. Die allgemeine Berechnung des Korrelationskoeffizienten ist im Formelverzeichnis zu finden (dort 4.1).

Definition

Der **Korrelationskoeffizient** ist eine Maßzahl, die darüber Auskunft gibt, ob ein linearer Zusammenhang zwischen zwei Variablen besteht und berechnet sich über die Abweichung der Beobachtungen vom Mittelwert der beiden Stichproben. Er nimmt Werte zwischen -1 und 1 an.

Wenn der berechnete Wert 0 ergibt, existiert kein linearer Zusammenhang zwischen den beiden Variablen.

Bei -1 sagt man, die beiden Variablen sind negativ linear abhängig.

Bei 1 sind beide Variablen positiv linear abhängig.

Abhängig davon, ob ein linearer Zusammenhang zwischen zwei Variablen angenommen werden kann, berechnet man entweder den **Korrelationskoeffizienten nach Pearson** oder einen Rangkorrelationskoeffizienten wie z. B. **Spearman's Korrelationskoeffizienten**.

Beispieldatensatz Die Vorgehensweise in SPSS wird in den folgenden Abschnitten anhand des Datensatzes *Zentren_gesamt.sav* beschrieben.

Vor der Wahl, welcher Korrelationskoeffizient berechnet werden soll, ist es ratsam den grafischen Zusammenhang zwischen den beiden Variablen zu verdeutlichen.

Mit Hilfe eines Streudiagramms lassen sich beide Variablen gegeneinander abtragen. Wie in **Abbildung 8.1** dargestellt, lässt sich ein nahezu perfekter linearer Zusammenhang zwischen dem Gewicht vor Reha-Beginn (*v13_1_gew_vor*) und der Körpergröße (*v12_größe*) grafisch vermuten: Die Beobachtungen liegen nahezu auf einer Geraden. Das heißt, je größer die Patienten und Patientinnen sind, desto höher war auch ihr Körpergewicht vor Beginn der Reha.

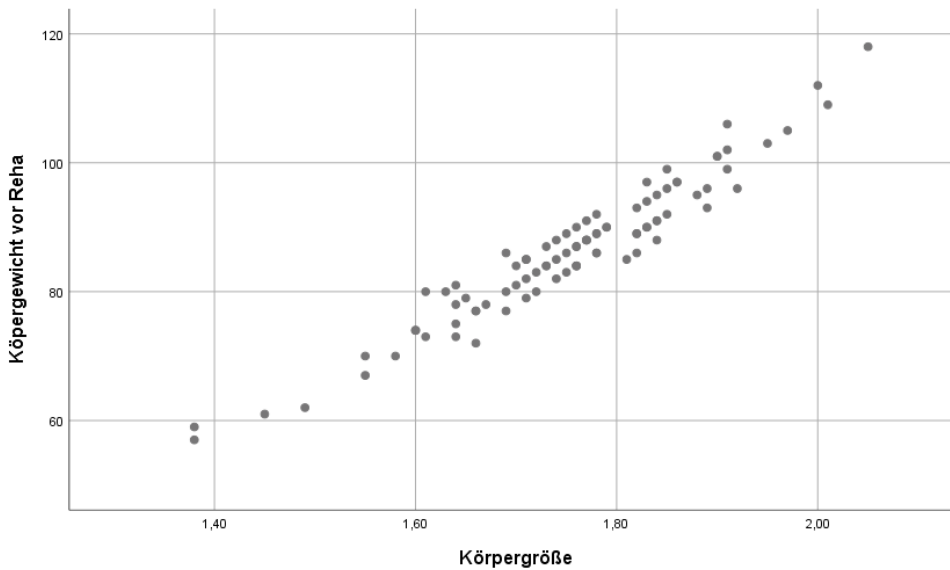


Abb. 8.1: Streudiagramm der Variablen Gewicht vor der Reha (v13_1_gew_vor) und Körpergröße (v12_größe)

Die Anleitung für den rechnerischen Beleg für den linearen Zusammenhang von zwei Variablen mit Hilfe von SPSS wird im nachfolgenden Abschnitt erklärt.

8.2 Korrelationskoeffizient nach Pearson

Wie schon im vorherigen Abschnitt angedeutet, wird der Korrelationskoeffizient nach Pearson bei Variablen berechnet, bei denen man einen linearen Zusammenhang vermutet. Dazu konnte in dem o. g. Beispiel (Zusammenhang zwischen Körpergröße und Gewicht vor der Reha), grafisch gezeigt werden, dass dieser Zusammenhang existiert.

Eine weitere Voraussetzung für die Berechnung des Korrelationskoeffizienten nach Pearson ist, dass beide Variablen stetig oder intervallskaliert sind (vgl. zu den Voraussetzungen Schendera 2004: 491; Bortz 2005: 224).

Die Hypothesen lassen sich wie folgt definieren:

H_0 : Es gibt keinen Zusammenhang zwischen Körpergröße und Gewicht vor der Reha.

H_1 : Es gibt einen Zusammenhang zwischen Körpergröße und Gewicht vor der Reha.

Voraussetzungen

Formulieren des Hypothesenpaars

In SPSS lässt sich die Funktion zur Berechnung des Korrelationskoeffizienten wie in **Abbildung 8.2** dargestellt, aufrufen: *Analysieren* → *Korrelation* → *Bivariat*.

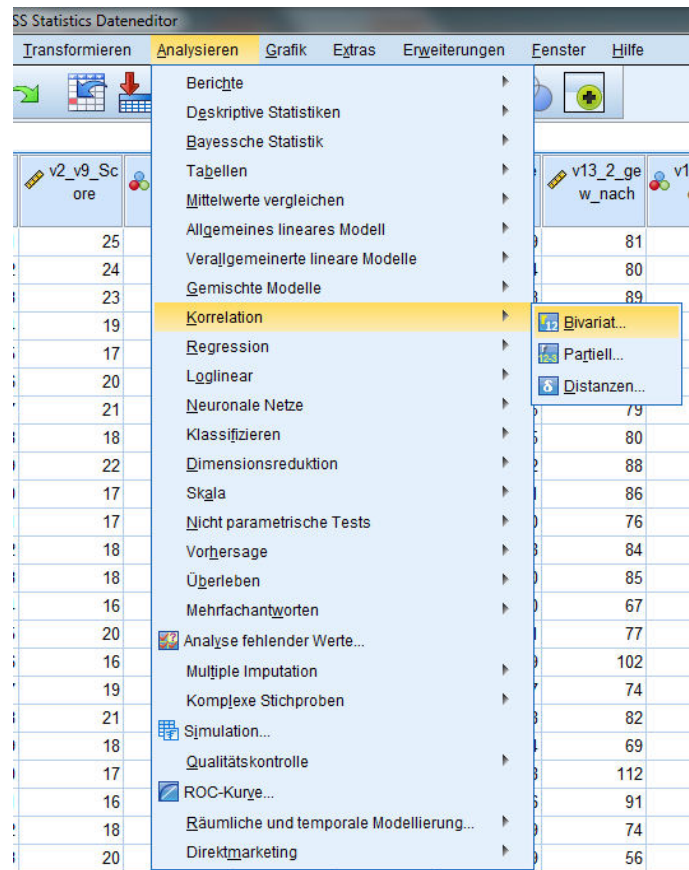


Abb. 8.2: Berechnung des Korrelationskoeffizienten nach Pearson

Auswahl der Variablen

Nachfolgend öffnet sich die Dialogbox *Bivariate Korrelationen* (siehe Abb. 8.3). In das Feld *Variablen* werden nun die beiden zu untersuchenden Variablen eingefügt. Außerdem wird ein Häkchen bei *Pearson* gemacht, damit nur der Korrelationskoeffizient nach Pearson ausgegeben wird. Über das Feld *Optionen* können deskriptive Werte berechnet werden. Wenn man ein Häkchen vor das Feld *Signifikante Korrelationen markieren* setzt, werden signifikante Ergebnisse in der Ausgabedatei mit Sternchen (*) gekennzeichnet.

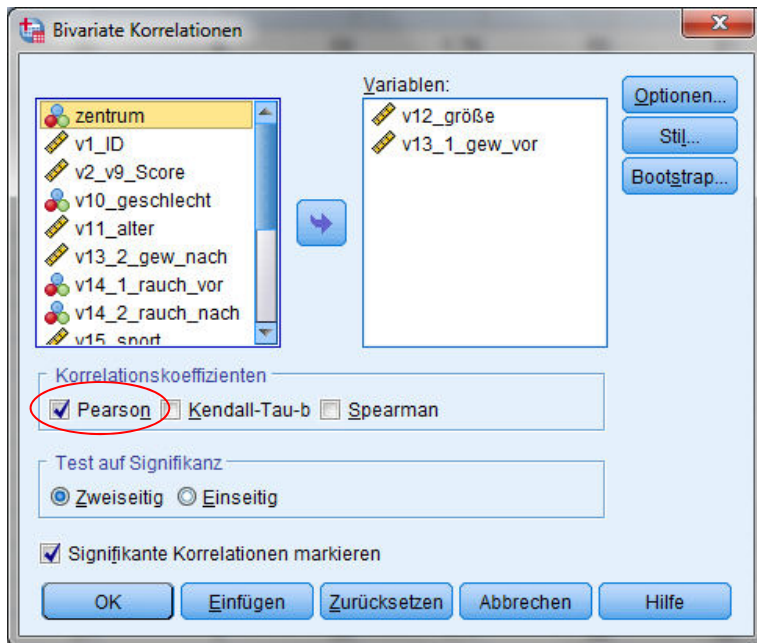


Abb. 8.3: Fenster zur Auswahl der Variablen und des Korrelationskoeffizienten

Nach Klick auf OK folgt das Ausgabefenster mit den Ergebnissen (siehe Abb. 8.4).

Korrelationen		Körpergröße	Körpergewicht vor Reha
Körpergröße	Korrelation nach Pearson	1	,965**
	Signifikanz (2-seitig)		,000
	N	90	90
Körpergewicht vor Reha	Korrelation nach Pearson	,965**	1
	Signifikanz (2-seitig)	,000	
	N	90	90

** . Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Abb. 8.4: Ergebnisfenster für die Korrelation zwischen der Körpergröße (v12_größe) und Gewicht vor der Reha (v13_1_gew_vor)

Das Ergebnisfenster für die berechneten Korrelationen ist in SPSS als Vierfeldertafel aufgebaut. Das bedeutet, man kann alle Korrelationen zwischen und innerhalb der beiden erhobenen Variablen Gewicht vor der Reha (v13_1_gew_vor) und Körpergröße (v12_größe) ablesen. Korrelationen innerhalb einer Variablen sind dabei zu vernachlässigen, da selbstverständlich ein perfekter Zusammenhang zwischen ein und derselben Variable besteht und damit auch ein Korrelationskoeffizient von 1 vorliegt.

Der Korrelationskoeffizient nach Pearson zwischen v12_größe und v13_1_gew_vor beträgt 0,965 und deutet auf einen starken positiv linearen Zusammenhang zwischen den beiden Variablen hin. Das bedeutet: Je größer der Patient ist, desto höher war auch sein Körpergewicht vor der Reha.

SPSS führt zusätzlich einen statistischen Test durch und kommt hier zu einem signifikanten Ergebnis des Zusammenhangs (Signifikanz: 0,000). Die formulierte Nullhypothese muss also verworfen werden. Es wird an dieser Stelle nicht weiter auf die

Interpretation der Ergebnisse

Durchführung des statistischen Tests eingegangen, da dieser für die Analyse des Zusammenhangs zweier Variablen nicht notwendig ist. Üblicherweise gibt man hierfür immer den Korrelationskoeffizienten an.

Im Eintrag links unten in der Vier-Felder-Tafel stehen dieselben Ergebnisse wie rechts oben. Wenn ein negativ linearer Zusammenhang vorliegen würde, würde der Korrelationskoeffizient mit entgegengesetztem Vorzeichen als rechts oben, aber mit demselben Wert angegeben werden.

Hinweis

Die Berechnung des **Rangkorrelationskoeffizienten nach Spearman** ist an dieser Stelle ebenfalls möglich. Ist der Zusammenhang zwischen zwei Variablen stark linear, so ist der Korrelationskoeffizient nach Spearman nahezu identisch mit dem Korrelationskoeffizienten nach Pearson (vgl. Fahrmeir et al. 2007: 145). In diesem Beispiel beträgt Spearmans Korrelationskoeffizient 0,954.

Die Umkehrung gilt allerdings nicht: Wenn sich der Korrelationskoeffizient nach Spearman berechnen lässt, kann nicht automatisch auch der nach Pearson berechnet werden.

8.3 Korrelationskoeffizient nach Spearman

Anwendung

Der Korrelationskoeffizient nach Spearman setzt im Gegensatz zu Pearson nicht den linearen Zusammenhang zwischen zwei Variablen voraus. Außerdem ist er geeignet, um Korrelationen bei ordinalen Variablen oder wenn eine Variable stetig und die andere ordinal ist, zu berechnen (vgl. zu den Voraussetzungen Schendera 2004: 497; Bortz 2005: 227).

Der Korrelationskoeffizient nach Spearman ist ein **Rangkorrelationskoeffizient**, dessen Berechnung auf Rängen basiert. Es gibt in SPSS noch einen weiteren Korrelationskoeffizienten, wenn zum Zusammenhang zwischen Variablen keine Annahme gemacht wird: Kendalls Tau-b.

An dieser Stelle soll beispielhaft der Zusammenhang zwischen zwei ordinal skalierten Variablen untersucht werden. Dazu wird der Korrelationskoeffizient nach Spearman für *v14_1_rauch_vor* und *v14_2_rauch_nach* berechnet (kodiert mit 0=Nichtraucher, 1=Gelegenheitsraucher und 2=Raucher). Mit diesen beiden Variablen soll untersucht werden, ob Patienten und Patientinnen, die bereits vor Beginn der Reha Raucher waren, dies auch nach der Reha noch sind, ob sich der Raucherstatus eventuell positiv verändert hat und Patienten weniger oder gar nicht mehr rauchen. Die Hypothesen lauten dementsprechend:

Formulieren des Hypothesenpaars

H_0 : Es gibt keinen Zusammenhang zwischen dem Raucherstatus vor und nach der Reha.

H_1 : Es gibt einen Zusammenhang zwischen dem Raucherstatus vor und nach der Reha.

Die Funktion zur Berechnung des Korrelationskoeffizienten lässt sich wie in **Abbildung 8.2** auswählen: *Analysieren* → *Korrelation* → *Bivariat*

Als *Variablen* werden die beiden besagten Merkmale ausgewählt (*v14_1_rauch_vor* und *v14_2_rauch_nach*). Bei den *Korrelationskoeffizienten* wird *Spearman* angeklickt (siehe *Abb. 8.5*).

Auswahl der Variablen

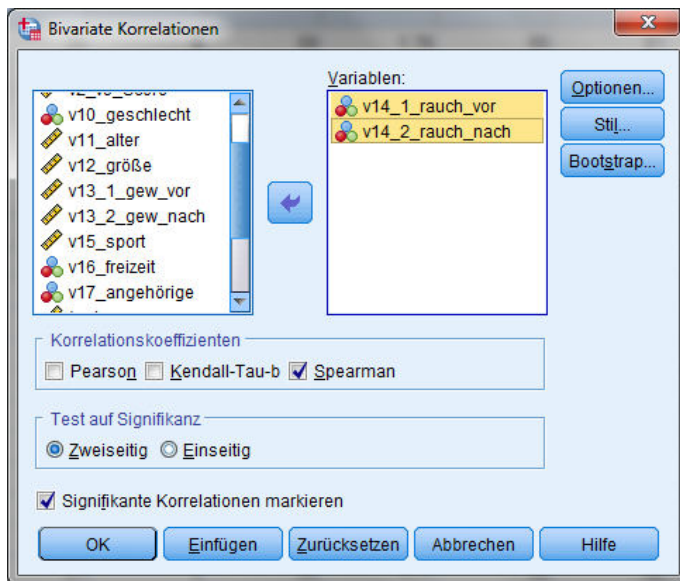


Abb. 8.5: Auswahl der Variablen und des Typs des Korrelationskoeffizienten

Mit *OK* folgt die Ergebnisausgabe (siehe *Abb. 8.6*). Das Ausgabefenster ähnelt der Ergebnisausgabe für den Korrelationskoeffizienten nach Pearson. Man sieht, dass der Korrelationskoeffizient zwischen dem Raucherstatus vor und nach der Reha 0,893 beträgt, was auf einen positiv linearen Zusammenhang zwischen den beiden Variablen schließen lässt. Das bedeutet, je höher der Statuswert vor der Reha war (Raucherstatus=2), desto höher ist der Statuswert nach der Reha, was ebenfalls einem Raucherstatus entspricht. Patienten und Patientinnen, die vor der Reha geraucht haben, tun dies also mit einer relativ hohen Korrelation auch nach der Reha. Dieser Test auf Zusammenhang zwischen den Variablen ist signifikant, was das Verwerfen der Nullhypothese bedeutet. Für eine Änderung des Raucherstatus hätte ein negativ linearer Zusammenhang vorliegen müssen.

Interpretation der Ergebnisse

Korrelationen			Raucherstatus vor Reha	Raucherstatus nach Reha
Spearman-Rho	Raucher- status vor Reha	Korrelationskoeffizient	1,000	,893**
		Sig. (2-seitig)	.	,000
		N	90	90
	Raucher- status nach Reha	Korrelationskoeffizient	,893**	1,000
		Sig. (2-seitig)	,000	.
		N	90	90

** . Die Korrelation ist auf dem 0,01 Niveau signifikant (zweiseitig).

Abb. 8.6: Ausgabefenster des Korrelationskoeffizienten nach Spearman für den Raucherstatus vor und nach der Reha

8.4 Bestimmtheitsmaß

Im Zusammenhang mit dem Korrelationskoeffizienten wird sehr häufig das Bestimmtheitsmaß mit ausgegeben. Dies liegt unter anderem daran, dass es einfach zu berechnen ist (Korrelationskoeffizient ins Quadrat gerechnet). Viel wichtiger ist jedoch die Interpretation, die dieses Maß zulässt.

Definition

Das **Bestimmtheitsmaß** ist ein Maß zur Erklärung, wie viel Varianz der einen Variablen sich durch eine andere Variable erklären lässt (vgl. Janssen, Laatz 2007: 427).

Die Berechnung des Bestimmtheitsmaßes lässt sich in SPSS über die Regression durchführen, deren Vorgehensweise in Studienbrief 4 ausführlich erklärt wird.

Beispielhaft soll das Bestimmtheitsmaß für die Variablen Körpergröße und Gewicht vor der Reha berechnet werden.

Zum Aufrufen der Funktion wird zunächst, wie in **Abbildung 8.7** dargestellt, die Regression aufgerufen: *Analysieren* → *Regression* → *Linear*.

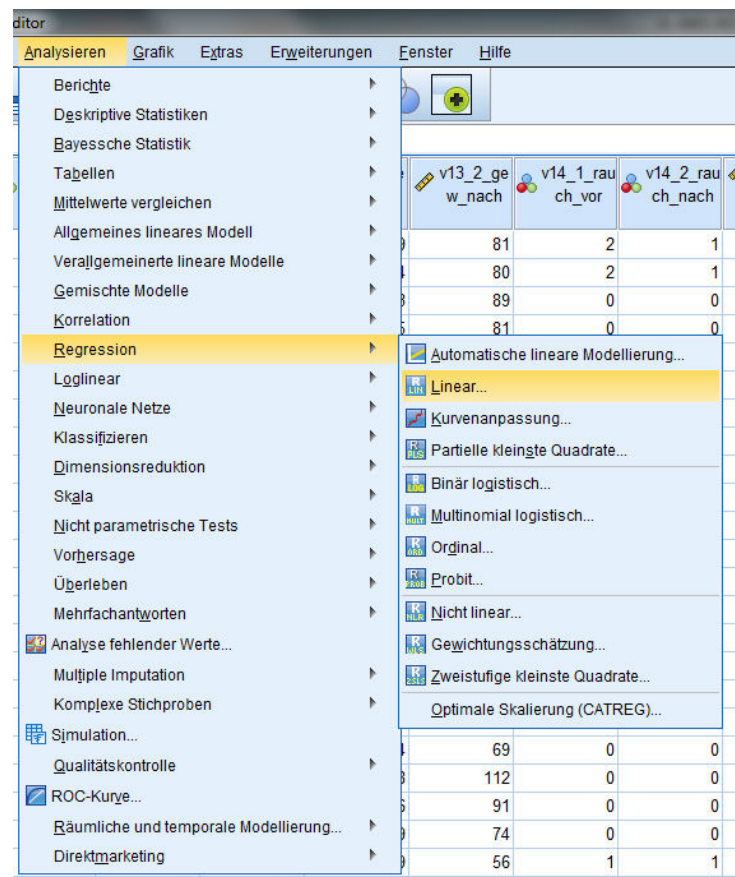


Abb. 8.7: Aufrufen der Funktion zur Berechnung des Bestimmtheitsmaßes

Auswahl der Variablen

Nach Anklicken der *Linearen Regression* erscheint das Auswahlfenster (siehe **Abb. 8.8**). Als *Abhängige Variable* wird *v13_1_gew_vor* und als *Unabhängige (Variable)* *v12_größe* ausgewählt. Für die Berechnung des Bestimmtheitsmaßes ist es nicht entscheidend, welche Variable als abhängige und welche als unabhängig gewählt wird. Weitere Optionen müssen für diese Berechnung nicht ausgewählt werden.

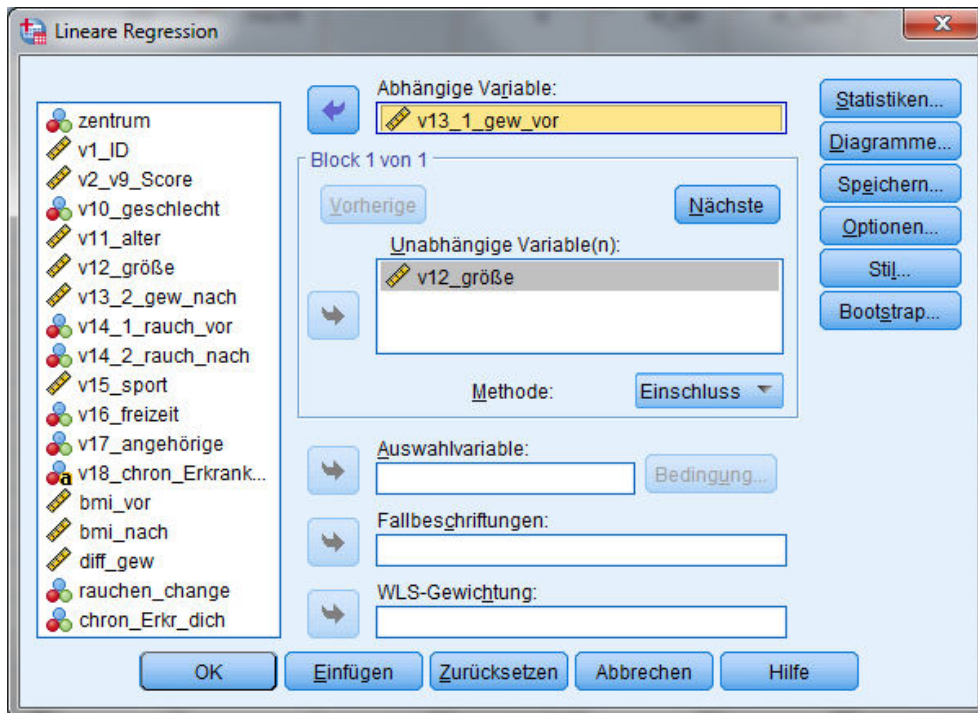


Abb. 8.8: Auswahl der Variablen zur Berechnung des Bestimmtheitsmaßes

Nach Klick auf *OK* erhält man das Ausgabefenster (siehe Abb. 8.9), wobei für die Berechnung des Bestimmtheitsmaßes nur die Tabelle „Modellzusammenfassung“ entscheidend ist.

Modellzusammenfassung				
Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,965 ^a	,931	,930	2,935

a. Einflussvariablen: (Konstante), Körpergröße

Abb. 8.9: Ausschnitt aus dem Ausgabefenster mit den Informationen über das Bestimmtheitsmaß

In dieser Tabelle erkennt man unter R erneut den Korrelationskoeffizienten nach Pearson, wie er auch in Abschnitt 8.1 berechnet wurde. Direkt daneben befindet sich das gesuchte Bestimmtheitsmaß (R-Quadrat) mit einem Wert von 0,931. Dieser Wert ist ziemlich hoch (nahe 1) und lässt sich so interpretieren, dass sich 93,1 % der Variation bei der Körpergröße in der Untersuchungsgruppe durch das Gewicht vor der Reha erklären lässt.

Zusätzlich gibt SPSS einen korrigierten Wert für das Bestimmtheitsmaß aus (Korrigiertes R-Quadrat). Dieser berücksichtigt auch die Anzahl der Beobachtungen (vgl. Janssen, Laatz 2007: 427).

Interpretation der Ergebnisse

Übungsaufgabe

- 8.1) Berechnen Sie den Korrelationskoeffizienten für die beiden Variablen *v15_sport* und *v2_v9_Score* sowie das Bestimmtheitsmaß. Ist der Zusammenhang positiv linear? Wie lässt sich der Anteil der erklärten Varianz beschreiben? Kann man schlussfolgern, dass mit dem Umfang der sportlichen Betätigung auch die Zufriedenheit der Patienten zunimmt?

Anhang 1: Kodeplan

Patientenfragebogen

Liebe Patientinnen, liebe Patienten,
wir hoffen, Sie haben Ihren Aufenthalt bei uns genossen. Bevor wir Sie nach Hause entlassen, würden wir Sie darum bitten, uns die folgenden Fragen zu Ihrer Zufriedenheit mit Ihrem Aufenthalt bei und zu Ihrer Person zu beantworten. Damit unterstützen Sie uns dabei, unsere Angebote und unseren Service stetig zu verbessern. Ihre Daten werden selbstverständlich vertraulich behandelt.

zentrum 0. Nr. des Zentrums			
0 (wird von Klinik ausgefüllt)			
v1_ID 1. Patientenidentifikationsnummer			
0 (wird bei Zustimmung des Patienten von dem Arzt ausgefüllt)			
v2_zuf 2. Wie würden Sie die Qualität der Behandlung, welche Sie erhalten haben, beurteilen?			
ausgezeichnet <input type="checkbox"/> _1	gut <input checked="" type="checkbox"/> _2	weniger gut <input type="checkbox"/> _3	schlecht <input type="checkbox"/> _4
v3_zuf 3. Haben Sie die Art von Behandlung erhalten, die Sie wollten?			
eindeutig nicht <input type="checkbox"/> _1	eigentlich nicht <input type="checkbox"/> _2	im Allgemeinen ja <input type="checkbox"/> _3	eindeutig ja <input checked="" type="checkbox"/> _4
v4_zuf 4. In welchem Maße hat unsere Klinik Ihren Bedürfnissen entsprochen?			
sie hat fast allen meinen Bedürfnissen entsprochen <input checked="" type="checkbox"/> _1	sie hat den meisten meiner Bedürfnisse entsprochen <input type="checkbox"/> _2	sie hat nur wenigen meiner Bedürfnisse entsprochen <input type="checkbox"/> _3	sie hat meinen Bedürfnissen nicht entsprochen <input type="checkbox"/> _4
v5_zuf 5. Würden Sie einem Freund / einer Freundin unsere Klinik empfehlen, wenn er / sie eine ähnliche Hilfe benötigen würde?			
eindeutig nicht <input checked="" type="checkbox"/> _1	ich glaube nicht <input type="checkbox"/> _2	ich glaube ja <input type="checkbox"/> _3	eindeutig ja <input type="checkbox"/> _4
v6_zuf 6. Wie zufrieden sind Sie mit dem Ausmaß der Hilfe, welche Sie hier erhalten haben?			
ziemlich unzufrieden <input type="checkbox"/> _1	leidlich oder leicht unzufrieden <input type="checkbox"/> _2	weitgehend zufrieden <input checked="" type="checkbox"/> _3	sehr zufrieden <input type="checkbox"/> _4
v7_zuf 7. Hat die Behandlung, die Sie hier erhielten, Ihnen dabei geholfen, angemessener mit Ihren Problemen umzugehen?			
ja, sie half eine ganze Menge <input checked="" type="checkbox"/> _1	ja, sie half etwas <input type="checkbox"/> _2	nein, sie half eigentlich nicht <input type="checkbox"/> _3	nein, sie hat mir die Dinge schwerer gemacht <input type="checkbox"/> _4
v8_zuf 8. Wie zufrieden sind Sie mit der Behandlung, die Sie erhalten haben, im Großen und Ganzen?			
sehr zufrieden <input type="checkbox"/> _1	weitgehend zufrieden <input checked="" type="checkbox"/> _2	leidlich oder leicht unzufrieden <input type="checkbox"/> _3	ziemlich unzufrieden <input type="checkbox"/> _4
v9_zuf 9. Würden Sie wieder in unsere Klinik kommen, wenn Sie eine Hilfe bräuchten?			
eindeutig nicht <input type="checkbox"/> _1	ich glaube nicht <input type="checkbox"/> _2	ich glaube ja <input checked="" type="checkbox"/> _3	eindeutig ja <input type="checkbox"/> _4
v10_geschlecht 10. Geschlecht			
<input checked="" type="checkbox"/> _0 weiblich <input type="checkbox"/> _1 männlich			
v11_alter 11. Alter			
55 Jahre			
v12_größe 12. Körpergröße			
1,64 m			
13. Körpergewicht			
v13_1_gew_vor vor Reha 85 kg			
v13_2_gew_nach nach Reha 80 kg			
14. Raucherstatus			
v14_1_rauch_vor vor Reha <input type="checkbox"/> _0 Nichtraucher <input type="checkbox"/> _1 Gelegenheitsraucher <input checked="" type="checkbox"/> _2 Raucher			
v14_2_rauch_vor nach Reha <input type="checkbox"/> _0 Nichtraucher <input checked="" type="checkbox"/> _1 Gelegenheitsraucher <input type="checkbox"/> _2 Raucher			
v15_sport 15. Durchschnittliche Anzahl an Minuten, die Sie sich während Ihres Aufenthalts täglich sportlich betätigt haben:			
60 Minuten (pro Tag)			
v16_freizeit 16. Anzahl an wahrgenommenen Freizeitangebote während Ihres Aufenthalts:			
<input type="checkbox"/> _0 bis 2 <input type="checkbox"/> _1 3-6 <input type="checkbox"/> _2 >6			
v17_angehörige 17. Einbezug von Angehörigen in Ihre Behandlung:			
<input checked="" type="checkbox"/> _1 ja <input type="checkbox"/> _0 nein			
v18_chron_Erkrankung 18. Leiden Sie neben Typ-II-Diabetes unter weiteren chronischen Erkrankungen? Falls ja, welche?			
COPD			



Vielen Dank für Ihre Unterstützung und eine gute Heimreise!
Ihr Team des Verbunds der Reha-Zentren „Nordstern“



Anmerkung: Fragen 2 bis 9 = ZUF-8 (Schmidt & Nübling, 2002; Schmidt et al. 1989, 1994).

Glossar

Fall: Erhebungseinheit in einer empirischen Untersuchung.

Grundgesamtheit: Bezeichnung der gesamten Population, für die die Aussagen einer Untersuchung gelten sollen und die sich in der Regel auf Grund der Größe und Dimension nicht erfassen lässt und deswegen geschätzt werden muss.

Median: Wert, der in der Mitte aller beobachteten Werte einer Variablen liegt.

Merkmalsausprägung: unterschiedliche Ausprägungen einer Variablen.

Mittelwert: Summe der beobachteten Werte geteilt durch die Anzahl der beobachteten Werte einer Variablen. Entspricht hier dem arithmetischen Mittelwert.

Modus: der am häufigsten auftretende Wert einer Variablen.

Nominal: Eigenschaft einer Variablen, wenn sich ihre Ausprägungen keiner Ordnung unterziehen lassen (z. B. Haarfarbe).

Ordinal: Eigenschaft einer Variablen, wenn sich ihre Ausprägungen einer Ordnung unterziehen lassen (z. B. Schulnoten).

Stetig: Eigenschaft einer Variablen, wenn ihre Ausprägungen messbar sind (z. B. Körpergröße). Andere Bezeichnung: intervallskaliert oder metrisch.

Stichprobe: zufälliges Ziehen aus der Grundgesamtheit.

Syntax: Programmiersprache.

Variable: ein bestimmtes Merkmal, das im Rahmen einer empirischen Untersuchung erhoben wird und das von Fall zu Fall unterschiedliche Ausprägungen aufweisen kann.

Zufallsstichprobe: zufällige Ziehung von Fällen aus einer definierten Gruppe von Fällen.

Lösungen zu den Übungsaufgaben

Lösungen zu Kapitel 1

- 1.1) **Nominalniveau:** Beispiel: *v17_angehörige*.

Begründung: Die Variable weist nur zwei Merkmalsausprägungen auf (ja und nein).

Ordinalniveau: Beispiel: *v16_freizeit*.

Begründung: Die Merkmalsausprägungen lassen sich nach steigender Anzahl der in Anspruch genommenen Freizeitangebote sortieren. Allerdings weisen die Merkmalsausprägungen keine gleichen Abstände auf: Man kann nicht sagen, dass der Abstand zwischen der Kategorie „bis 2“ und „3-6“ genauso groß ist wie zwischen „3-6“ und „>6“.

Intervallniveau: Beispiel: *v8_zuf*.

Begründung: Man nimmt an, dass die Abstände zwischen den vier Merkmalsausprägungen gleich sind. Jedoch ist nicht anzunehmen, dass jemand der die Ausprägung „weitgehend zufrieden“ (2) angekreuzt hat halb so zufrieden ist im Vergleich zu einer Person, die „sehr zufrieden“ (1) angekreuzt hat.

Verhältnissniveau: Beispiel: *v11_alter*.

Begründung: Die Variable weist einen natürlichen Nullpunkt auf. Wenn ein Befragter theoretisch eine 0 angibt, bedeutet dies, dass er noch kein Jahr gelebt hat. Und wenn jemand 50 angibt, ist er doppelt so alt im Vergleich zu einer Person, die 25 Jahre alt ist.

- 1.2) Die Variable *v12_größe* weist Intervallniveau auf, so dass alle bekannten Lage- und Streuungsmaße berechnet werden können (Modus, Median, Mittelwert, Quartile, Minimum, Maximum, Spannweite, Standardabweichung, Varianz).

Die Variable *v3_zuf* weist Intervallniveau auf, so dass alle bekannten Lage- und Streuungsmaße berechnet werden können (s. o.).

Die Variable *v17_angehörige* weist Nominalniveau auf, so dass es nur sinnvoll ist, hier den Modus anzugeben bzw. eine Häufigkeitstabelle zu erstellen.

Lösungen zu Kapitel 2

- 2.1) Das Ausführen der Prozedur Häufigkeiten für die Variable *v3_zuf* (*Art Behandlung*) bewirkt, dass im Viewer die folgenden Tabellen erscheinen:

Statistiken

Art Behandlung

N	Gültig	30
	Fehlend	0

Art Behandlung

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	eindeutig nicht	9	30,0	30,0	30,0
	eigentlich nicht	10	33,3	33,3	63,3
	im Allgemeinen ja	9	30,0	30,0	93,3
	eindeutig ja	2	6,7	6,7	100,0
	Gesamt	30	100,0	100,0	

Der ersten Tabelle ist zu entnehmen, dass alle 30 Patienten in der Datendatei gültige Angaben zu der Variablen gemacht haben. Der Häufigkeitstabelle ist zu entnehmen, wie sich die Antworten auf die Merkmalsausprägungen verteilen. So haben 9 Personen auf die Frage „Haben Sie die Art von Behandlung erhalten, die Sie wollten?“ „eindeutig nicht“ angekreuzt (30,0%). 10 Patienten haben die Kategorie „eigentlich nicht“ (33,3%) angekreuzt, 9 Patienten die Kategorie „im Allgemeinen ja“ (30,0%) und 2 Patienten „eindeutig ja“ (6,7%).

- 2.2) Nachdem die temporäre Auswahl der weiblichen Patienten (*v10_geschlecht=0*) getroffen und die Prozedur Häufigkeiten ausgewählt wurde, erscheinen im Viewer die folgenden Tabellen

Statistiken

Einbezug von Angehörigen

N	Gültig	19
	Fehlend	0

Einbezug von Angehörigen

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	nein	16	84,2	84,2	84,2
	ja	3	15,8	15,8	100,0
	Gesamt	19	100,0	100,0	

Der ersten Tabelle ist zum einen zu entnehmen, dass insgesamt 19 Frauen an der Befragung in Zentrum 1 teilgenommen haben, die alle gültige Angaben bei der Variablen *v17_angehörige* gemacht haben. Der Häufigkeitstabelle ist zu entnehmen, dass bei 16 Frauen kein Angehöriger in die Behandlung einbezogen war – dies entspricht 84,2% der befragten Frauen.

- 3.1) Da es sich bei der Variablen *v11_alter* um eine stetige Variable mit Verhältnisniveau handelt, können hier alle bekannten Lage- und Streuungsmaße berechnet werden. Alle 30 Personen haben gültige Angaben gemacht. Der Mittelwert des Alters liegt bei 57,13 Jahren mit einer Standardabweichung von 3,192. Die Standardabweichung ist relativ gering, was bereits andeutet, dass das Alter der Patienten und Patientinnen sehr ähnlich ist. Dies bestätigt sich, wenn man einen Blick auf das Minimum von 51 Jahren und das Maximum von 63 Jahren wirft. Das heißt, die Spannweite liegt bei 12 Jahren. Die meisten Patienten und Patientinnen sind 56 Jahre alt (Modus).

Lösungen zu Kapitel 3

Statistiken

Alter

N	Gültig	30
	Fehlend	0
Mittelwert		57,13
Median		57,00
Modus		56
Std.-Abweichung		3,192
Varianz		10,189
Spannweite		12
Minimum		51
Perzentile	25	55,00
	50	57,00
	75	60,00

Um die gleiche Tabelle nun für die Patienten und Patientinnen zu erstellen, die vor der Reha angegeben haben, Raucher zu sein (*v14_1_rauch_vor = 2*), ist zunächst eine temporäre Fallauswahl vorzunehmen. Anschließend sind wieder die entsprechenden Lage- und Streuungsmaße zu berechnen, so dass im Viewer die folgende Ausgabe erscheint:

Statistiken

Alter

N	Gültig	3
	Fehlend	0
Mittelwert		56,00
Median		57,00
Modus		53 ^a
Std.-Abweichung		2,646
Varianz		7,000
Spannweite		5
Minimum		53
Maximum		58
Perzentile	25	53,00
	50	57,00
	75	.

a. Mehrere Modi vorhanden. Der kleinste Wert wird angezeigt.

Der Tabelle ist zu entnehmen, dass nur 3 Personen angegeben haben vor der Reha Raucher gewesen zu sein. Der Mittelwert des Alters liegt in dieser Subgruppe bei 56,0 Jahren mit einer Standardabweichung von 2,646 Jahren. Da es in der Subgruppe nur noch 3 Patienten sind, gibt es anscheinend genauso viele unterschiedliche Messwerte wie Patienten. Dies erkennt man z. B. an der Anmerkung, dass es mehrere Modi gibt und dass der kleinste Wert angezeigt wird.

- 3.2) Da es sich bei der Variablen *v14_2_rauch_nach* um ein Merkmal mit Ordinalniveau handelt, sind die folgenden Lage- und Streuungsmaße geeignet: Modus, Median, Quartile, Minimum, Maximum. Die berechneten Maße werden im Viewer folgendermaßen angezeigt (siehe Tabelle „Statistiken“). Um die Angaben nun sinnvoll interpretieren zu können, ist es erforderlich, sich den Kodeplan daneben zu legen. So entspricht die 0 sowohl dem Median als auch der am häufigsten vorkommenden Merkmalsausprägung (Modus). Die 0 entspricht der Kategorie Nichtraucher. Das heißt, nach der Reha haben die meisten Personen angegeben, Nichtraucher zu sein. Aber anscheinend ist auch noch mindestens ein Raucher dabei, da als Merkmalsausprägungen die 2 (= Raucher) immer noch vorkommt.

Statistiken

Raucherstatus nach Reha

N	Gültig	27
	Fehlend	3
Median		,00
Modus		0
Spannweite		2
Minimum		0
Maximum		2
Perzentile	25	,00
	50	,00
	75	,00

Lösungen zu Kapitel 4

- 4.1) Da es schwierig ist, direkt aus dem Kreisdiagramm abzulesen, wie die genauen relativen Häufigkeiten der Merkmalsausprägungen lauten, wird empfohlen, die Häufigkeitstabellen mit zum Kreisdiagramm ausgeben zu lassen. Dann sieht die Ausgabe für die Variable *v17_angehörige* folgendermaßen aus: Insgesamt haben 42,2% der Patienten und Patientinnen angegeben, dass ihre Angehörige mit in die Behandlung einbezogen wurden.

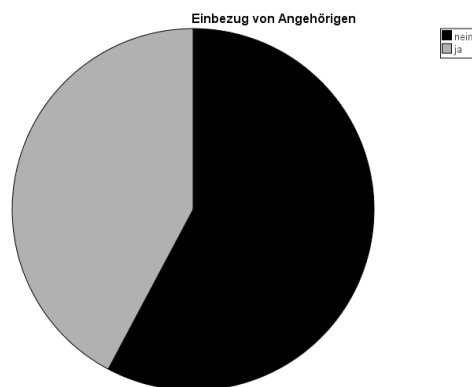
Statistiken

Einbezug von Angehörigen

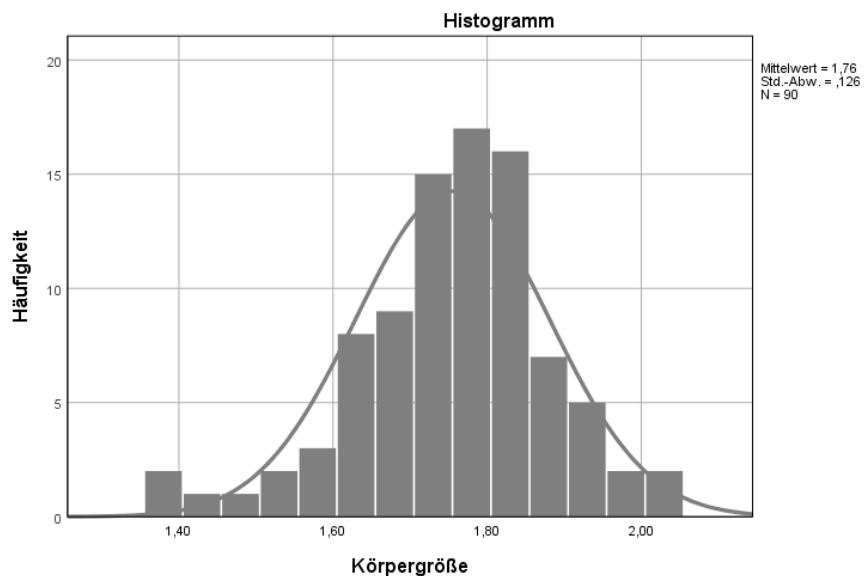
N	Gültig	90
	Fehlend	0

Einbezug von Angehörigen

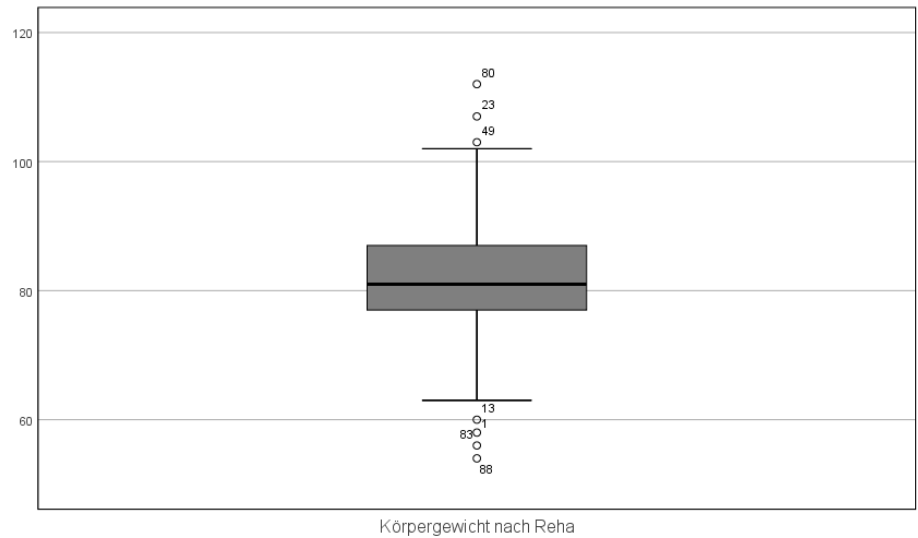
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	nein	52	57,8	57,8	57,8
	ja	38	42,2	42,2	100,0
	Gesamt	90	100,0	100,0	



- 4.2) Das Histogramm für die Variable *v12_körpergröße* zeigt deutlich die Form einer Normalverteilung

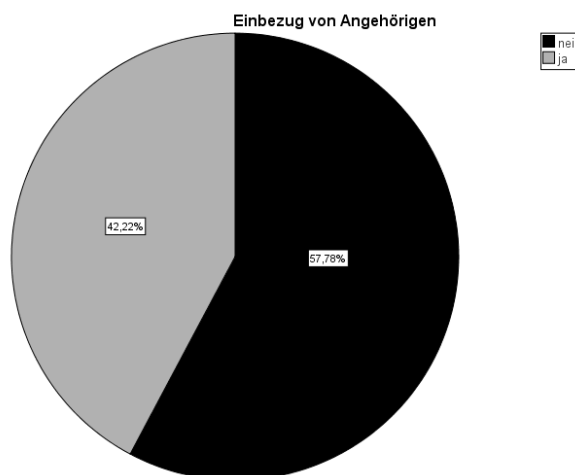


- 4.3) Der Boxplot zu der Variablen *v13_2_gew_nach* spricht für eine gleichmäßige Verteilung der Messwerte um den Median. Es fallen insbesondere die vielen Ausreißer auf, die es aber sowohl oberhalb als auch unterhalb des Medians gibt.

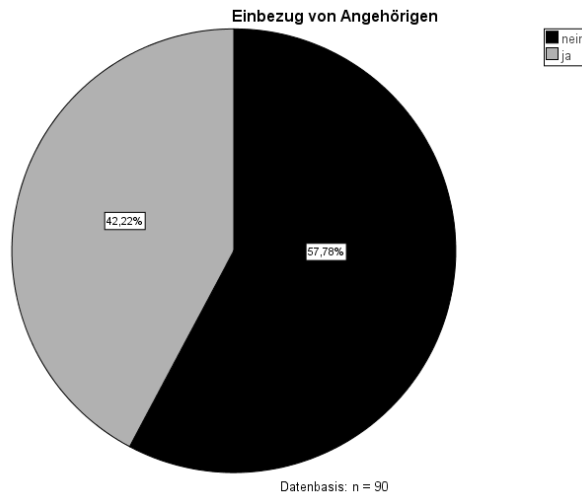


Lösungen zu Kapitel 5

- 5.1) Um sich die relativen Häufigkeiten anzeigen zu lassen, ist der Diagramm-Editor zu öffnen. Dann ist das gesamte Kreisdiagramm so zu markieren, dass es sich automatisch gelb umrandet. Anschließend ist die rechte Maustaste zu klicken, so dass sich das Kontextmenü öffnet. Im Menü ist der vorletzte Menüpunkt *Elemente* → *Datenbeschriftung einblenden* auszuwählen.



- 5.2) Um links unten im Kreisdiagramm den Text „Datenbasis: n = 90“ einzufügen, ist zunächst der Diagramm-Editor zu öffnen. Anschließend ist das Kontextmenü zu öffnen und der Punkt *Hinzufügen Textfeld* ist auszuwählen. Daraufhin erscheint ein Textfeld im Diagramm. Dort ist der Text „Datenbasis: n = 90“ einzutragen. Abschließend ist das Textfeld durch die „Klicken-und Ziehen-Methode“ an die richtige Stelle zu verschieben.



- 5.3) Um die Zeilen und Spalten in der Häufigkeitstabelle zu der Variablen *v3_zuf* zu vertauschen, ist diese zunächst einmal im Viewer so zu markieren, dass ein roter Pfeil links neben ihr erscheint. Anschließend ist das Kontextmenü zu öffnen und der Punkt *Inhalt bearbeiten* → *In separatem Fenster* auszuwählen. Darauf hin öffnet sich ein separates Fenster zum Bearbeiten von Pivottabellen. Dort ist im Menü die folgende Einstellung vorzunehmen: *Pivot* → *Zeilen und Spalten transponieren*. Diese Funktion wird direkt umgesetzt, so dass nach dem Schließen des Fensters zur Pivottable die veränderte Tabelle in der Ausgabe erscheint.

Art Behandlung

	Gültig				
	eindeutig nicht	eigentlich nicht	im Allgemeinen ja	eindeutig ja	Gesamt
Häufigkeit	9	10	9	2	30
Prozent	30,0	33,3	30,0	6,7	100,0
Gültige Prozente	30,0	33,3	30,0	6,7	100,0
Kumulierte Prozente	30,0	63,3	93,3	100,0	

- 6.1) Um die Variable *v2_v9_Score* anhand des Medians zu dichotomisieren, ist dieser zunächst zu bestimmen (z. B. über die Prozedur *Analysieren* → *deskriptive Statistik* → *Häufigkeiten*). Dieser beträgt hier 19. Mit Hilfe der Funktion *Transformieren* → *Variable berechnen* kann nun eine neue dichotomisierte Variable erstellt werden (im Folgenden *score_dich* genannt). Zur Durchführung der Dichotomisierung kann Studienbrief 1, Abschnitt 4.1.2 (Umkodieren in andere Variable) wiederholend hinzugezogen werden. Als Dichotomisierung wird folgende Einteilung verwendet:

score_dich = 2 wenn *v2_v9_Score* ≤ 19, d. h. wenn der Zufriedenheitsscore kleiner oder gleich 19 ist, wird der dichotomisierte Score auf 2 gesetzt und anschließend die Variablenausprägung „unzufrieden“ vergeben.

score_dich = 1 wenn *v2_v9_Score* > 19, d. h. wenn der Zufriedenheitsscore größer als 19 ist, wird der dichotomisierte Score auf 1 gesetzt und die Variablenausprägung „zufrieden“ vergeben.

Zur Kontrolle der erfolgreichen Umkodierung, wird empfohlen, einen Blick in den Dateneditor zu werfen oder sich eine Häufigkeitstabelle ausgeben zu lassen.

- 6.2) Die Forschungshypothesen für die Fragestellung „Unterscheiden sich Personen, bei denen Angehörige in die Reha-Maßnahme mit einbezogen wurden bezüglich ihrer Zufriedenheit?“ und unter Berücksichtigung der durchgeführten Dichotomisierung lauten:

H_0 : Das Einbeziehen von Angehörigen während der Reha bewirkt keinen Unterschied hinsichtlich der Rate an zufriedenen Patienten.

H_1 : Das Einbeziehen von Angehörigen während der Reha bewirkt einen Unterschied hinsichtlich der Rate an zufriedenen Patienten.

Formal lassen sich die Hypothesen folgendermaßen beschreiben:

$H_0: p_{\text{zufAng}} = p_{\text{zufAngohne}}$

$H_1: p_{\text{zufAng}} \neq p_{\text{zufAngohne}}$

Lösungen zu Kapitel 6

- 6.3) Die Kreuztabelle ist über folgende Prozedur zu erstellen: *Analysieren* → *deskriptive Statistiken* → *Kreuztabellen*. In die Zeilen ist die Variable *v17_angehörige* zu klicken und in die Spalten die Variable *score_dich*. Zudem sind unter Zellen, die Zeilenprozente zu aktivieren. Die Tabelle in der Ausgabe „Verarbeitete Fälle“ zeigt, dass es keine fehlenden Werte gibt und die Kreuztabelle auf Basis aller 90 Fälle in der Datendatei erstellt wurde. Der Kreuztabelle ist zu entnehmen, dass 68,4 % der Personen, bei denen die Angehörigen mit in die Behandlung einbezogen wurden, zufrieden sind im Gegensatz zu 31,6 %, bei denen die Angehörigen nicht mit in die Behandlung einbezogen wurden. Das heißt, die Stichprobenergebnisse zeigen einen Unterschied in Bezug auf die Zufriedenheit der Patienten in Abhängigkeit vom Einbezug der Angehörigen.

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
Einbezug von Angehörigen * score_dich	90	100,0%	0	0,0%	90	100,0%

Einbezug von Angehörigen * score_dich Kreuztabelle

			score_dich		Gesamt
			,00	1,00	
Einbezug von Angehörigen	nein	Anzahl	36	16	52
		% innerhalb von Einbezug von Angehörigen	69,2%	30,8%	100,0%
	ja	Anzahl	12	26	38
		% innerhalb von Einbezug von Angehörigen	31,6%	68,4%	100,0%
Gesamt	Anzahl		48	42	90
	% innerhalb von Einbezug von Angehörigen		53,3%	46,7%	100,0%

- 6.4) Um den Chi-Quadrat-Test durchzuführen, ist in dem Dialogfenster, welches über die Prozedur *Analysieren* → *deskriptive Statistik* → *Kreuztabellen* erscheint, unter *Statistiken* der Chi-Quadrat-Test zu aktivieren. Der Tabelle in der Ausgabe ist für die Chi-Quadrat-Tests folgendes zu entnehmen: der Chi-Quadrat-Test nach Pearson hat einen Testwert von 12,506. Bei der Asymptotischen Signifikanz (zweiseitig) – was dem p-Wert entspricht – ist ein Wert von 0,000 zu finden. Dies bedeutet, dass der p-Wert nicht nur kleiner als 0,05 ist sondern kleiner als 0,000. Das heißt, hier liegt ein signifikantes Ergebnis vor, so dass die H_0 („Das Einbeziehen von Angehörigen während der Reha bewirkt keinen Unterschied hinsichtlich der Rate an zufriedenen Patienten.“) zugunsten der H_1 („Das Einbeziehen von Angehörigen während der Reha bewirkt einen Unterschied hinsichtlich der Rate an zufriedenen Patienten.“) abgelehnt werden kann. Patient(inn)en, deren Angehörige mit in die Behandlung einbezogen werden, sind zufriedener.

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	12,506 ^a	1	,000		
Kontinuitätskorrektur ^b	11,039	1	,001		
Likelihood-Quotient	12,775	1	,000		
Exakter Test nach Fisher				,001	,000
Zusammenhang linear-mit-linear	12,367	1	,000		
Anzahl der gültigen Fälle	90				

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 17,73.

b. Wird nur für eine 2x2-Tabelle berechnet

- 7.1) Die Nullhypothese lautet: Es gibt keinen Unterschied zwischen Männern und Frauen hinsichtlich des Zufriedenheitscores.

Lösung zu Kapitel 7

Die beiden Variablen *v2_v9_Score* und *v10_geschlecht* werden unter *Analysieren* → *Mittelwerte vergleichen* → *T-Test bei unabhängigen Stichproben* ausgewählt. Die Gruppenvariable (Gruppe 1=0, Gruppe 2=1) wird bestimmt. Es ergibt sich folgendes Ausgabefenster:

Gruppenstatistiken

	Geschlecht	N	Mittelwert	Std.-Abweichung	Standardfehler des Mittelwertes
ZUF 8 Score	weiblich	64	19,88	4,600	,575
	männlich	26	18,54	4,785	,938

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	95% Konfidenzintervall der Differenz	
									Untere	Obere
ZUF 8 Score	Varianzen sind gleich	,054	,817	1,235	88	,220	1,337	1,082	-,814	3,487
	Varianzen sind nicht gleich			1,214	44,788	,231	1,337	1,101	-,880	3,554

Laut Levene-Test sind die Varianzen gleich und der *t*-Test führt zu einem nicht signifikanten Ergebnis mit einem *p*-Wert von 0,220. Die Nullhypothese wird also beibehalten. Es konnte kein Unterschied hinsichtlich des Zufriedenheitscores nach Geschlecht festgestellt werden.

- 8.1) Die Berechnung des Korrelationskoeffizienten lässt sich folgendermaßen berechnen: *Analysieren* → *Korrelation* → *Bivariate Korrelationen*. Es wird der Korrelationskoeffizient nach Pearson verwendet, da von einem linearen Zusammenhang ausgegangen wird und beide Variablen stetig sind. Das Ausgabefenster zeigt folgendes Ergebnis:

Lösung zu Kapitel 8

Korrelationen

		ZUF 8 Score	Sportliche Betätigung
ZUF 8 Score	Korrelation nach Pearson	1	,305**
	Signifikanz (2-seitig)		,003
	N	90	90
Sportliche Betätigung	Korrelation nach Pearson	,305**	1
	Signifikanz (2-seitig)	,003	
	N	90	90

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Der Korrelationskoeffizient beträgt 0,305. Es kann also nur von einem schwachen positiv linearen Zusammenhang ausgegangen werden.

Das Bestimmtheitsmaß lässt sich folgendermaßen berechnen: *Analysieren* → *Regression* → *linear*.

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,305 ^a	,093	,083	4,469

a. Einflußvariablen : (Konstante), Sportliche Betätigung

Der Anteil der erklärten Varianz ist mit 0,093 sehr klein.

Literaturverzeichnis

- Bortz, J. (2005). *Statistik für Human-und Geisteswissenschaftler* (6. Auflage). Berlin, Heidelberg: Springer.
- Bühl, A. (2010). *SPSS 18. Einführung in die moderne Datenanalyse* (12. Auflage). München: Pearson Studium.
- Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (2007). *Statistik* (6. Auflage). Berlin, Heidelberg: Springer.
- Janssen, J. & Laatz, W. (2007). *Statistische Datenanalyse mit SPSS für Windows* (6. Auflage). Berlin/Heidelberg: Springer.
- Martens, J. (2003). *Statistische Datenanalyse mit SPSS für Windows* (2. Auflage). München: Oldenbourg.
- Rasch, B., Frieze, M., Hofmann, W. & Naumann, E. (2010a). *Quantitative Methoden. Band 1* (3. Auflage). Berlin, Heidelberg: Springer.
- Rasch, B., Frieze, M., Hofmann, W. & Naumann, E. (2010b). *Quantitative Methoden. Band 2* (3. Auflage). Berlin, Heidelberg: Springer.
- Schendera, C. F. G. (2004). *Datenmanagement und Datenanalyse mit dem SAS-System*. München: Oldenbourg.
- Weiß, C. (2002). *Basiswissen Medizinische Statistik* (2. Auflage). Berlin, Heidelberg: Springer.
- Zöfel, P. (2002). *SPSS-Syntax: Die ideale Ergänzung für effiziente Datenanalyse*. München: Pearson Studium.
- Zöfel, P. & Bühl, A. (2000). *Statistik verstehen: Ein Begleitbuch zur computerunterstützten Anwendung*. München: Addison-Wesley.