

Skript  
Computergestützte Statistik I

Prof. Dr. Bernd Rönz



Humboldt-Universität zu Berlin  
Wirtschaftswissenschaftliche Fakultät  
Institut für Statistik und Ökonometrie  
2001



# Inhaltsverzeichnis

<b>1. Einführung</b>	<b>1</b>
<b>2. Entdeckung und Identifikation von Ausreißern</b>	<b>17</b>
2.1. Stem-and-Leaf Plot . . . . .	17
2.2. Boxplot . . . . .	22
2.3. Scatterplot . . . . .	25
2.4. Andrews-Plot . . . . .	33
2.5. Ausreißertests . . . . .	35
2.6. Robuste Schätzer für die Lokalisation und Streuung . . . . .	43
<b>3. Prüfung der Verteilungsform von Variablen</b>	<b>65</b>
3.1. Explorative Datenanalyse . . . . .	66
3.2. Beschreibende Datenanalyse . . . . .	91
3.3. Statistische Tests . . . . .	100
3.3.1. Kolmogorov-Smirnov-Test . . . . .	101
3.3.2. Chi-Quadrat-Anpassungstest . . . . .	109
3.3.3. Binomial-Test . . . . .	120
3.4. Transformationen . . . . .	125
<b>4. Parametervergleiche bei unabhängigen Stichproben</b>	<b>135</b>
4.1. Explorative Analyse . . . . .	136
4.2. Statistische Tests . . . . .	142
4.2.1. Prüfung der Gleichheit der Varianzen . . . . .	142
4.2.2. Prüfung der Gleichheit der Mittelwerte mittels parametrischer Tests . . . . .	149
4.2.2.1. Test der Mittelwerte zweier Grundgesamtheiten mit gleichen, unbekannten Varianzen . . . . .	149
4.2.2.2. Test der Mittelwerte zweier Grundgesamtheiten mit ungleichen, unbekannten Varianzen . . . . .	152

## Inhaltsverzeichnis

4.2.2.3. Test der Mittelwerte mehrerer Grundgesamtheiten . . . . .	159
4.2.2.4. Multiple Mittelwertvergleiche . . . . .	164
<b>Anhang A: Testentscheidung unter Verwendung statistischer Software</b>	<b>195</b>
<b>Anhang B: Kolmogorov-Smirnov-Test: Quantile <math>d_{n;1-\alpha}</math> der Teststatistik <math>D_n</math></b>	<b>201</b>
<b>Anhang C: <math>\chi^2</math>-Verteilung: Quantile <math>\chi_{n;1-\alpha^2}</math> der Verteilungsfunktion F</b>	<b>203</b>
<b>Anhang D: Zur Varianzanalyse</b>	<b>205</b>
<b>Literaturverzeichnis</b>	<b>211</b>
Index . . . . .	215

# Abbildungsverzeichnis

1.1. Dialogfeld „Sort Cases“ . . . . .	5
1.2. Dialogfeld „Split File“ . . . . .	6
1.3. Dialogfeld „Recode into Different Variables“ . . . . .	7
1.4. Dialogfeld „Recode into Different Variables: Old and new Values“ . . . . .	8
1.5. Dialogfeld „Select Cases“ . . . . .	9
1.6. Dialogfeld „Select Cases: If“ . . . . .	10
1.7. Dialogfeld „Select Cases: Random Sample“ . . . . .	10
1.8. Dialogfeld „Select Cases: Range“ . . . . .	11
1.9. Dialogfeld „Weight Cases“ . . . . .	12
1.10. Dialogfeld „Compute Variable“ . . . . .	13
1.11. Dialogfeld „Compute Variable: Type and Label“ . . . . .	13
1.12. Dialogfeld „Compute Variable: If Cases“ . . . . .	14
1.13. Dialogfeld „Automatic Recode“ . . . . .	15
2.1. Dialogfeld „Explore“ . . . . .	18
2.2. Dialogfeld „Explore: Plots“ . . . . .	18
2.3. Dialogfeld „Explore: Statistics“ . . . . .	22
2.4. Schematische Darstellung eines Boxplots . . . . .	23
2.5. Dialogfeld „Boxplot“ . . . . .	24
2.6. Dialogfeld „Define Simple Boxplot: Summaries of Separate Variables“ . . . . .	24
2.7. Beispiel eines Boxplots . . . . .	25
2.8. Dialogfeld „Scatterplot“ . . . . .	26
2.9. Dialogfeld „Simple Scatterplot“ . . . . .	27
2.10. Beispiel für einen einfachen Scatterplot . . . . .	27
2.11. Point Selection Ikon . . . . .	28
2.12. Dialogfeld „Scatterplot Matrix“ . . . . .	28
2.13. Beispiel einer Scatterplot-Matrix . . . . .	29

## Abbildungsverzeichnis

2.14. Dialogfeld „3-D Scatterplot“ . . . . .	30
2.15. Beispiel eines 3-D Scatterplot . . . . .	30
2.16. Dialogfeld „3-D Scatterplot: Options“ . . . . .	31
2.17. Beispiel eines 3-D Scatterplot mit Floor-Projektion . . . . .	31
2.18. Menüleiste des Chart Editors . . . . .	32
2.19. Dialogfeld „3-D Rotation“ . . . . .	32
2.20. Menüleiste des Spin Mode . . . . .	32
2.21. Beispiel eines rotierten 3-D Scatterplots . . . . .	33
2.22. Andrews-Plot . . . . .	35
2.23. Dialogfeld „Descriptives“ . . . . .	37
2.24. Dialogfeld „Descriptives: Options“ . . . . .	37
2.25. Boxplot für Beispiel 2.5 . . . . .	38
2.26. $\theta(z)$ für den Huber-1-Schätzer . . . . .	51
2.27. $\xi(z)$ für den Huber-1-Schätzer . . . . .	51
2.28. Gewichtsfunktion $w(z)$ des Huber-1-Schätzers . . . . .	52
2.29. $\theta(z)$ des Hampel-Schätzers für $a = 1,7$ , $b = 3,4$ und $c = 8,5$ . . . . .	54
2.30. $\xi(z)$ des Hampel-Schätzers für $a = 1,7$ , $b = 3,4$ und $c = 8,5$ . . . . .	54
2.31. Gewichtsfunktion $w(z)$ des Hampel-Schätzers für $a = 1,7$ , $b = 3,4$ und $c = 8,5$ . . . . .	54
2.32. $\theta(z)$ des Andrews-Schätzers für $a = 1$ . . . . .	55
2.33. $\xi(z)$ des Andrews-Schätzers für $a = 1$ . . . . .	56
2.34. Gewichtsfunktion $w(z)$ des Andrews-Schätzers für $a = 1$ . . . . .	56
2.35. Funktion $\theta(z)$ des Tukey's biweight für $a = 1$ . . . . .	57
2.36. Funktion $\xi(z)$ des Tukey's biweight für $a = 1$ . . . . .	57
2.37. Gewichtsfunktion $w(z)$ des Tukey-Schätzers für $a = 1$ . . . . .	58
3.1. Haushaltsgrößen im früheren Bundesgebiet . . . . .	66
3.2. Dialogfeld „Bar Charts“ . . . . .	68
3.3. Dialogfeld „Define Simple Bar: Summaries for Groups of Cases“ . . . . .	68
3.4. Dialogfeld „Define Simple Bar: Values of Individual Cases“ . . . . .	69
3.5. Dialogfeld „Frequencies“ . . . . .	70
3.6. Dialogfeld „Frequencies: Charts“ . . . . .	70
3.7. Dialogfeld „Define Clustered Bar: Summaries for Groups of Cases“ . . . . .	71
3.8. Gruppiertes Balkendiagramm für Beispiel 3.1 . . . . .	72
3.9. Boxplot des monatlichen persönlichen Nettoeinkommens nach dem Geschlecht . . . . .	76
3.10. Dialogfeld „Histogram“ . . . . .	77
3.11. Dialogfeld „Interval Axis“ . . . . .	78
3.12. Dialogfeld „Interval Axis: Define Custom Intervals“ . . . . .	79

3.13. Histogramm mit Normalverteilung für Beispiel 3.2, Klassenbreite 200 . . . . .	80
3.14. Histogramm mit Normalverteilung für Beispiel 3.2, Klassenbreite 450 . . . . .	81
3.15. Stabdiagramm der relativen Häufigkeiten . . . . .	83
3.16. Normalkerne über den Beobachtungen . . . . .	83
3.17. Kerndichteschätzer mit Normalkern und Bandbreite $w = 0,8$ . . . . .	84
3.18. Dialogfeld „Q-Q Plots“ . . . . .	87
3.19. Histogramm für Beispiel 3.4 . . . . .	87
3.20. Normal Q-Q Plot für Beispiel 3.4 . . . . .	88
3.21. Trendbereinigter Normal Q-Q Plot für Beispiel 3.4 . . . . .	88
3.22. Normal P-P Plot für Beispiel 3.4 . . . . .	89
3.23. Trendbereinigter Normal P-P Plot für Beispiel 3.4 . . . . .	89
3.24. Histogramm für Beispiel 3.4 mit den ln-transformierten Werten . . . . .	90
3.25. Normal Q-Q Plot für Beispiel 3.4 mit den ln-transformierten Werten . . . . .	90
3.26. Trendbereinigter Normal Q-Q Plot für Beispiel 3.4 mit den ln-transformierten Werten . . . . .	90
3.27. Dialogfeld „One-Sample Kolmogorov-Smirnov-Test“ . . . . .	105
3.28. Dialogfeld „Exact Tests“ . . . . .	108
3.29. Nichtablehnungsbereich und Ablehnungsbereich der $H_0$ des $\chi^2$ -Anpassungstests .	114
3.30. Dialogfeld „Chi-Square Test“ . . . . .	115
3.31. Balkendiagramm der Variablen persönliches Nettoeinkommen (klassiert) . . . . .	118
3.32. Dialogfeld „Binomial Test“ . . . . .	121
3.33. Potenzfunktionen $T_p(x)$ für ausgewählte p-Werte . . . . .	128
3.34. Histogramm der Variablen SQR(barrel) des Beispiels 3.10 . . . . .	132
3.35. Histogramm der Variablen ln(barrel) des Beispiels 3.10 . . . . .	132
3.36. Histogramm der Variablen z des Beispiels 3.10 . . . . .	133
3.37. Boxplots der Variablen barrel und z des Beispiels 3.10 . . . . .	133
4.1. Histogramme des monatlichen persönlichen Nettoeinkommens nach Geschlecht .	137
4.2. Dialogfeld „Error Bar“ . . . . .	139
4.3. Dialogfeld „Define Simple Error Bar: Summaries for groups of cases“ . . . . .	139
4.4. Fehlerbalkendiagramm (95%iges Konfidenzintervall für den Mittelwert) für Nettoeinkommen nach Geschlecht . . . . .	141
4.5. Fehlerbalkendiagramm (95%iges Konfidenzintervall für den Mittelwert) für Nettoeinkommen nach dem Schulabschluß . . . . .	141
4.6. Spread vs. Level Plot von monatlichem Nettoeinkommen nach Geschlecht . . . . .	147
4.7. Spread vs. Level Plot von monatlichem Nettoeinkommen nach Schulabschluß .	148
4.8. Dialogfeld „Independent-Samples T Test“ . . . . .	153

## Abbildungsverzeichnis

4.9. Dialogfeld „Define Groups“ . . . . .	154
4.10. Verteilung der Zufallsvariablen $X_1$ , $X_2$ und $X_3$ in den drei Grundgesamtheiten . . . . .	161
4.11. Dialogfeld „One-Way ANOVA“ . . . . .	178
4.12. Dialogfeld „One-Way ANOVA: Options“ . . . . .	179
4.13. Dialogfeld „One-Way ANOVA: Post Hoc Multiple Comparisons“ . . . . .	180
4.14. Dialogfeld „One-Way ANOVA: Contrasts“ . . . . .	180
4.15. Means Plot für Beispiel 4.3 . . . . .	183
A.1. Signifikanzniveau $\alpha$ und Entscheidungsbereiche beim rechtsseitigen Test . . . . .	196
A.2. Überschreitungswahrscheinlichkeit $P = P(V > v \vartheta_0)$ bei Gültigkeit der $H_0$ . . . . .	197
A.3. Signifikanzniveau $\alpha = P(V > c \vartheta_0)$ und Überschreitungswahrscheinlichkeit $P = P(V > v \vartheta_0)$ bei Gültigkeit der Nullhypothese $H_0$ für einen rechtsseitigen Test . . . . .	198
A.4. Signifikanzniveau $\alpha = P(V > c \vartheta_0)$ und Überschreitungswahrscheinlichkeit $P = P(V > v \vartheta_0)$ bei Gültigkeit der Nullhypothese $H_0$ für einen rechtsseitigen Test . . . . .	199

# 1. Einführung

Statistische Datenanalyse in Wissenschaft, Wirtschaft, Verwaltung und Politik ist heutzutage ohne Unterstützung durch Computer kaum durchführbar. Rechenaufwendige Analysen können in kürzester Zeit bewältigt werden; andererseits sind eine Reihe von statistischen Methoden erst durch schnellere Computer mit ausreichender Speicherkapazität praktisch durchführbar geworden. Für die Bewältigung solcher Datenanalysen stehen eine Vielzahl von statistischen Software-Paketen zur Verfügung. Die Anwendung dieser statistischen Datenanalysesysteme erfordert jedoch ein umfangreiches Wissen in Statistik, um entsprechend der fachlichen Problemstellung die angemessenen statistischen Verfahren unter Berücksichtigung ihrer Voraussetzungen auszuwählen und aus den Ergebnissen, die der Computer als Output liefert, die richtigen Schlüsse zu ziehen und Fehlinterpretationen zu vermeiden. Computergestützte Statistik (Computer-Statistik, Computational Statistics) wird deshalb als Vermittlung von Kenntnissen über die Aufbereitung und Analyse statistischer Daten verstanden, zu deren Durchführung der Computer und ein statistisches Datenanalysesystem (Software-Paket) als Hilfsmittel eingesetzt wird.

Für die Lehrveranstaltung „Computergestützte Statistik“ wurde SPSS for Windows Release 10.0.7 (im weiteren kurz SPSS genannt, <http://www.spss.com>) ausgewählt, nicht weil es den anderen Statistik-Softwarepaketen wesentlich überlegen ist, sondern weil es u.E. in vielen Wirtschaftsbereichen praktisch eingesetzt wird. Zum anderen sind spezielle Kenntnisse der Programmsyntax für die grundlegende Handhabung von SPSS nicht erforderlich, da es im wesentlichen menü- und dialoggesteuert ist und viele Aufgaben durch Anklicken mit der Maus ausgewählt und abgearbeitet werden können. Grundlagen der Handhabung von SPSS werden im weiteren vorausgesetzt.

In anderen Statistik-Lehrveranstaltungen im Hauptstudium wird auch andere Statistik-Software herangezogen, wie z.B. XploRe (<http://www.xplore-stat.de/index.js.html>) und S-Plus (<http://www.splus.mathsoft.com>), so dass keine Einseitigkeit in der Handhabung von statistischen Datenanalysesystemen auftreten kann.

## 1. Einführung

Man sollte sich jedoch vor dem Gedanken hüten, dass man nunmehr ein Statistik-Softwarepaket hat, es mit Daten füttert, statistische Prozeduren mechanisch abarbeitet und anschließend etwas Vernünftiges herauskommt.

Zum einen ist das Wissen um den Inhalt der verwendeten Daten und die diesen Daten zugrundeliegende Problemstellung eine wesentliche Voraussetzung für eine sachgerechte statistische Auswertung. Solange man zur Beantwortung einer wissenschaftlichen oder praktischen Fragestellung selbst die dafür benötigten Daten erhebt (Primärerhebung), ist dies im allgemeinen gegeben. Werden jedoch Sekundärdaten verwendet, so entfallen entscheidende Etappen der statistischen Arbeit:

- die Planungs- und Definitionsphase mit der Formulierung der Problem- und Zielstellung der Untersuchung und ihrer theoretischen Begründung, mit der Festlegung der Grundgesamtheit, der statistischen Einheiten, der zu erhebenden Variablen (einschließlich ihrer Adäquation), der Art und des Umfangs der Erhebung,
- die Erhebungsphase mit der Erarbeitung des Erhebungsinstrumentes (z.B. eines Fragebogens), der Festlegung des Erhebungsplans und der Durchführung der eigentlichen Erhebung,
- die Aufbereitungsphase, worunter hier nur die technische Aufbereitung (wie Codierung der Daten, Übertragung auf computerlesbare Medien, Datenprüfung und -korrektur) verstanden wird.

Alle Festlegungen dieser Arbeitsetappen können mit der eigenen Zielstellung der statistischen Untersuchung differieren. Auch bei einer statistischen Sekundäranalyse sollte man sich deshalb soweit wie möglich einen Überblick über diese Etappen verschaffen, um Fehlverwendungen der Daten vorzubeugen und eine sachgerechte statistische Auswertung zu gewährleisten.

Eine weitere Fehlerquelle für unrealistische oder gar unsinnige statistische Ergebnisse ist eine nicht adäquate Auswahl der anzuwendenden statistischen Verfahren und Modelle, die oftmals in der Unkenntnis der Voraussetzungen dieser Methoden begründet liegt und besonders im Zusammenhang mit der gegebenen Menü- und Dialogsteuerung der Software durch unbedarfe Handhabung hervorgerufen wird. Es ist immer noch der Nutzer der Software, der entscheidet, welches Verfahren auf welche Daten angewendet werden soll. Die statistische Software liefert auch bei Nichteinhaltung der Voraussetzung (fast immer) irgendwelche Ergebnisse. Zum Beispiel werden auch für eine nominalskalierte Variable, deren Ausprägungen in Form von Schlüsselzahlen vorliegen, (arithmetischer) Mittelwert und Streuung berechnet.

Vor Beginn der statistischen Datenanalyse (Analysephase) und somit der Methodenauswahl ist deshalb, ausgehend von der fachlichen Problemstellung, das statistische Ziel der Untersuchung zu definieren. Dies beinhaltet zwei grundsätzliche Aspekte:

- zum einen die Entscheidung über die Anzahl der gleichzeitig zu untersuchenden Variablen, d.h. ob
  - eine **univariate Analyse** (Einbeziehung nur einer Variablen)
  - eine **bivariate Analyse** (Einbeziehung zweier Variablen) oder
  - eine **multivariate Analyse** (Einbeziehung von mehr als zwei Variablen)

durchzuführen ist,

- zum anderen die Festlegung der statistischen Herangehensweise, d.h. ob
  - eine Beschreibung der Untersuchungsgesamtheit anhand ausgewählter Variablen erfolgen soll. Hierbei kommen vor allem Methoden der **deskriptiven Statistik** zur Anwendung.
  - die Analyse der Generierung von statistisch überprüfbaren Hypothesen dienen soll. Im Vordergrund stehen dabei Verfahren der **explorativen Datenanalyse**.
  - aus Forschungshypothesen abgeleitete statistische Hypothesen getestet werden sollen. Dies erfordert Methoden und Modelle der **induktiven Statistik**.

Diese drei Herangehensweisen sind nicht im Sinne „entweder ... oder“ zu verstehen, sondern sie werden vielmehr Schritte einer stufenweisen Analyse sein.

Diese Aspekte entscheiden über die Auswahl der statistischen Methode(n), wobei in jeder Phase der statistischen Datenanalyse stets erneut zu prüfen ist, ob die Voraussetzungen ihrer Anwendung auch gegeben sind.

Die Lehrveranstaltung will diese konzeptionelle Vorgehensweise computergestützter Statistik vermitteln. Sie ist kein Kurs zur Vermittlung der ausgewählten Software. Das eingesetzte Software-Paket soll möglichst im Hintergrund bleiben, d.h., es soll wirklich als ein technisches Hilfsmittel begriffen werden, das jederzeit gegen ein anderes ausgetauscht werden kann.

Im Mittelpunkt der Lehrveranstaltung „Computergestützte Statistik I“ stehen folgende drei Problemkreise:

- Entdeckung und Identifizierung von möglichen Ausreißern

Im Datenmaterial enthaltene Ausreißer (atypische Variablenwerte) können zu einer tendenziellen Verzerrung der Ergebnisse statistischer Methoden führen. Für die Entdeckung und Identifizierung von möglichen Ausreißern werden vor allem explorative graphische Verfahren (Stem-and-Leaf-Plot, Boxplot) sowie eine Auswertung der empirischen Häufigkeitsverteilung und Ausreißertests herangezogen.

## 1. Einführung

- Formulierung und Überprüfung von Verteilungshypothesen für bestimmte Variablen  
Um entsprechend dem Skalenniveau der Variablen eine überprüfbare statistische Hypothese formulieren zu können, werden auch hierfür explorative graphische Mittel (u.a. Histogramm, Wahrscheinlichkeitsplot) eingesetzt sowie deskriptive statistische Maßzahlen (Mittelwerte, Schiefe, Kurtosis) berechnet. Die Prüfung der Verteilungshypothesen erfolgt mittels statistischer Tests (Kolmogorov-Smirnov-Test, Chi-Quadrat-Anpassungstest, Binomial-Test). Bei metrisch skalierten Variablen ist vor allem die Prüfung auf Normalverteilung von Bedeutung. Mittels geeigneter Transformationen lässt sich u.U. die Verteilung der Variablen einer symmetrischen Verteilung annähern.
- Parametervergleiche bei unabhängigen Stichproben  
Oftmals können Variable nach einer zweiten Variablen (Faktorvariablen) gruppiert werden, z.B. das Haushaltsnettoeinkommen nach der Haushaltsgröße oder persönliches Einkommen nach dem Geschlecht. Zur Feststellung von Unterschieden zwischen den Gruppen bezüglich der betrachteten Variablen werden neben der graphischen Veranschaulichung und der deskriptiven Auswertung von statistischen Maßzahlen verschiedene Tests zur Prüfung von Unterschieden in den Verteilungen und deren Parameter (z.B. F-Test, t-Test, einfache Varianzanalyse, multiple Mittelwertvergleiche) behandelt.

Diese Problemkreise werden an Beispielen ausführlich diskutiert und mit SPSS demonstriert. Während die Problemkreise 1 und 2 als notwendig für jede statistische Analyse angesehen werden, stellt der dritte Problemkreis eine Auswahl aus der Vielzahl der möglichen statistischen Analysen dar.

Weitere Analysen mittels SPSS werden in der Lehrveranstaltung „Computergestützte Statistik II“ vorgestellt:

- Überprüfung von Zusammenhängen zwischen Variablen,
- Feststellung der Abhängigkeit von Variablen (Regressionsanalyse),
- Reliabilitäts- und Homogenitätsanalyse von Konstrukten.

Grundlagen der Handhabung von statistischen Software-Paketen und damit auch von SPSS werden im weiteren vorausgesetzt. An dieser Stelle sollen jedoch einige Hinweise zur Datenmodifikation und Datenselektion unter SPSS gegeben werden, die oftmals im Verlaufe einer statistischen Analyse erforderlich sind.

Im weiteren wird davon ausgegangen, dass

- der Untersuchungsgegenstand fachwissenschaftlich begründet formuliert wurde,
- die Definition der statistischen Elemente und der Variablen erfolgte,

- die statistischen Daten vorliegen, d.h. durch eine primärstatistische Erhebung gewonnen oder als Sekundärstatistik aus Veröffentlichungen entnommen wurden, und in einer SPSS-Datei (mit der Erweiterung .sav) gespeichert sind.

Für die formale Darstellung wird vereinbart, dass Beobachtungen an n statistischen Elementen ( $i = 1, \dots, n$ ) und für m Variablen ( $j = 1, \dots, m$ ) gegeben sind.

Unter SPSS werden die Elemente als Fälle, Variablen mit zahlenmäßigen Beobachtungswerten als numerische Variable, alle anderen Variablen als Zeichenketten-Variablen (String-Variablen) bezeichnet. Von besonderer Bedeutung ist, dass man sich als Nutzer der Daten über das Skalenniveau<sup>1</sup> der statistischen Variablen im klaren ist, da davon ganz entscheidend die Auswahl der statistischen Prozeduren abhängt.

SPSS bietet im Data Editor ein spreadsheet-ähnliches Arbeitsmittel zur effizienten Eingabe, Darstellung und Editierung von Daten. Dieses ist in Zeilen und Spalten aufgeteilt. Die Zeilen entsprechen den Fällen, die Spalten den Variablen. Die Zellen nehmen die Beobachtungswerte auf. Fehlende Beobachtungswerte heißen Missing-Werte (Missings). Da in dem Spreadsheet in den Bereichsgrenzen der Datendatei keine Zelle leer bleibt, werden sie im Falle fehlender Werte durch systemdefinierte Missing-Werte (system missings) belegt. Ein System-Missing ist im Falle numerischer Variablen ein Punkt (d.h. in der leer gelassenen Zelle erscheint ein Punkt), im Falle von String-Variablen Leerzeichen. Der Nutzer kann jedoch selbst die Werte der Missings festlegen (benutzerdefinierte Missings), die dann wie alle anderen Werte einzugeben sind.

## Fälle sortieren

Oftmals wünscht man sich eine andere Sortierung der Fälle, als sie durch die Eingabe realisiert ist, bzw. es ist eine andere Reihenfolge notwendig. Durch Anwahl von

### ■ Data

#### ■ Sort Cases...

öffnet sich das folgende Dialogfeld.

Abbildung 1.1.: Dialogfeld „Sort Cases“



Aus der Variablenliste kann eine Variable oder können mehrere Variablen in das Textfeld „Sort

<sup>1</sup>Vgl. u.a. Röenz, B., Strohe, H.G. (Hrsg.) (1994), S. 241 f., S. 328 ff.

## 1. Einführung

by:“ gebracht werden. Für jede dieser Variablen kann die Sortierung entweder als Aufsteigend (Ascending) oder Absteigend (Descending) erfolgen. Werden mehrere Variablen zur Sortierung ausgewählt, so erfolgt die Sortierung in der Reihenfolge, wie die Variablen in der Sortierungsliste genannt sind. Das heißt, es erfolgt zunächst die Sortierung nach der ersten Variablen, dann wird für die gleichen Werte der ersten Variablen nach der zweiten Variablen sortiert usw. Dabei kann durchaus eine Variable aufsteigend, die andere absteigend sortieren. Durch Betätigen von „OK“ wird der Sortierungsprozeß in Gang gesetzt.

### Unterteilung der Fälle nach Gruppen (Split File)

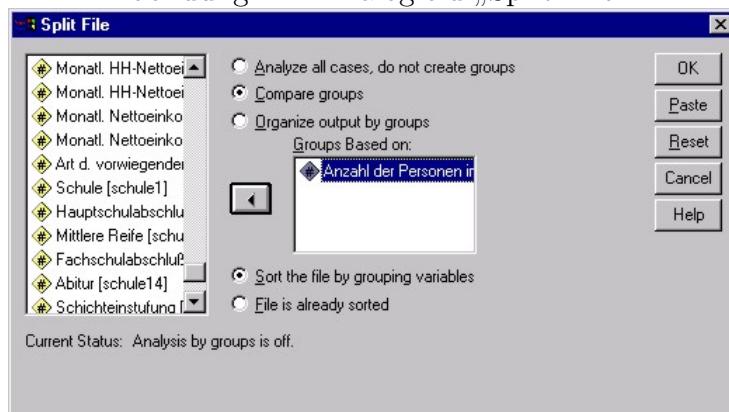
Statistische Analysen werden häufig getrennt für Gruppen von Fällen durchgeführt. SPSS bietet eine einfache Möglichkeit zur Schaffung von Gruppen. Dazu ist

#### ■ Data

##### ■ Split File...

zu wählen. Es erscheint nachstehendes Dialogfeld.

Abbildung 1.2.: Dialogfeld „Split File“



Aus der Variablenliste auf der linken Seite wird die Faktor-Variablen bzw. werden die Faktor-Variablen (maximal 8) in das Feld „Groups Based on:“ gebracht. Die Gruppierung erfolgt nach den Ausprägungen dieser Variable(n). Die Faktor-Variablen sind i.a. nominal- oder ordinalskalierte Variablen bzw. metrische Variablen, die in Klassen eingeteilt sind. Man sollte jedoch darauf achten, dass nicht zu wenig Fälle je Gruppe auftreten.

Für die Präsentation der Ergebnisse im SPSS for Windows Viewer gibt es zwei Möglichkeiten:

- Compare groups

In diesem Fall werden die Ergebnisse einer SPSS-Prozedur in einer Tabelle unterteilt nach den Ausprägungen der Gruppierungsvariablen ausgegeben. Bei Grafiken wird je Ausprägung der Gruppierungsvariablen eine separate Grafik erzeugt.

- Organize output by groups

In diesem Fall wird für jede Ausprägung der Gruppierungsvariablen eine separate Ergebnistabelle im SPSS for Windows Viewer erstellt.

Die Verarbeitung nach Gruppen durch andere SPSS-Prozeduren erfordert eine nach dieser (diesen) Variablen sortierte Datei. Die Voreinstellung von „Sort the file by grouping variables“ sollte deshalb belassen werden. Nach Klicken auf „OK“ wird die Sortierung vorgenommen. Als Kennzeichnung dafür, dass die Datei zur Analyse in Gruppen aufgeteilt ist, steht im Data Editor in der Statuszeile unten rechts „Split File On“.

Wenn die Statistik-Prozeduren wieder für alle Fälle gemeinsam durchgeführt werden sollen, ruft man das Dialogfeld „Split File“ nochmals auf und klickt auf „Analyze all cases, do not create groups“ und auf „OK“. Der Vermerk „Split File On“ in der Statuszeile verschwindet.

### Klassierung einer metrisch skalierten Variablen

Die Klassierung einer metrisch skalierten Variablen erfordert zunächst, dass der Wertebereich einer jeden Gruppe festgelegt wird. Es ist dafür eine neue Variable zu schaffen, die für jeden Fall die Klassen-Nummer aufnimmt, zu der der Wert der Variablen bei dem betrachteten Fall gehört. Dieser Vorgang ist also eine Umkodierung der Ausgangsvariablen.

Unter SPSS kann eine solche Klassierung vorgenommen werden, indem

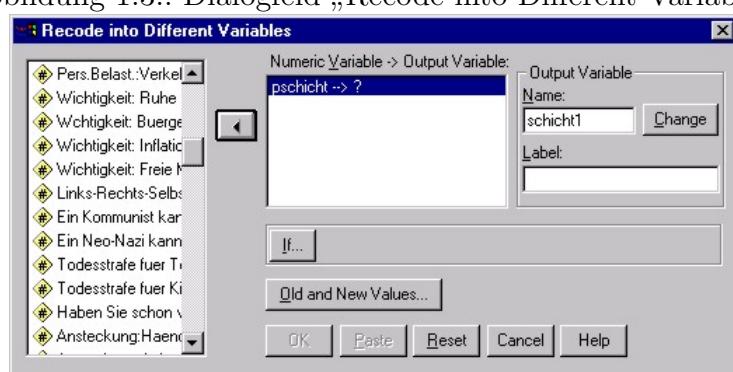
#### ■ Transform

##### ■ Recode

###### ■ Into Different Variables...

aufgerufen wird. Die Option „Recode Into Same Variables...“ sollte vermieden werden, um die Ausgangsvariable zu erhalten.

Abbildung 1.3.: Dialogfeld „Recode into Different Variables“



Aus der linken Variablenliste ist die Variable, die umkodiert werden soll, in das Feld „Input Variable → Output Variable:“ zu bringen. Das Fragezeichen ist durch die Eingabe eines Namens

## 1. Einführung

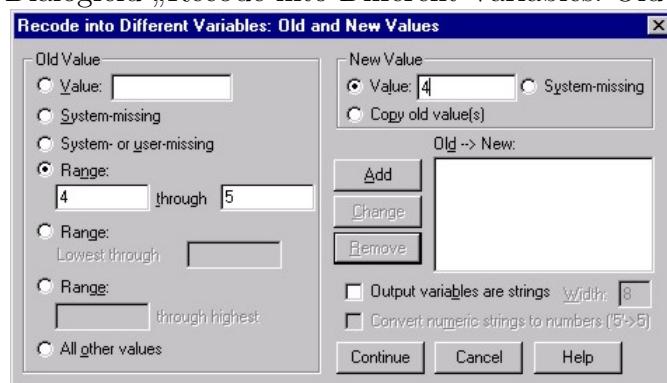
für die Output Variable in der rechten Textbox und Klick auf „Change“ zu ersetzen. Im Feld „Label“ kann noch eine Bezeichnung für die Variable eingegeben werden. Nun ist noch zu vereinbaren, wie die alten Werte in die neuen zu überführen sind. Durch Anklicken von „Old and new Values...“ öffnet sich das Dialogfeld „Recode into Different Variables: Old and new Values“ (siehe Abb. 1.4).

Bei der Umkodierung kann

- für einen alten Wert (Value)
- für einen Bereich von ... bis ... der alten Werte (Range ... through ...)
- vom niedrigsten bis ... alten Wert (Range Lowest through ...)
- von ... bis zum höchsten alten Wert (Range ... through highest)
- für alle anderen alten Werte (All other values)

ein neuer Wert (d.h. eine Klassen-Nummer) gesetzt werden.

Abbildung 1.4.: Dialogfeld „Recode into Different Variables: Old and new Values“



Diese Eingaben sind auf der linken Seite in dem Textfeld „Old Value“ nach Anklicken der gewünschten Möglichkeit zu machen. Anschließend ist in dem Textfeld „New Value“ entweder

- ein neuer Wert einzugeben (Value) oder
- anzuseigen, dass der alte Wert(ebereich) übernommen werden soll (Copy old value(s)), oder
- der alte Wert(ebereich) soll zu einem System-Missing werden.

Durch Klicken auf „Add“ werden diese Entscheidungen in das Feld „Old → New:“ gebracht. Dieser Vorgang wird solange wiederholt, bis alle Gruppen eingegeben sind. Des weiteren kann

auch für ein altes System-Missing bzw. ein altes benutzerdefiniertes Missing (System- oder user-missing) ein neuer Wert vereinbart werden.

Über die Schaltfläche „Continue“ kehrt man zurück in das vorangegangene Dialogfeld, in dem durch Klick auf „OK“ die Umkodierung ausgelöst und der Datei eine neue Variable hinzugefügt wird. Es erweist sich stets als günstig für die neue Variable durch Doppelklick auf den Variablenamen Label für die Klassen einzutragen.

## Fälle auswählen

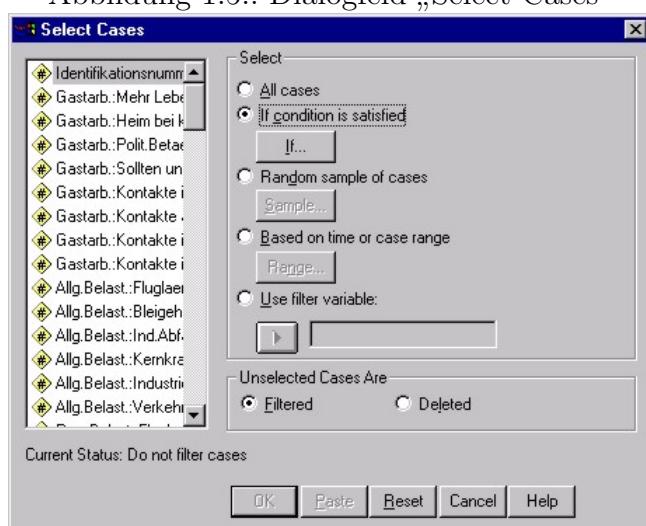
SPSS bietet ein breites Spektrum von Möglichkeiten, aus einer Datei Fälle auszuwählen, worauf hier nur kurz eingegangen werden soll. Durch Anwahl von

### ■ Data

#### ■ Select Cases...

öffnet sich das dazugehörige Dialogfeld „Select Cases“, in dem es vier verschiedene Möglichkeiten gibt, die Auswahl der Fälle festzulegen.

Abbildung 1.5.: Dialogfeld „Select Cases“

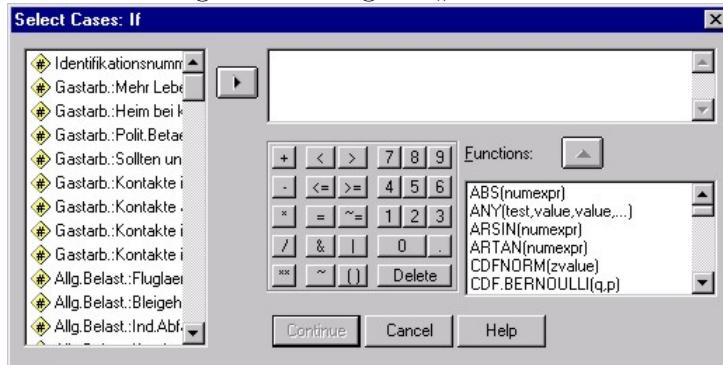


#### 1. Auswahl, falls Bedingung zutrifft (If condition is satisfied):

Durch Anwahl dieser Auswahlmöglichkeiten und Klicken auf die Schaltfläche „If...“ öffnet sich ein weiteres Dialogfeld.

## 1. Einführung

Abbildung 1.6.: Dialogfeld „Select Cases: If“



In dem freien Textfeld können die Bedingungen für die Auswahl eingetragen werden, wobei Variablen, Funktionen, arithmetische Operationen, Konstanten, logische und relationale Operatoren verwendet werden können, die entweder in das Bedingungsfeld eingetippt oder mittels der Maus aus der Variablenliste und der Funktionsliste durch Doppelklick und aus der Rechnertastatur (Taschenrechner) durch einfaches Anklicken in das Bedingungsfeld gebracht werden. Zur ausführlichen Beschreibung der Erstellung solcher Bedingungen sei auf das SPSS-Handbuch verwiesen. Das Bedingungsfeld kann wie eine Textdatei editiert werden.

Grundsätzlich gilt, dass nur diejenigen Fälle ausgewählt werden, für die die Bedingung erfüllt (wahr) ist. Ist für einen Fall die Bedingung nicht erfüllt (false) oder missing, dann wird der Fall nicht ausgewählt.

Durch Betätigen der Schaltfläche „Continue“ gelangt man in das vorherige Dialogfeld zurück.

## 2. Zufallsstichprobe (Random sample of cases):

Durch Anwahl dieser Auswahlmöglichkeiten und Klicken auf die Schaltfläche „Sample...“ öffnet sich ein weiteres Dialogfeld.

Abbildung 1.7.: Dialogfeld „Select Cases: Random Sample“



In diesem Dialogfeld kann entweder

- ein (etwa) Prozentsatz von auszuwählenden Fällen (Approximately ... % of all cases)

oder

- eine genaue Anzahl aus den ersten ... Fällen (Exactly ... cases from the first ... cases)

angegeben werden, der/die durch die Zufallsstichprobe zu realisieren ist. Im zweiten Fall ist auch die Anzahl der Fälle anzugeben, aus denen die Zufallsstichprobe gezogen werden soll. Diese Zahl muß kleiner oder gleich der Gesamtzahl der Fälle in der Datei und größer als die davorstehende Anzahl sein.

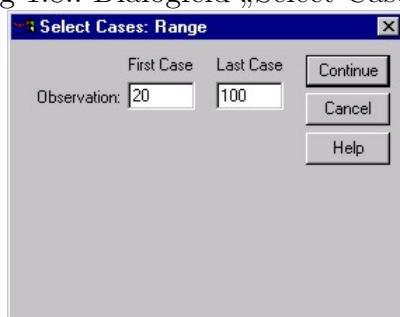
Da SPSS für jede Zufallsstichprobe einen anderen Startwert verwendet, wird jedesmal eine andere Stichprobe realisiert. Allerdings kann der Startwert auch beeinflußt werden (siehe SPSS-Handbuch).

Durch Betätigen der Schaltfläche „Continue“ gelangt man in das vorherige Dialogfeld zurück.

### 3. Auswahl nach Zeit- oder Fallbereich (Based on time or case range):

Durch Anwahl dieser Auswahlmöglichkeit und Klicken auf die Schaltfläche „Range...“ öffnet sich ein weiteres Dialogfeld.

Abbildung 1.8.: Dialogfeld „Select Cases: Range“



Hier kann der Bereich der auszuwählenden Fälle festgelegt werden. Wird nur ein „Last Case“ eingetragen, so heißt das: alle Fälle vom ersten bis zum letzten Fall. Durch Betätigen der Schaltfläche „Continue“ gelangt man in das vorherige Dialogfeld zurück.

### 4. Auswahl unter Verwendung einer Filtervariablen (Use filter variable):

Bei dieser Auswahlmöglichkeit ist eine numerische Variable aus der Variablenliste anzugeben. Alle Fälle, bei denen diese Variable einen Wert von Null oder Missing hat, werden nicht ausgewählt.

Die letzte Entscheidung im Dialogfeld „Select Cases“ betrifft die Behandlung von nicht ausgewählten Fällen. Sie können gefiltert (Voreinstellung) oder gelöscht werden. Filtern (Filtered) bedeutet, dass nicht ausgewählte Fälle in der Datei verbleiben, aber bei den Analysen nicht

## 1. Einführung

berücksichtigt werden. Der Datei wird eine neue Variable filter\_\\$ hinzugefügt, die den Wert 1 für alle ausgewählten Fälle aufweist, ansonsten 0 ist. Löschen (Deleted) bedeutet, daß die nicht ausgewählten Fälle aus der Datei entfernt werden. Bei großen Dateien wird dadurch Rechenzeit eingespart. Man sollte jedoch dann diese reduzierte Datei nicht oder auf einen anderen Dateinamen abspeichern.

Durch Betätigen von „OK“ gelangt man zurück zum Data Editor. Wurde „Filtered“ gewählt, sind die Nummern der nicht ausgewählten Fälle am linken Rand durchgestrichen und in der Statuszeile ist „Filter On“ vermerkt.

Wenn eine Analyse wieder mit allen Fällen durchgeführt werden soll und wurde auf „Filtered“ entschieden, kann die Auswahl von Fällen rückgängig gemacht werden, indem

### ■ Data

#### ■ Select Cases...

und im Dialogfeld „All cases“ gewählt wird. In der Statuszeile verschwindet die Ausschrift „Filter On“. Wurde auf „Deleted“ entschieden, so muß die Datei neu geladen werden.

## Fälle gewichten

Oftmals liegen bei einer statistischen Untersuchung für eine numerische Variable nicht die Ausgangswerte, sondern die Häufigkeitsverteilung mit absoluten oder relativen Häufigkeiten vor. Dann ist neben der Variablen mit den verschieden aufgetretenen Beobachtungswerten eine Variable mit den Häufigkeiten in die Datei aufzunehmen. Vor der Durchführung statistischer Prozeduren ist demzufolge eine Gewichtung unter Verwendung der Häufigkeitsvariablen durchzuführen.

Ebenso kann es vorkommen, dass bestimmten Werten numerischer Variablen ein besonderes Gewicht beigemessen wird und diese Gewichte in einer Gewichtsvariablen in der Datei enthalten sind.

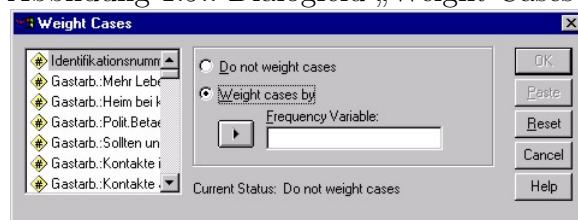
Um die Gewichtung zu aktivieren, wird

### ■ Data

#### ■ Weight Cases...

aufgerufen.

Abbildung 1.9.: Dialogfeld „Weight Cases“



In diesem Dialogfeld ist „Weight cases by“ anzuklicken und dann diejenige Variable, die die gewünschten Gewichte enthält, in das Feld „Frequency Variable:“ zu bringen. Fälle, für die die Häufigkeitsvariable Null oder Missing enthält, werden von der weiteren Analyse ausgeschlossen. Nach Betätigen der Schaltfläche „OK“ und Rückkehr zum Data Editor ist in der Statuszeile „Weight On“ vermerkt.

Die Gewichtung wird wieder deaktiviert, indem in dem Dialogfeld der Abb. 1.9 auf „Do not weight cases“ entschieden wird.

### Berechnung von neuen Variablen

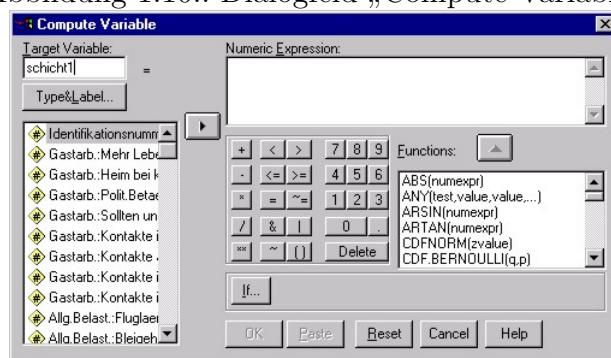
Unter SPSS können die Werte einer neuen Variablen durch numerische Transformationen aus den Werten anderer Variablen einer Datei berechnet werden. Dazu ist

#### ■ Transform

##### ■ Compute...

zu wählen. Das Dialogfeld „Compute Variable“ wird angezeigt.

Abbildung 1.10.: Dialogfeld „Compute Variable“



Zuerst ist ein Name für die neue Variable (Zielvariable) im Feld „Target Variable:“ festzulegen (es wird in der Regel ein neuer Variablenname sein, um keine bereits existierende Variable zu überschreiben). Für diese Zielvariable wird der Typ und ein Label vergeben, indem die Schaltfläche „Type & Label...“ geklickt wird. Das Dialogfeld der Abb. 1.11 ermöglicht die Eingabe.

Abbildung 1.11.: Dialogfeld „Compute Variable: Type and Label“



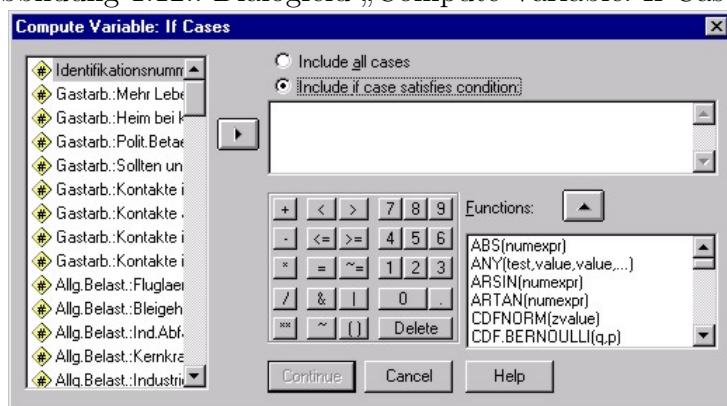
## 1. Einführung

Über die Schaltfläche „Continue“ kehrt man in das vorherige Dialogfeld zurück. Hier wird jetzt im Textfeld „Numeric Expression:“ die Berechnungsvorschrift für die neue Variable eingegeben (zu Einzelheiten siehe SPSS-Handbuch). Dazu stehen zur Verfügung:

- die Liste der bereits existierenden Variablen der Datei in dem linken Feld,
- arithmetische, relationale und logische Operatoren sowie Konstanten über den 'Taschenrechner',
- Funktionen aus dem gleichnamigen Menü „Functions:“; diese unterteilen sich u.a. in arithmetische Funktionen, statistische Funktionen, Verteilungsfunktionen, logische Funktionen, Datums- und Zeitfunktionen, Missing-Werte-Funktionen, String-Funktionen.

Diese Berechnung kann zum einen für alle Fälle der Datei durchgeführt werden. Andererseits kann aber auch eine bedingte Auswahl von Fällen erfolgen. Dazu ist die Schaltfläche „If...“ zu betätigen. Es öffnet sich das Dialogfeld „Compute Variable: If Cases“, in dem „Include if case satisfies condition:“ anzuklicken ist.

Abbildung 1.12.: Dialogfeld „Compute Variable: If Cases“



Für die Formulierung der Bedingung stehen die gleichen Hilfsmittel wie vorher zur Verfügung. Über die Schaltfläche „Continue“ kehrt man in das vorherige Dialogfeld zurück. Bei komplexen Berechnungen sollten die getroffenen Entscheidungen über die Schaltfläche „Paste“ in den Syntax Editor gebracht und abgespeichert werden, um sie gegebenenfalls schnell und einfach reproduzieren zu können. Zur Abarbeitung der Befehle ist der gesamte Text im Syntax Editor zu markieren und auf die Schaltfläche (Run Current) zu klicken.

## Vergabe von Rängen oder Schlüsselnummern

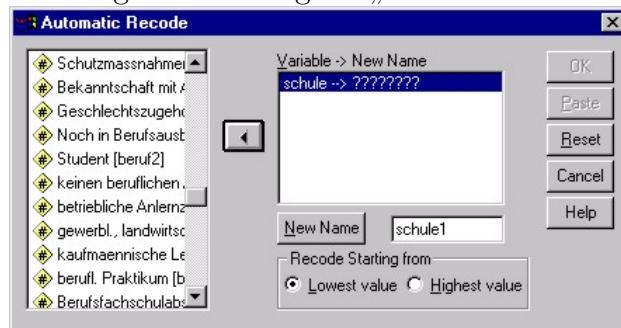
Über

■ Transform

## ■ Automatic Recode...

können metrisch skalierte in ordinalskalierte Variablen überführt oder nominalskalierten Variablen Schlüsselnummern zugewiesen werden.

Abbildung 1.13.: Dialogfeld „Automatic Recode“



Zunächst ist die umzukodierende Variable aus der Liste in das Feld „Variable → New Name“ zu bringen. Der neue Name wird in das Textfeld neben der Schaltfläche „New Name“ einge-tipt und anschließend diese Schaltfläche betätigt. Der neue Name ersetzt die Fragezeichen im oberen Feld. Dieser Vorgang kann für beliebig viele Variablen wiederholt werden. Des weiteren kann die Umkodierung beim kleinsten Wert (Lowest value) oder beim größten Wert (Highest value) beginnen. Allerdings gilt dies dann für alle im oberen Feld enthaltenen Variablen. Nach Betätigung der Schaltfläche „OK“ wird im SPSS for Windows Viewer die Umkodierung für jede Variable angezeigt und im Data Editor wurden die neuen Variablen der Datei hinzugefügt.

Für numerische Variablen kann eine Umkodierung, jedoch mit differenzierteren Möglichkeiten, erfolgen (zu Einzelheiten siehe SPSS-Handbuch).

Möglichkeiten für die Modifizierung von Zeitreihenwerten werden über „Transform“ und „Create Time Series...“ bzw. „Replace Missing-Values...“ durch SPSS angeboten, worauf an dieser Stellen jedoch nicht eingegangen werden kann.

## *1. Einführung*

## 2. Entdeckung und Identifikation von Ausreißern

Ausreißer<sup>2</sup> sind extreme Beobachtungswerte in einer statistischen Reihe, die qualitativ von der Gesamtheit abweichende statistische Elemente signalisieren. Die Frage nach Ausreißern im Datenmaterial ist nur sinnvoll bei metrisch skalierten Variablen. Die Statistik kennt kein exaktes Verfahren, um festzustellen, ob ein für die Beobachtungsreihe bzw. Stichprobe atypischer Wert ein Ausreißer ist oder nicht. Solche atypischen Werte können durch Erfassungsfehler, Eingabefehler oder Rechenfehler verursacht worden sein. Es können aber auch seltene Werte sein, d.h. Werte, die aufgrund der Verteilung der Grundgesamtheit mit einer sehr kleinen Wahrscheinlichkeit auftreten. Was die Statistik jedoch tun kann und muß, ist die Entdeckung und Identifizierung von Ausreißern mit allen ihr zur Verfügung stehenden Mitteln und eine Ausreißerbehandlung vorzunehmen, da solche Ausreißer zu schwerwiegenden Verzerrungen der Ergebnisse statistischer Methoden und ihrer Interpretation führen können. Diese Phase der Datenkontrolle darf also in keinem Falle vernachlässigt werden. Um potentielle Ausreißer zu entdecken und zu identifizieren, werden in einem ersten Schritt, vor allem bei Variablen mit vielen Beobachtungswerten, graphische bzw. pseudo-graphische Verfahren der explorativen Datenanalyse herangezogen.

Bei univariaten eindimensionalen Datenanalysen sind dies Stem-and-Leaf Plot und Boxplot.

### 2.1. Stem-and-Leaf Plot

Ein Stem-and-Leaf Plot ist eine halbgraphische Darstellung der Werte einer Beobachtungsreihe eines metrisch skalierten Merkmals, für das es verschiedene Modifikationen gibt. SPSS bietet den Stem-and-Leaf Plot unter

- Analyze
  - Descriptive Statistics

---

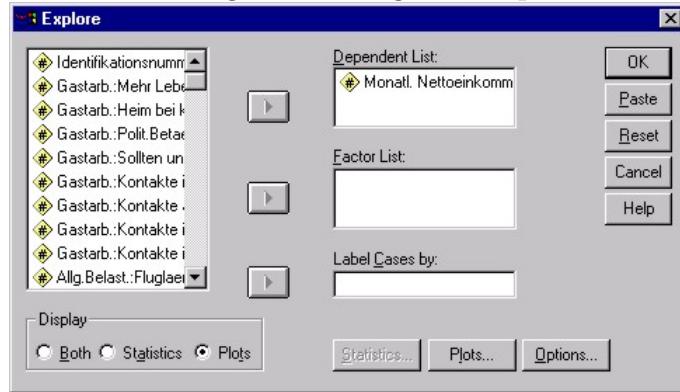
<sup>2</sup>Vgl. u.a. Rönnz, B., Strohe, H.G. (Hrsg.)(1994), S. 28 f.

## 2. Entdeckung und Identifikation von Ausreißern

### ■ Explore...

an. Es öffnet sich das Dialogfeld „Explore“

Abbildung 2.1.: Dialogfeld „Explore“



Die gewünschte Variable wird aus der linken Variablenliste durch Klick auf diese Variable (Markieren) und auf die Schaltfläche in das Feld „Dependent List:“ gebracht. Dies kann für andere Variablen wiederholt werden. Für jede Variable wird ein Stem-and-Leaf Plot erzeugt. Da zunächst nur die Grafik interessiert, wird im Feld „Display“ der Kreis vor Plots markiert und anschließend die Schaltfläche „Plots...“ aktiviert. Es öffnet sich ein weiteres Dialogfeld.

Abbildung 2.2.: Dialogfeld „Explore: Plots“



Im Feld „Descriptive“ ist der Stem-and-Leaf Plot voreingestellt. Im Feld „Boxplots“ wird der Kreis vor None markiert. Über die Schaltfläche „Continue“ kommt man zurück in das vorangegangene Dialogfeld und nach „OK“ erscheint im SPSS for Windows Viewer der gewünschte Plot.

Zunächst wird in einer Tabelle eine Übersicht über die Anzahl (N) und den Prozentsatz (Percent) der gültigen Fälle (valid), der Missings und der Fälle insgesamt (total) gegeben.

Unter einer Vorspalte zum Stem-and-Leaf Plot steht unter der Überschrift Frequency die absolute Häufigkeit der Fälle des Stammes. SPSS wählt intern eine Stamm-Einheit (stem width),

die unterhalb des Diagramms angegeben wird. Zum Beispiel beinhalten bei einer stem width von 10 der Stamm die Zehner-Ziffern, die Blätter die Einer-Ziffern. Die Dezimalstellen werden vernachlässigt. So hat beispielsweise der Beobachtungswert 34,5 den Stamm 3 und das Blatt 4. Analog erhält man bei einer stem width von 100 als Stamm die Hunderter-Ziffern, als Blätter die Zehnerziffern und alle weiteren Ziffern werden vernachlässigt (Zum Beispiel hat 186 den Stamm 1 und das Blatt 8).

Die Darstellung des Stem-and-Leaf Plots variiert je nach der Anzahl der Stamm-Ziffern und der Anzahl der Blatt-Ziffern. Bei der kleineren Version des Stem-and-Leaf Plots wird jeder Stamm auf zwei Zeilen aufgeteilt. Die erste Zeile nimmt die Blätter von 0 bis 4 auf, die zweite Zeile die Blätter von 5 bis 9. Bei der ausgedehnteren Version des Stem-and-Leaf Plots wird jeder Stamm auf 5 Zeilen aufgeteilt. Die erste Zeile nimmt die Blätter 0 und 1 auf, die zweite Zeile die Blätter 2 und 3, die dritte Zeile die Blätter 4 und 5, die vierte Zeile die Blätter 6 und 7 und die fünfte Zeile die Blätter 8 und 9. Die letzte Angabe unterhalb des Diagramms, bezeichnet mit each leaf, gibt an, wieviele Fälle durch eine Blatt-Ziffer repräsentiert werden.

Wenn „extrem“ kleine und/oder „extrem“ große Werte auftreten, so wird das in einer ersten und/oder letzten Zeile des Stem-and-Leaf Plots durch die Ausschrift Extremes gekennzeichnet. Diese Werte werden dann in Klammern angegeben.

- Beispiel 2.1:

In der Datei europa.sav, die von der bei Bühl, A., Zöfel, P. (1994) beiliegenden Diskette entnommen wurde, stehen Daten für 28 europäische Länder zur Verfügung. Für die Variable Kindersterblichkeit (bei 1000 Geburten) soll ein Stem-and-Leaf Plot erstellt werden. Hat man die Entscheidungen in der oben dargestellten Weise vollzogen, erhält man nachstehenden Output (siehe SPSS-Output 2.1).

Da die Stamm-Einheit 10 ist, beinhalten die Blatt-Ziffern die Einer-Ziffern und die Dezimalstellen entfallen. Zum Beispiel ist die erste Zeile des Stem-and-Leaf Plots wie folgt zu lesen: Es gibt ein Land mit einer Kindersterblichkeit von 7 bis unter 8 bei 1000 Geburten und ein Land mit einer Kindersterblichkeit von 8 bis unter 9 bei 1000 Geburten. Aus dieser Interpretation wird deutlich, dass mit dem Stem-and-Leaf Plot eine Klassenbildung der Beobachtungswerte einhergeht. Es wird kein extremer Wert im Stem-and-Leaf Plot angezeigt.

## 2. Entdeckung und Identifikation von Ausreißern

### SPSS-Output 2.1: Kleinere Version des Stem-and-Leaf Plots

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Kindersterblichkeit bei 1000 Geburten	28	100,0%	0	,0%	28	100,0%

### Kindersterblichkeit bei 1000 Geburten Stem-and-Leaf Plot

Frequency Stem & Leaf

2,00	0	.	78
10,00	1	.	0000113444
5,00	1	.	56679
3,00	2	.	034
5,00	2	.	66779
1,00	3	.	1
2,00	3	.	68

Stem width: 10,0

Each leaf: 1 case(s)

- Beispiel 2.2:

Es wird der Datensatz allbus.sav<sup>3</sup> zugrunde gelegt. Er beinhaltet Ergebnisse der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (ALLBUS). Für die Variable einkomp1 (monatliches persönliches Nettoeinkommen in DM) soll ein Stem-and-Leaf Plot erstellt werden (siehe SPSS-Output 2.2-1).

Da die Stamm-Einheit (stem width) 1000 ist, sind die Blatt-Ziffern die Hunderter. & als Blatt beinhaltet eine restliche Anzahl von Fällen. Zum Beispiel sind beim Stamm von 4 in der zweiten Zeile vier Fälle (befragte Personen) registriert. Davon haben zwei Personen (da jedes Blatt 2 Fälle angibt) ein monatliches Nettoeinkommen von 4200 bis unter 4300 DM bei der Befragung angegeben. Von den restlichen 2 Personen hat eine ein Nettoeinkommen von 4200 bis unter 4300 DM und die andere von 4300 bis unter 4400 DM. Es kann also weder ein Blatt mit der Ziffer 2, noch ein Blatt mit der Ziffer 3 angegeben werden, da jedes Blatt zwei Fälle zu repräsentieren hat. Dies wird durch & gekennzeichnet.

Dagegen kann beispielsweise in der 5. Zeile des Stammes 3 kein & (fractional leaf) angegeben werden, obwohl dafür 5 Fälle registriert sind, von denen 4 Personen (da jedes Blatt 2 Fälle bedeutet) ein monatliches Nettoeinkommen von 3800 bis unter 3900 DM bei der Befragung angegeben haben. Es verbleibt somit nur ein Fall. Bei Angabe eines fractional leaf würde nicht klar sein, ob dazu ein Blatt von 8 oder ein Blatt von 9 gehören würde.

<sup>3</sup>Die Datei allbus.sav wurde entnommen aus: Wittenberg, R. (1991)

**SPSS-Output 2.2-1:** Ausgedehntere Version des Stem-and-Leaf Plots

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Monatliches Nettoeinkommen in DM	716	23,5%	2336	76,5%	3052	100,0%

## Monatl. Nettoeinkommen in DM Stem-and-Leaf Plot

Frequency      Stem    &amp;   Leaf

2,00	0	.	1
21,00	0	.	2233333333
35,00	0	.	4444444555555555
47,00	0	.	666666666666666677777777
41,00	0	.	88888888888899999999
45,00	1	.	0000000000000000111111
38,00	1	.	222222222222233333
63,00	1	.	4444444444555555555555555555
45,00	1	.	6666666666667777777777
72,00	1	.	88888888888888888888889999999999
78,00	2	.	000000000000000000000000000000000000001111111
46,00	2	.	222222222222233333333333
32,00	2	.	44455555555555
28,00	2	.	666666677777777
23,00	2	.	888888899999
28,00	3	.	00000000000011
10,00	3	.	2233
16,00	3	.	4455555
8,00	3	.	6677
5,00	3	.	88
12,00	4	.	00000&
4,00	4	.	2&
17,00	Extremes		(>= 4400)

Stem width: 1000

Each leaf: 2 case(s)

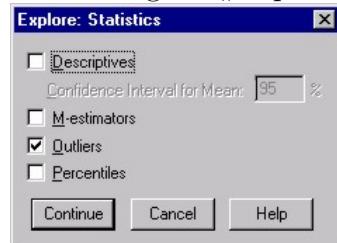
&amp; denotes fractional leaves.

Insgesamt werden 17 extreme Werte angezeigt, von denen nicht alle als potentielle Ausreißer zu klassifizieren sind. Weitere Prüfungen sind hier notwendig.

Zusätzlich zum Stem-and-Leaf Plot kann man sich extreme Beobachtungswerte ausgeben lassen, und zwar die 5 größten und die 5 kleinsten Werte. Obwohl die Fallnummer (Case Number) der extremen Werte ausgegeben wird, kann zusätzlich im Dialogfeld „Explore“ (siehe Abb. 2.1) diejenige Variable in das Feld „Label Cases by:“ gebracht werden, die zur Kennzeichnung der Fälle dient, um diese extremen Werte besser identifizieren zu können. Weiterhin ist im Feld „Display“ Statistics zu aktivieren und die Schaltfläche „Statistics...“ zu betätigen.

## 2. Entdeckung und Identifikation von Ausreißern

Abbildung 2.3.: Dialogfeld „Explore: Statistics“



In diesem Dialogfeld wird nur auf Outliers (Ausreißer) entschieden.

- Beispiel 2.2 (Fortsetzung):

Für die Datei allbus.sav wird die Variable id (Identifikationsnummer, Fragebogennummer) zur Identifikation verwendet. Die 5 größten und die 5 kleinsten Werte für die Variable einkomp1 sind im nachfolgenden Output enthalten.

### SPSS-Output 2.2-2: Extreme Werte

Extreme Values					
		Case Number	Identifikationsnummer der Befragten	Value	
Monatl. Nettoeinkommen in DM	Highest	1	245	247	15000
		2	882	886	7000
		3	1090	1094	6500
		4	1048	1052	6400
		5	1058	1062	5900
	Lowest	1	866	870	120
		2	214	216	150
		3	1611	1625	200
		4	1802	1818	200
		5	992	996	206

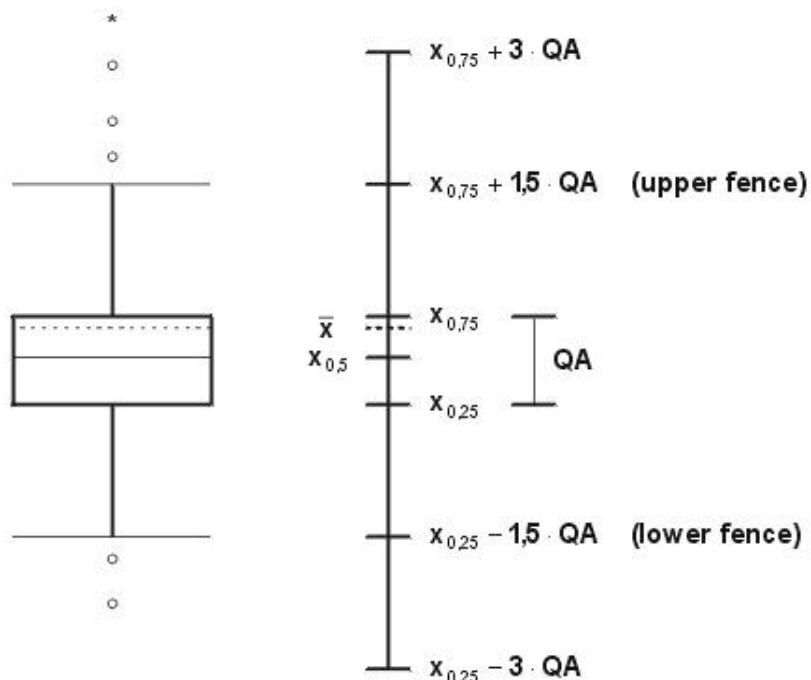
## 2.2. Boxplot

Im Gegensatz zum Stem-and-Leaf Plot enthält der Boxplot<sup>4</sup> (Box-Whisker-Plot, Schachtelzeichnung) nicht mehr Kennzeichnungen für Einzelwerte, sondern bereits summarische Kenngrößen der Häufigkeitsverteilung eines metrisch skalierten Merkmals. Das Schema eines Boxplots beinhaltet als wesentliche Kenngrößen den kleinsten und den größten Beobachtungswert  $x_{(1)}$  und  $x_{(n)}$  sowie die drei Quartile  $x_{0,25}$ ,  $x_{0,5}$  und  $x_{0,75}$ , wobei das zweite Quartil mit dem Median identisch ist. Auf einer Skala werden die Quartile durch waagerechte Striche abgetragen und die Striche des 1. und 3. Quartils zu einer Box vervollständigt. Die Linie innerhalb der Box

<sup>4</sup>Vgl. u.a. Rönnz, B., Strohe, H.G. (Hrsg.)(1994), S. 60 f.

kennzeichnet den Median. Die Höhe der Box entspricht dem Interquartilsabstand (interquartile range), der als Differenz zwischen dem oberen Quartil  $x_{0,75}$  und dem unteren Quartil  $x_{0,25}$  definiert ist:  $IQR = x_{0,75} - x_{0,25}$ . Innerhalb der Box liegen die mittleren 50% aller beobachteten Werte. Weitere waagerechte Linien werden entweder beim kleinsten und größten Wert, sofern sie keine extremen Werte sind, bzw. beim 1,5fachen des Quartilsabstandes von der Boxbegrenzung gezogen. Diese werden mit der Box durch senkrechte Linien (whiskers) verbunden. Die Grenzen  $x_{0,25} - 1,5 \cdot IQR$  und  $x_{0,75} + 1,5 \cdot IQR$  werden als „lower fence“ bzw. „upper fence“ bezeichnet. Werte, die mehr als das 1,5fache des Interquartilsabstandes von der Boxbegrenzung entfernt liegen, werden als Extremwerte gekennzeichnet. Der Boxplot gibt anschaulich Verteilung und Struktur der Beobachtungsdaten wieder.

Abbildung 2.4.: Schematische Darstellung eines Boxplots



Unter SPSS kann ein Boxplot auf zwei verschiedene Möglichkeiten erstellt werden.

### 1. Möglichkeit

Es ist zu wählen

■ Analyze

■ Descriptive Statistics

■ Explore ...

Analog wie vorher wird im Dialogfeld „Explore“ (siehe Abb. 2.1) die zu analysierende Variable

## 2. Entdeckung und Identifikation von Ausreißern

in das Feld „Dependent List.“ gebracht, in Feld „Display“ der Kreis vor Plots markiert und anschließend die Schaltfläche „Plots...“ aktiviert. Im Dialogfeld „Explore: Plots“ (siehe Abb. 2.2) wird die Voreinstellung bei Boxplots auf „Factor levels together“ belassen und im Feld „Descriptive“ der Stem-and-leaf Plot deaktiviert.

### 2. Möglichkeit

Es wird im Data Editor

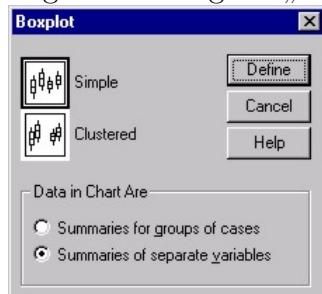
#### ■ Graphs

und im sich öffnenden Pull-Down-Menü

#### ■ Boxplot...

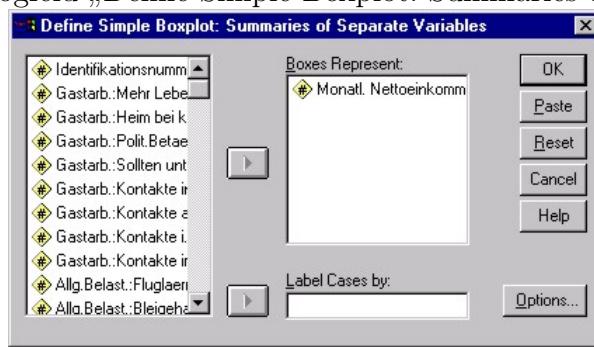
gewählt.

Abbildung 2.5.: Dialogfeld „Boxplot“



Im Dialogfeld „Boxplot“ bleibt es bei der Voreinstellung „Simple“, im Feld „Data in Chart Are“ wird auf „Summaries of separate variables“ entschieden und dann die Schaltfläche „Define“ betätigt.

Abbildung 2.6.: Dialogfeld „Define Simple Boxplot: Summaries of Separate Variables“



In diesem Dialogfeld wird die zu analysierende Variable aus der linken Variablenliste in das Feld „Boxes Represent:“ gebracht und „OK“ betätigt.

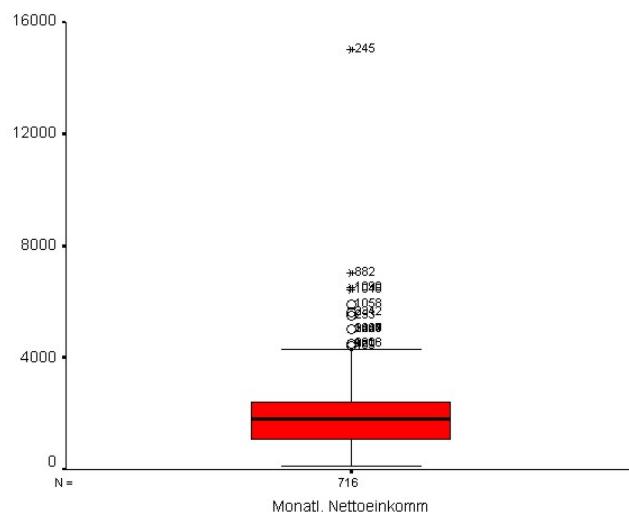
Im SPSS for Windows Viewer kann die Grafik durch einen Doppelklick in den Chart Editor gebracht werden, wo der Boxplot bearbeitet werden kann.

SPSS kennzeichnet im Boxplot extreme Werte durch einen Kreis, wenn sie zwischen dem 1,5fachen und dem 3fachen des Interquartilsabstandes von der Box entfernt liegen, und durch einen Stern, wenn sie mehr als das 3fache des Interquartilsabstandes von der Box entfernt sind. An diesen Extremwerten werden die Fallnummern angegeben, so dass man sie schnell in der Datei identifizieren kann. Unterhalb der Grafik ist noch die Anzahl der gültigen Fälle mit N= vermerkt.

- Beispiel 2.2 (Fortsetzung):

Für die Variable einkomp1 (monatliches Nettoeinkommen in DM) der Datei allbus.sav wird der Boxplot erstellt.

Abbildung 2.7.: Beispiel eines Boxplots



Im Boxplot weisen 6 Fälle extreme Beobachtungswerte auf, die durch einen Kreis gekennzeichnet sind, und 4 Fälle extreme Beobachtungswerte, die durch einen Stern markiert sind. Vor allem Fall 245 könnte ein Ausreißer sein, was jedoch weiter geprüft werden muss.

## 2.3. Scatterplot

Um potentielle Ausreißer bei bivariaten und (in eingeschränktem Sinne auch bei) multivariaten Datenanalysen zu entdecken, bietet sich als graphisches Instrument der Scatterplot bzw. die Scatterplot-Matrix<sup>5</sup> an.

Der Scatterplot (Streuungsdiagramm) ist eine graphische Darstellung der Beobachtungswerte zweier metrisch skalierten Variablen X und Y in einem kartesischen Koordinatensystem.

---

<sup>5</sup>Vgl. u.a. Röenz, B., Strohe, H.G. (Hrsg.) (1994), S. 353, 320 f.

## 2. Entdeckung und Identifikation von Ausreißern

Jedes Paar von Beobachtungswerten ( $x_i, y_i$ ) erscheint als Punkt in der Variablenebene.

Die Scatterplot-Matrix (Draftsman-Display) ist ein graphisches Verfahren zur Veranschaulichung von paarweisen Zusammenhängen zwischen mehr als zwei Variablen. Gegeben sind  $r$  ( $r > 2$ ) metrisch skalierte Variable  $X_1, \dots, X_r$ , für die Beobachtungen an  $n$  statistischen Elementen vorliegen. Für jeweils zwei Variable werden die Beobachtungswerte in einem Scatterplot dargestellt. Es ergeben sich insgesamt  $r \cdot (r - 1)$  Plots<sup>6</sup>. Diese werden in Form einer Matrix angeordnet, wobei die Hauptdiagonale leere Flächen enthält, da die Variable  $X_j$  nicht gegen sich selbst abgetragen wird. Die sich ergebende Scatterplot-Matrix ermöglicht, visuell Beziehungen bzw. Anomalitäten in den Daten zwischen den Variablen bzw. Gruppen einander ähnlicher statistischer Elemente zu entdecken.

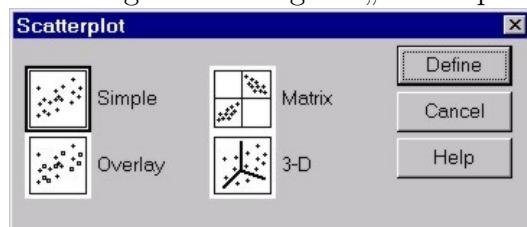
Geht man zunächst von zwei Variablen aus, so kann unter SPSS ein einfacher Scatterplot über

### ■ Graphs

#### ■ Scatter...

erstellt werden. Im Dialogfeld „Scatterplot“ wird die Voreinstellung für einen einfachen Scatterplot (Simple) belassen und die Schaltfläche „Define“ betätigt.

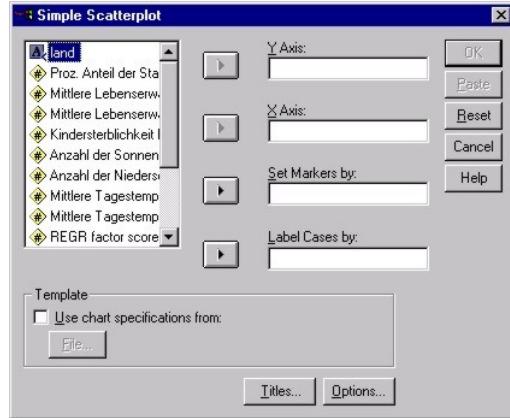
Abbildung 2.8.: Dialogfeld „Scatterplot“



Im Dialogfeld „Simple Scatterplot“ werden die beiden Variablen, für die ein Plot erstellt werden soll, in die Felder „Y Axis:“ bzw. „X Axis:“ gebracht, wobei der Nutzer entscheidet, welche Variable auf welcher Achse erscheinen soll. Eine Fallbeschriftung ist möglich, sollte jedoch nur bei

<sup>6</sup>Die Anzahl der Plots ergibt sich als die Anzahl der Variationen von  $r$  Elementen zur zweiten Klasse ohne Wiederholung der Elemente:  $V(r, 2) = r!/(r - 2)!$ , da die Anordnung der Elemente berücksichtigt werden muß.

Abbildung 2.9.: Dialogfeld „Simple Scatterplot“

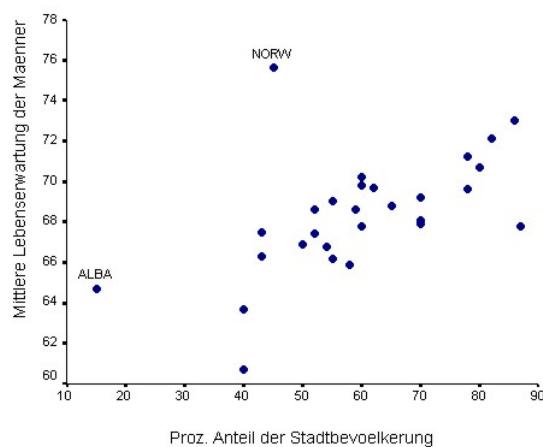


einem Plot mit wenigen Punkten verwendet werden. Nach Klick auf „OK“ erscheint im SPSS Viewer der Scatterplot. Die Grafik kann editiert werden, indem sie durch einen Doppelklick in den SPSS Chart Editor gebracht wird. Im Chart Editor stehen vielfältige Möglichkeiten für Veränderungen bzw. Ergänzungen des Plots zur Verfügung, worauf hier im einzelnen nicht eingegangen werden soll.

- Beispiel 2.3:

Aus der Datei europa.sav, die bereits im Beispiel 2.1 verwendet wurde, wird die Variable sb (prozentualer Anteil der Stadtbevölkerung an der Gesamtbevölkerung) für die X-Achse und die Variable lem (mittlere Lebenserwartung der Männer) für die Y-Achse ausgewählt. Als Fallbeschriftung wird die Variable land verwendet.

Abbildung 2.10.: Beispiel für einen einfachen Scatterplot



Zwei Punkte fallen etwas aus der resultierenden Punktwolke heraus. SPSS ermöglicht die Identifizierung der Punkte im Chart Editor durch das Point Selection Ikon .

## 2. Entdeckung und Identifikation von Ausreißern

Abbildung 2.11.: Point Selection Ikon



Ein Klick auf dieses Ikon aktiviert den Point Selection Mode, wodurch der Cursor sein Aussehen zu einem „Fadenkreuz“ verändert. Mit diesem Cursor geht man auf den interessierenden Punkt im Scatterplot und klickt ihn an. Die Identifizierung kann in verschiedener Weise erfolgen:

- Wurde bei der Erstellung des Scatterplots keine Variable für „Label Cases by“ (siehe Abb. 2.9) angegeben, dann erscheint die zu diesem Punkt gehörige Fallnummer.
- Wurde eine Variable für „Label Cases by“ spezifiziert, erfolgt die Identifizierung mit der zu diesem Punkt gehörenden Ausprägung der Variablen.

Wenn der Datenfile, aus dem heraus der Scatterplot erstellt wurde, noch geöffnet ist, wird der Fall markiert und kann durch einen Wechsel zum SPSS Data Editor inspiziert werden. Ein nochmaliger Klick auf den Punkt entfernt die Punkt-Identifikation und ein nochmaliger Klick auf das Point Selection Ikon deaktiviert den Selection Mode.

Die Punkt-Identifizierung zeigt an, dass die beiden auffälligen Punkte zu den Ländern Norwegen (Fallnummer 17) und Albanien (Fallnummer 1) gehören.

Im multivariaten Fall kann über eine Scatterplot-Matrix geprüft werden, ob bestimmte Werte in bivariaten Variablenkombinationen immer wieder als Extremwerte bzw. Ausreißer sichtbar werden. Dazu wird im Dialogfeld „Scatterplot“ (siehe Abb. 2.8) auf „Matrix“ entschieden. Im sich öffnenden Dialogfeld „Scatterplot Matrix“ werden alle gewünschten Variablen aus der Variablenliste in das Feld „Matrix Variables:“ gebracht.

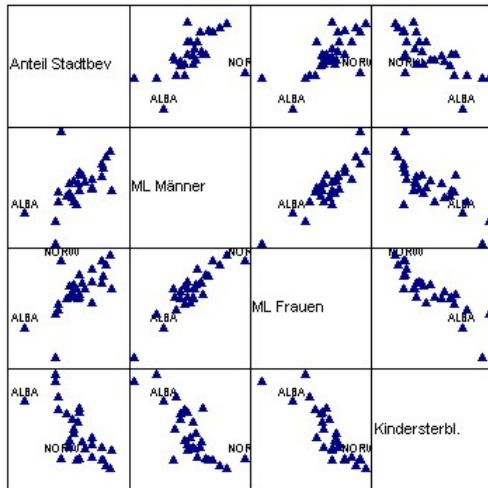
Abbildung 2.12.: Dialogfeld „Scatterplot Matrix“



- Beispiel 2.3 (Fortsetzung):

Es soll überprüft werden, ob die gleichen Fälle auch bei anderen bivariaten Variablenkombinationen in der Datei europa.sav auffällige Beobachtungspunkte aufweisen. Dazu werden die Variablen mittlere Lebenserwartung der Männer (lem), mittlere Lebenserwartung der Frauen (lew), Kindersterblichkeit je 1000 Geburten (ks) und prozentualer Anteil der Stadtbevölkerung (sb) paarweise in einer Scatterplot-Matrix veranschaulicht. Für die Auswertung sind nur die Scatterplots unterhalb der Matrixdiagonalen von Interesse, da die Plots oberhalb der Diagonalen nur die entsprechenden Spiegelbilder sind. Der Scatterplot der Abb. 2.10 ist im Matrix-Feld (2,1) enthalten. Wird im SPSS Chart Editor in diesem Scatterplot die gleiche Punktidentifizierung wie vorher durchgeführt, so werden in den anderen Plots der Scatterplot-Matrix diese Fälle ebenfalls markiert.

Abbildung 2.13.: Beispiel einer Scatterplot-Matrix



Die Länder Norwegen (Fallnummer 17) und Albanien (Fallnummer 1) sind auch in dem Scatterplot von prozentuellem Anteil der Stadtbevölkerung und mittlere Lebenserwartung der Frauen im Matrix-Feld (3,1) auffällig, jedoch nicht in anderen Plots.

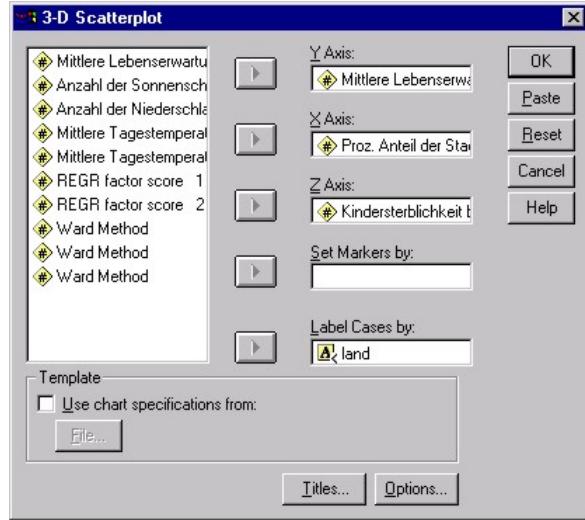
Für die Kombination von 3 Variablen kann man den 3-D Scatterplot im Dialogfeld „Scatterplot“ (siehe Abb. 2.8) wählen. Im Dialogfeld „3-D Scatterplot“ (siehe Abb. 2.14) werden die Variablen den drei Achsen zugewiesen. Wie vorher ist eine Fallbeschriftung möglich. Ebenso kann in dem erzeugten 3-D Scatterplot eine Punkt-Identifizierung vorgenommen werden.

- Beispiel 2.3 (Fortsetzung):

Es werden die Variablen lem (mittlere Lebenserwartung der Männer) für die Y-Achse, ks (Kindersterblichkeit je 1000 Geburten) für die Z-Achse und sb (prozentualer Anteil der Stadt-

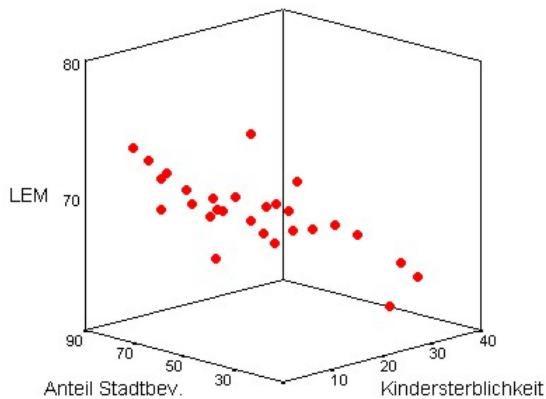
## 2. Entdeckung und Identifikation von Ausreißern

Abbildung 2.14.: Dialogfeld „3-D Scatterplot“



bevölkerung) für die X-Achse aus der Datei europa.sav für die Darstellung im 3-D Scatterplot verwendet. Mit der in Abb. 2.15 gegebenen Sicht in den dreidimensionalen Raum sind potentielle Ausreißer nicht auszumachen.

Abbildung 2.15.: Beispiel eines 3-D Scatterplot



Da es bei der Betrachtung eines 3-D Scatterplots oft schwierig ist, sich die Lage der Punkte im dreidimensionalen Raum vorzustellen, gibt es die Möglichkeit von Projektionen. Dazu muß sich der 3-D Scatterplot im Chart Editor befinden. Durch Anwahl von „Chart“ in der Menü-Leiste des Chart-Editors öffnet sich ein Pull-Down-Menü, in dem „Option...“ auszuwählen ist. Es kann auch das Ikon „Chart Options“ angewählt werden. Es öffnet sich das in Abb. 2.16 enthaltene Dialogfeld.

Abbildung 2.16.: Dialogfeld „3-D Scatterplot: Options“

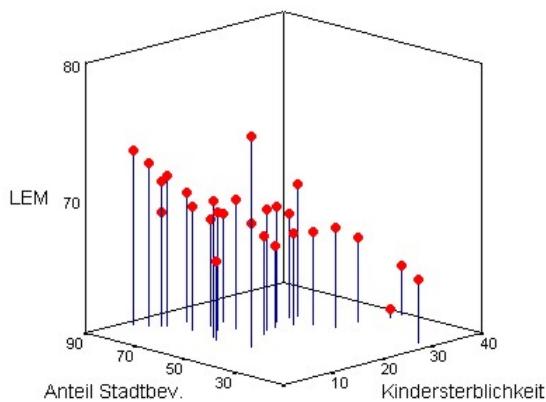


Bei „Spikes:“ kann unter drei verschiedenen Projektionsmöglichkeiten gewählt werden:

- Floor (Parallel): Es erfolgt eine Projektion auf die X-Ebene.
- Centroid (Zentroid): Es erfolgt eine Projektion zum Mittelpunkt der Punktwolke.
- Origin (Ursprung): Es erfolgt eine Projektion zum Koordinatenursprung.

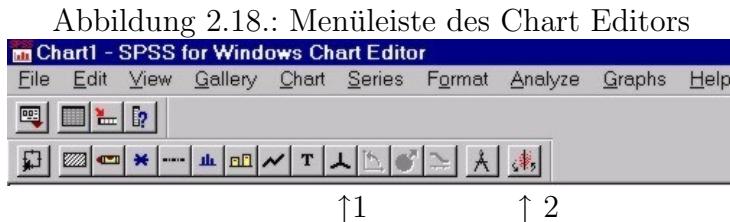
Die folgende Abbildung zeigt den 3-D Scatterplot der Abb. 2.15 mit der Projektion Floor.

Abbildung 2.17.: Beispiel eines 3-D Scatterplot mit Floor-Projektion



Die 3-D Darstellung bietet weiterhin die Möglichkeit der Rotation der Punktwolke über die verschiedenen Achsen. Auf diese Weise ist es möglich, die Punktwolke aus einem anderen Blickwinkel zu betrachten, wobei möglicherweise aus der Punktwolke herausfallende Punkte besser erkennbar sind. Dazu muß sich der 3-D Scatterplot ebenfalls im Chart Editor befinden.

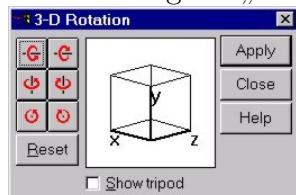
## 2. Entdeckung und Identifikation von Ausreißern



Für die Rotation bieten sich zwei Möglichkeiten:

1. Eine Möglichkeit der Rotation wird über das Ikon „3-D Rotation“ (siehe mit 1 gekennzeichnetes Ikon in Abb. 2.18) bzw. durch Anwahl von „Format“ in der Menü-Leiste des Chart Editors und anschließender Auswahl von „3-D Rotation“ in dem Pull-Down-Menü angeboten. Es öffnet sich das Dialogfeld der Abb. 2.19.

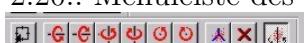
Abbildung 2.19.: Dialogfeld „3-D Rotation“



Über die Schaltflächen auf der linken Seite kann man bei gedrückter Maustaste das Koordinatensystem in die gewünschte(n) Richtung(en) drehen. Nach Betätigung von „Apply“ (Zuweisen) wird die Rotation ausgeführt. Der Nachteil dieser Rotationsmöglichkeit ist, dass der gedrehte 3-D Scatterplot erst nach dem Zuweisen sichtbar wird. Über die Schaltfläche „Reset“ kann die Rotation wieder rückgängig gemacht werden.

2. Die andere Möglichkeit der Rotation ist über das Ikon „Set/exit spin mode“ (siehe mit 2 gekennzeichnetes Ikon in Abb. 2.18) bzw. durch Anwahl von „Format“ in der Menü-Leiste des Chart-Editors und anschließender Auswahl von „Spin Mode“ in dem Pull-Down-Menü gegeben, wodurch Ikons in der Menü-Leiste erscheinen, die das Drehen des 3-D Scatterplots in verschiedene Richtungen erlauben.

Abbildung 2.20.: Menüleiste des Spin Mode

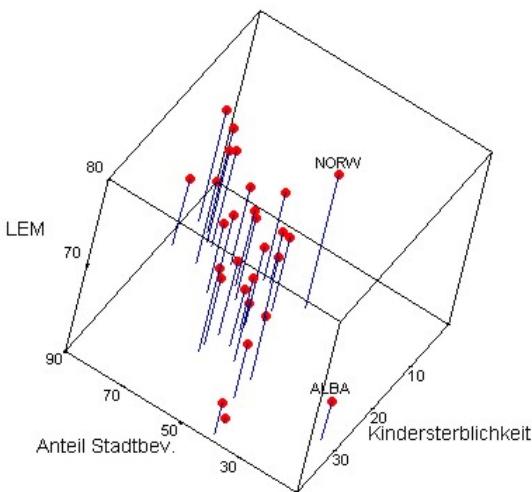


Wählt man eine dieser Schaltflächen an und hält die linke Maustaste gedrückt, so wird die Drehung sofort ausgeführt, wodurch eine unmittelbare Verfolgung der Veränderung der

Punktwolke wesentlich erleichtert wird. Bei nochmaliger Betätigung des Ikon „Set/exit spin mode“ verbleibt der 3-D Scatterplot in der letzten Rotationsposition. Wählt man das Ikon „Reset“, so wird der 3-D Scatterplot in die Ausgangsposition gebracht, wobei der Spin Mode weiter aktiv bleibt. Wählt man das Ikon „Cancel“, so wird der 3-D Scatterplot in die Ausgangsposition zurückgesetzt und der Spin Mode deaktiviert.

Abb. 2.21 zeigt den 3-D Scatterplot der Abb. 2.15 in rotierter Form.

Abbildung 2.21.: Beispiel eines rotierten 3-D Scatterplots



Auf diese Weise ist es möglich, potentielle Ausreißer zu finden, die in der Ursprungslage des Koordinatensystems nicht sichtbar sind. Hier sind es die Beobachtungspunkte der Länder Norwegen und Albanien, die etwas aus der Punktwolke herausfallen, damit aber nicht zwangsläufig Ausreißer sein müssen.

## 2.4. Andrews-Plot

Für eine multivariate Analyse schlug Andrew<sup>7</sup> einen Plot vor, bei dem jeder Datenpunkt  $\mathbf{x} = (x_1, \dots, x_p)$ , wobei  $p$  die Anzahl der Variablen bezeichnet, mittels der Funktion

$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \dots \quad (2.1)$$

dargestellt und diese Funktion im Bereich  $-\pi < t < \pi$  gezeichnet wird. Ein multivariater Datenpunkt wird dadurch in einem zweidimensionalen Raum  $(t, f_x(t))$  abgebildet. Jedes statistische Element (Fall) wird durch eine Kurve in demselben Koordinatensystem repräsentiert.

---

<sup>7</sup>Andrews, D.F. (1972); siehe auch du Toit, S.H.C., Steyn, A.G.W., Stumpf, R.H. (1986, S. 59 - 63).

## 2. Entdeckung und Identifikation von Ausreißern

Sie können somit leicht miteinander verglichen werden. Elemente mit „ähnlichen“ Werten über die p Variablen werden ähnliche Kurven haben.

Der Andrews-Plot ist damit für die Entdeckung von Ausreißern und auch zum Auffinden von Clustern geeignet. Allerdings ist der Plot nur bei einer kleineren Anzahl n von Fällen gut überschaubar.

Der Andrews-Plot weist u.a. folgende zwei wichtige Eigenschaften auf:

- Die Funktionsdarstellung bewahrt die Mittelwerte der p Merkmale. Die Funktion des Mittelwertvektors  $\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_p)$  ist der punktweise Mittelwert der n Funktionen  $f_{x1}(t), \dots, f_{xn}(t)$ :

$$f_{\bar{x}}(t) = \frac{1}{n} \sum_{i=1}^n f_{x_i}(t). \quad (2.2)$$

- Die Funktionsdarstellung bewahrt Distanzen. Die Distanz zwischen zwei Funktionen ist

$$\|f_{x_i}(t) - f_{x_j}(t)\|_{L_2} = \int_{-\pi}^{\pi} [f_{x_i}(t) - f_{x_j}(t)]^2 dt. \quad (2.3)$$

Diese Distanz ist proportional zur Euklidischen Distanz zwischen den Beobachtungsvektoren  $\mathbf{x}_i$  und  $\mathbf{x}_j$ :

$$\|f_{x_i}(t) - f_{x_j}(t)\|_{L_2} = \pi \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \pi \sum_{k=1}^p (x_{ik} - x_{jk})^2, \quad i, j = 1, \dots, n; i \neq j. \quad (2.4)$$

Der Andrews-Plot ist nicht in SPSS, jedoch u.a. in SYSTAT implementiert, lässt sich aber relativ leicht in jedem Datenanalysesystem programmieren.

### • Beispiel 2.4:

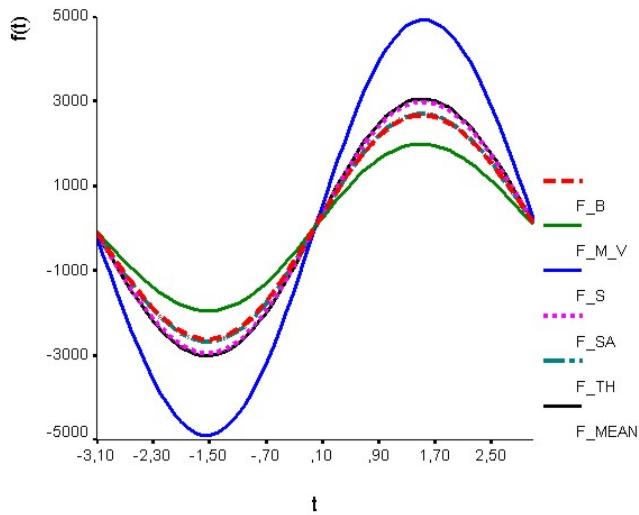
Als einfaches Beispiel wird ein Andrews-Plot für die Variablen Fläche, Einwohner, Arbeitslosenquote und Wahlbeteiligung der Datei beisp.sav erstellt, wobei die statistischen Elemente die neuen Bundesländer sind.

Zum Beispiel ergibt sich für das Land Brandenburg der Wert von  $f_x(t)$  an der Stelle  $t = -3, 1$  gemäß (2.1) zu:

$$\begin{aligned} f_B(t = -3, 1) &= \frac{29, 1}{\sqrt{2}} + 2641 \cdot \sin(-3, 1) + 10, 3 \cdot \cos(-3, 1) + 67, 1 \cdot \sin(2 \cdot -3, 1) \\ &= 20, 58 + (-109, 81) + (-10, 29) + 5, 58 = -93, 94 \end{aligned}$$

Im Andrews-Plot steht B für Brandenburg, M\_V für Mecklenburg-Vorpommern, S für Sachsen, SA für Sachsen-Anhalt, TH für Thüringen und Mean für den Mittelwert.

Abbildung 2.22.: Andrews-Plot



Die zu Sachsen gehörige Funktion ist deutlich verschieden von den anderen Funktionen, so dass dieses Land zumindest ein auffälliges Element ist, bei dem es zu prüfen gilt, ob es ein Ausreißer ist.

## 2.5. Ausreißertests

Wenn über die univariate graphische Exploration der Verdacht auf Ausreißer erhärtet wurde, steht als nächstes das Problem der weiteren statistischen Prüfung mittels Ausreißertests. Ein Verdacht auf Ausreißer besagt, dass die jeweilige(n) Beobachtung(en) in der Stichprobe aus einer anderen Grundgesamtheit als die anderen Beobachtungen stammt, d.h. aus einer Grundgesamtheit mit einer anderen Verteilung. Dabei bedeutet „andere Verteilung“ nicht zwangsläufig einen anderen Verteilungstyp, sondern es kann sich um den gleichen Verteilungstyp, jedoch mit anderen Parametern handeln (z.B. andere Werte der Parameter  $\mu$  und  $\sigma^2$  der Normalverteilung). Vorsicht ist jedoch bei der Einschätzung von potentiellen Ausreißern in der Hinsicht geboten, dass die betrachtete Stichprobe aus einer inhomogenen Grundgesamtheit stammen kann und somit die auffälligen Beobachtungswerte durchaus zu dieser Grundgesamtheit gehören können. Diese Ausreißertests kranken in zweierlei Hinsicht:

- Sie setzen im allgemeinen voraus, dass die Stichprobe aus einer normalverteilten Grundgesamtheit stammt. Es ist aber nur in den seltensten Fällen bekannt, ob eine Grundgesamtheit normalverteilt ist. Die Prüfung auf Normalverteilung kann somit nur auf der Basis der gezogenen Stichprobe erfolgen. Sind in der Stichprobe aber ausreißerverdächtige Werte enthalten, so werden sie diese Prüfung auf Normalverteilung stark beeinflussen,

## 2. Entdeckung und Identifikation von Ausreißern

d.h. man wird wohl eher die Hypothese auf Normalverteilung verwerfen.

Wird für den Ausreißertest keine Normalverteilung vorausgesetzt, so ist die Verteilung der Teststatistik nicht bekannt. Exakte kritische Werte können nicht angegeben, sondern nur über Simulationsstudien ermittelt werden.

- Für eine ganze Reihe von Ausreißertests wurden Tabellen mit kritischen Werten erstellt. Sie sind jedoch nicht in den Standardwerken der Statistik enthalten und somit bei praktischen Testdurchführungen nur mit größerem Aufwand zugänglich.

Trotzdem sollem im weiteren einige univariate Ausreißertests<sup>8</sup> vorgestellt werden. Sie setzen alle Anordnungswerte (order statistics) voraus. Die Sortierung von Fällen unter SPSS wurde bereits im ersten Kapitel behandelt.

Die nachfolgend behandelten Ausreißertests sind nicht in SPSS implementiert. Ihre Durchführung ist jedoch durch die Ausgabe bestimmter Kenngrößen unter SPSS und unter Verwendung eines Taschenrechners möglich.

### Grubbs-Test

Dieser Ausreißertest, der von einer normalverteilten Grundgesamtheit ausgeht, prüft, ob der kleinste (1) oder der größte (n) Beobachtungswert ein Ausreißer ist:

$H_0 (1)$  :  $x_{(1)}$  ist kein Ausreißer bzw.

$H_0 (n)$  :  $x_{(n)}$  ist kein Ausreißer.

Die Teststatistik basiert auf der studentisierten Differenz des kleinsten bzw. größten Wertes vom Mittelwert:

$$T_1 = \frac{\bar{x} - x_{(1)}}{s} \quad \text{bzw.} \quad T_n = \frac{x_{(n)} - \bar{x}}{s} \quad (2.5)$$

mit s als Standardabweichung und  $\bar{x}$  als Mittelwert der Stichprobe.  $H_0$  wird zum vorgegebenen Signifikanzniveau  $\alpha$  verworfen, wenn  $T_1 > T_{n;\alpha}$  bzw.  $T_n > T_{n;\alpha}$  ist, wobei  $T_{n;\alpha}$  kritische Werte des Grubbs-Test (Grubbs/Beck, 1972) sind.

Die für die Formel (2.5) benötigten Kenngrößen kann man unter SPSS u.a. über

■ Analyze

■ Descriptive Statistics

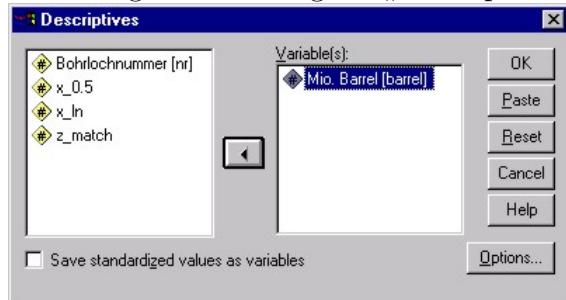
■ Descriptives...

erhalten.

---

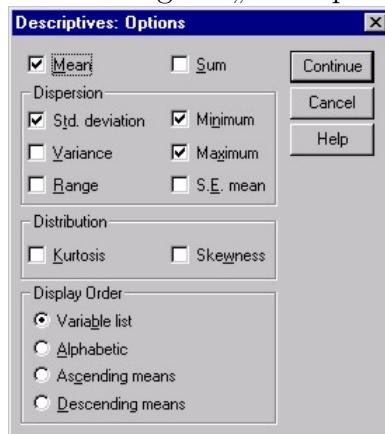
<sup>8</sup>Die Ausführungen zu den Ausreißertests sind angelehnt an die am Ende des Kapitels angegebene Literatur, vor allem Hartung/Elpelt/Klösener (1993)

Abbildung 2.23.: Dialogfeld „Descriptives“



In diesem Dialogfeld wird (werden) die gewünschte(n) Variable(n) aus der linken Variablenliste in das Feld „Variable(s)“ gebracht. Über die Betätigung der Schaltfläche „Options...“ gelangt man in das Dialogfeld „Descriptives: Options“, in dem man diejenigen Kenngrößen anklicken kann, die berechnet werden sollen: Mittelwert (Mean), Standardabweichung (Std. deviation), Minimum und Maximum.

Abbildung 2.24.: Dialogfeld „Descriptives: Options“



Nach Klick auf „Continue“ und anschließend auf „OK“ erhält man den SPSS-Output. Mit diesen Kenngrößen kann der Wert der Grubbs-Teststatistik berechnet werden.

- Beispiel 2.5:

In der Datei erdöl.sav ist eine Zufallsstichprobe vom Umfang  $n = 58$  Bohrlöchern eines Erdölfeldes mit der je Bohrloch geförderten Menge (in Mio. Barrel) enthalten. Diese Stichprobe wird auf potentielle Ausreißer untersucht, wobei zunächst ein Stem-and Leaf Plot und ein Boxplot erstellt wird und die Ausgabe der Outliers erfolgt. Die Datei wurde nach der Größe der Beobachtungswerte der Variablen Barrel geordnet. Die Variable Nr (Bohrlochnummer) wird für Label Cases by verwendet.

## 2. Entdeckung und Identifikation von Ausreißern

### SPSS-Output 2.5-1: Stem-and-Leaf Plot und Outliers für Beispiel 2.5 Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Mio Barrel	58	100,0%	0	,0%	58	100,0%

Mio. Barrel Stem-and-Leaf Plot

Frequency      Stem & Leaf

32,00	0 .	00000000011111122222223334444
8,00	0 .	55567789
5,00	1 .	00112
5,00	1 .	57778
1,00	2 .	1
7,00	Extremes	(>=335)

Stem width: 100,0

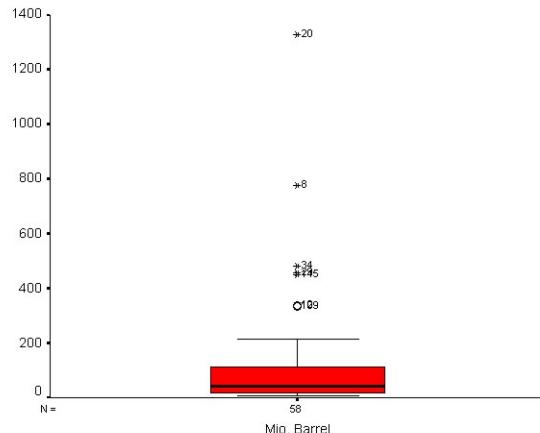
Each leaf: 1 case(s)

### Extreme Values

		Case Number	Bohrlochnummer	Value
Mio. Barrel	Highest	1	58	1328,0
		2	57	775,0
		3	56	482,0
		4	55	455,0
		5	54	450,0
	Lowest	1	1	152
		2	3	52
		3	2	209
		4	4	81
		5	5	141

a. Only partial list of cases with the value 9 are shown in the table of lower extremes

Abbildung 2.25.: Boxplot für Beispiel 2.5



Die im Stem-and-Leaf Plot angezeigten 7 Extremwerte scheinen bei näherer Betrachtung des Datensatzes bezüglich der Anzahl etwas übertrieben zu sein. Dagegen signalisiert der Boxplot zwei potentielle Ausreißer: die Bohrlöcher mit den Nummern 20 und 8.

Es soll nun mittels des Grubbs-Tests geprüft werden, ob diese Bohrlöcher tatsächlich aus statistischer Sicht als Ausreißer einzustufen sind. Dabei soll die Prüfung auf die durch den Test geforderte Normalverteilung hier zunächst weggelassen werden, da diese Prüfung erst im nachfolgenden Kapitel behandelt wird.

Für dieses Beispiel reichen die voreingestellten Kenngrößen Mittelwert, Standardabweichung, Minimum und Maximum im Dialogfeld „Descriptives: Options“ aus.

### SPSS-Output 2.5-2: Kenngrößen für den Grubbs-Test

**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
Mio Barrel	58	5,9	1328,0	120,721	217,676
Valid N (listwise)	58				

Die Nullhypothese soll auf dem 5%-Niveau geprüft werden und lautet:  $x_{58}$ (Bohrloch 20) ist kein Ausreißer.

Nach (2.5.) folgt:

$$T_{18} = (1328 - 120,721) / 217,676 = 5,546.$$

Bei Grubbs/Beck (1972), S. 848 f., findet man für  $\alpha = 0,05$  den kritischen Wert  $T_{58;0,05} = 3,013$ .  $H_0 : „x_{(58)}“$  ist kein Ausreißer“ wird somit auf dem 5%-Niveau abgelehnt.

### Grubbs/Beck-Test

Der Grubbs/Beck-Test ist ein Test auf ein Ausreißerpaar. Bei Unterstellung der Normalverteilung in der Grundgesamtheit prüft dieser Test die Nullhypothese

$H_0$ : Die beiden kleinster Werte ( $x_{(1)}$  und  $x_{(2)}$ ) sind keine Ausreißer bzw.

$H_0$ : Die beiden größten Werte ( $x_{(n)}$  und  $x_{(n-1)}$ ) sind keine Ausreißer bzw.

mittels der Teststatistik

$$T_{1,2} = \frac{SQA_{1,2}}{SQA} \quad \text{bzw.} \quad T_{n,n-1} = \frac{SQA_{n,n-1}}{SQA}, \quad (2.6)$$

worin

$$SQA = \sum_{i=1}^n (x_{(i)} - \bar{x})^2, \quad \bar{x} = \frac{\sum_{i=1}^n x_{(i)}}{n},$$

## 2. Entdeckung und Identifikation von Ausreißern

$$\begin{aligned}
 SQA_{1,2} &= \sum_{i=3}^n (x_{(i)} - \bar{x}_{1,2})^2, & \bar{x}_{1,2} &= \frac{\sum_{i=3}^n x_{(i)}}{n-2}, \\
 SQA_{n,n-1} &= \sum_{i=1}^{n-2} (x_{(i)} - \bar{x}_{n,n-1})^2, & \bar{x}_{n,n-1} &= \frac{\sum_{i=1}^{n-2} x_{(i)}}{n-2},
 \end{aligned} \tag{2.7}$$

sind und SQA als Abkürzung für Summe der quadratischen Abweichungen steht. Da durch die Herausnahme der beiden potentiellen Ausreißer die SQA im Zähler kleiner sein sollte als diejenige im Nenner der Teststatistik, führen kleine Werte der Teststatistik zur Ablehnung der  $H_0$ .  $s_{n;\alpha}$  sind von Grubbs und Beck (1972) angegebene kritische Werte zu diesem Test. Wenn also  $T_{1,2} < s_{n;\alpha}$  bzw.  $T_{n,n-1} < s_{n;\alpha}$  ist, wird die  $H_0$  auf dem vorgegebenen Signifikanzniveau abgelehnt.

- Beispiel 2.5 (Fortsetzung):

Für die Variable Barrel der Datei erdöl.sav wird geprüft, ob die Beobachtungswerte der Bohrlöcher 20 und 8 Ausreißer sind. Auch hier soll die Prüfung auf die durch den Test geforderte Normalverteilung zunächst weggelassen werden. Im Dialogfeld „Descriptives: Options“ (siehe Abb. 2.24) wird die Varianz zur Ausgabe angekreuzt. Sie beträgt  $s^2 = 47382,679$ . Daraus lässt sich die Summe der quadratischen Abweichungen SQA berechnen:

$$SQA = s^2(n-1) = 47382,679 \cdot 57 = 2.700.812,703.$$

Nun werden in der nach der Größe sortierten Datei die beiden Fälle 57 und 58 ausgeschlossen (siehe Fälle auswählen im Kapitel 1). Über das Dialogfeld „Descriptives: Options“ lässt man erneut die Varianz berechnen:  $s_{58,57}^2 = 13696,833$ . Daraus lässt sich die Summe der quadratischen Abweichungen  $SQA_{n,n-1} = SQA_{58,57}$  berechnen:

$$SQA_{58,57} = s_{58,57}^2(n-2-1) = 13696,833 \cdot 55 = 753.325,815.$$

Es folgt für die Grubbs/Beck Teststatistik:

$$T_{58,57} = 753.325,815 / 2.700.812,702 = 0,2789.$$

Wegen  $s_{58;0,05} = 0,7489$  (Grubbs/Beck, 1972, S. 852) wird  $H_0$ : „ $x_{(58)}$  und  $x_{(57)}$  sind keine Ausreißer“ abgelehnt.

### Dixon's r-Statistiken

Der Ausreißertest von Dixon unterstellt Normalverteilung der Grundgesamtheit und prüft die Nullhypothese

$H_0(1)$ :  $x_{(1)}$  ist kein Ausreißer bzw.

$H_0(n)$ :  $x_{(n)}$  ist kein Ausreißer.

Für die Prüfung, dass  $x_{(1)}$  kein Ausreißer ist, wird die Teststatistik

$$r_{kg}(1) = \frac{x_{(1+k)} - x_{(1)}}{x_{(n-g)} - x_{(1)}}, \quad k = 1, 2; g = 0, 1, 2 \quad (2.8)$$

und für die Prüfung, dass  $x_{(n)}$  kein Ausreißer ist die Teststatistik

$$r_{gk}(n) = \frac{x_{(n)} - x_{(n-g)}}{x_{(n)} - x_{(1+k)}}, \quad g = 1, 2; k = 0, 1, 2 \quad (2.9)$$

verwendet. Hierin bezeichnet g die Anzahl der potentiellen Ausreißer mit großen Werten und k die Anzahl der potentiellen Ausreißer mit kleinen Werten. Bei Dixon's r-Statistiken werden somit Spannweiten (ranges) ins Verhältnis gesetzt. Die Teststatistik hängt also davon ab, wie viele „kleinere“ bzw. „größere“ Werte mögliche Ausreißer sind, d.h. relativ weit von den anderen Werten entfernt sind. Wenn es z.B. einen kleinen und zwei große potentielle Ausreißer gibt, dann sind g = 2 und k = 1 und die Teststatistiken lauten:

$$r_{12}(1) = \frac{x_{(2)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}$$

$$r_{21}(n) = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}}$$

Durch die Berücksichtigung von weiteren potentiellen Ausreißern in der Teststatistik soll eine Maskierung vermieden werden. Eine Maskierung tritt bei dicht nebeneinander liegenden potentiellen Ausreißern auf, da die nicht berücksichtigten Ausreißer den Mittelwert an  $x_{(n)}$  bzw.  $x_{(1)}$  heranziehen und die Streuung groß bleibt.

Bei Dixon (1951) findet man kritische Werte  $r_{kg;n;\alpha}$  bzw.  $r_{gk;n;\alpha}$  für die Teststatistik, jedoch nur bis zu einem Stichprobenumfang  $n = 30$ . Wenn  $r_{kg}(1) > r_{kg;n;\alpha}$  bzw.  $r_{gk}(n) > r_{gk;n;\alpha}$  ist, wird  $H_0$  auf dem vorgegebenen Signifikanzniveau  $\alpha$  abgelehnt.

- Beispiel 2.5 (Fortsetzung):

Für die Datei erdöl.sav wird eine Zufallsstichprobe vom Umfang  $n = 30$  aus den verfügbaren Fällen gezogen (siehe Fälle auswählen, random sample of cases, wie im Kapitel 1 beschrieben). Es habe sich folgende Stichprobe ergeben.

## 2. Entdeckung und Identifikation von Ausreißern

**SPSS-Output 2.5-3:** Zufallsstichprobe vom Umfang  $n = 30$  aus der Datei erdöl.sav  
**Case Summaries<sup>a</sup>**

	Bohrlochnummer	Mio. Barrel		Bohrlochnummer	Mio. Barrel
1	152	5,9	16	58	33,0
2	52	6,9	17	105	42,0
3	209	6,9	18	162	49,0
4	176	8,8	19	46	58,0
5	202	8,8	20	33	89,0
6	203	8,8	21	131	93,0
7	174	10,0	22	18	113,0
8	189	12,0	23	9	114,0
9	71	15,0	24	195	125,0
10	178	17,0	25	90	170,0
11	92	19,0	26	145	450,0
12	88	25,0	27	29	455,0
13	210	25,0	28	34	482,0
14	7	26,0	29	8	775,0
15	11	31,0	30	20	1328,0

a. Limited to first 58 cases.

Prüft man für die Variable Barrel die  $H_0$ : „ $x_{(30)}$  ist kein Ausreißer“ nach Dixon's r-Statistik unter Berücksichtigung der Tatsache, dass auch  $x_{(29)}$  ein potentieller Ausreißer ist und kein kleiner Wert als Ausreißer in Frage kommt, so ist  $g = 2$  und  $k = 0$  und die Testgröße lautet:

$$r_{gk}(n) = r_{20}(30) = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}} = \frac{x_{(30)} - x_{(28)}}{x_{(30)} - x_{(1)}} = \frac{1328 - 482}{1328 - 5,9} = 0,63989.$$

Bei Dixon (1951), S. 76, findet man  $r_{20;30;0,05} = 0,322$ .  $H_0$  wird auf dem 5%-Niveau abgelehnt.

### David-Hartley-Pearson-Test

Dieser Ausreißertest, der ebenfalls von einer normalverteilten Grundgesamtheit ausgeht, prüft die Nullhypothese

$H_0 : x_{(1)}$  bzw.  $x_{(n)}$  ist kein Ausreißer

mittels der Teststatistik

$$T = \frac{R}{s} = \frac{x_{(n)} - x_{(1)}}{\sqrt{\frac{\sum_{i=1}^n (x_{(i)} - \bar{x})^2}{n-1}}}, \quad (2.10)$$

worin R die Spannweite (range), s die Standardabweichung und  $\bar{x}$  der Mittelwert der Stichprobe sind.

Die  $H_0$  wird auf dem Signifikanzniveau  $\alpha$  nicht verworfen, wenn  $Q_{L;n;\alpha} \leq T \leq Q_{U;n;\alpha}$  ist, wobei  $Q_{L;n;\alpha}$  der untere kritische Wert (lower percentage point) und  $Q_{U;n;\alpha}$  der obere kritische

Wert (upper percentage point) des David-Hartley-Pearson-Tests (David et al., 1954) sind. Bei Ablehnung der  $H_0$  wird der am weitesten vom Mittelwert entfernt liegende Wert (also entweder der kleinste oder der größte Wert) als Ausreißer betrachtet.

- Beispiel 2.5 (Fortsetzung):

Ausgehend von der Datei erdöl.sav und unter Verwendung aller gültigen Fälle wird die Variable Barrel mittels dieses Tests geprüft, ob der kleinste oder der größte Wert ein Ausreißer ist. Im Dialogfeld „Descriptives: Options“ (siehe Abb. 2.24) wird die Spannweite und die Standardabweichung zur Ausgabe angekreuzt. Aus der Ausgabe entnimmt man:

$$R = 1322,1 \text{ und } s = 217,676, \text{ womit } T = 1322,1 / 217,676 = 6,0737$$

ist. Bei David/Hartley/Pearson (1954), S. 491, findet man für den Stichprobenumfang  $n = 58$  keine kritischen Werte. Es werden deshalb diejenigen für  $n = 60$  verwendet:

$$Q_{L;60;0,05} = 3,95 \text{ und } Q_{U;60;0,05} = 5,50.$$

Die  $H_0$  wird auf dem 5%-Niveau abgelehnt. Der größte Wert  $x_{(58)}$  (Bohrloch 20) kann als ein Ausreißer angesehen werden.

Bohrloch 20 und auch Bohrloch 8 wurden im statistischen Sinne durch die Ausreißer-Tests als Ausreißer gekennzeichnet. Ob diese beiden Stichprobenelemente aus einer anderen Grundgesamtheit stammen, kann an dieser Stelle nicht beantwortet werden.

Damit zusammenhängende Fragen wären z.B.:

- Gehören diese beiden Bohrlöcher zu einem (geologisch) anderen Erdölfeld?
- Wurde bei diesen beiden Bohrlöchern eine andere (moderne) Fördertechnik eingesetzt?

## 2.6. Robuste Schätzer für die Lokalisation und Streuung

Wurden durch o.g. graphische Verfahren potentielle Ausreißer bzw. durch einen Ausreißertest Ausreißer im statistischen Sinne identifiziert, so stellt sich die Frage ihrer weiteren Behandlung.

- Eliminierung

Wenn als Ausreißer erkannte Werte aus dem Datensatz (Stichprobe) entfernt werden, sollte dies bei den weiteren Auswertungen angegeben werden. Die Herausnahme der Ausreißer ist unproblematisch, wenn es sich um Erhebungs- oder Eingabefehler handelt, ist aber problematisch, wenn ein Ausreißer Ausdruck der Variabilität in der Grundgesamtheit ist. In letzterem Sinne dürfte er somit nicht entfernt werden, sondern bedarf der speziellen Aufmerksamkeit und Interpretation. Wurden Ausreißer eliminiert, sollte dies bei der Auswertung auf jeden Fall verwerkt werden.

## 2. Entdeckung und Identifikation von Ausreißern

- Ersetzung

Als Ausreißer erkannte Beobachtungswerte können durch andere Werte ersetzt werden, z.B. durch den nächstliegenden Beobachtungswert oder durch das arithmetische Mittel der Stichprobe. Wurden Ausreißer ersetzt, sollte dies bei der Auswertung auf jeden Fall vermerkt werden.

- Ausreißerabschätzung

Die Auswirkungen von Ausreißern auf ein statistisches Verfahren sollten in dem Sinne abgeschätzt werden, indem das Verfahren jeweils unter Einschluß und unter Ausschluß der Ausreißer durchgeführt wird.

- Robuste statistische Verfahren

Wurden Ausreißer entdeckt, so sind sogenannte robuste Schätz- und Testverfahren den sonst üblichen (wie z.B. Mittelwert  $\bar{x}$  und Standardabweichung  $s$ ) vorzuziehen. Als robust werden Schätzer bezeichnet, die relativ unempfindlich gegenüber Ausreißern und Abweichungen von den eigentlich erforderlichen Modellvoraussetzungen sind und noch hinreichend zuverlässige Ergebnisse liefern. Der Robustheitsbegriff bezieht sich auf die Eigenschaft einer Stichprobenfunktion (Schätzer)  $T_n(X_1, \dots, X_n)$  einer Stichprobe  $X_1, \dots, X_n$  für eine Zufallsvariable mit der Verteilungsfunktion  $F(x)$ . Der Schätzer  $T_n$  heißt robust, wenn sich seine Verteilungsfunktion für Veränderungen von  $F(x)$  nur geringfügig verändert, wobei diese Veränderung mit einem geeigneten Abstandsmaß gemessen wird. Hinsichtlich der Robustheit bedeutsame Punktschätzer sind L-Schätzer, M-Schätzer und R-Schätzer, wobei hier nur die L-Schätzer und M-Schätzer behandelt werden sollen.

## Robuste Schätzer für die Lokalisation

### L-Schätzer

Ein L-Schätzer ist eine Linearkombination der Anordnungswerte (order statistics) der Stichprobe  $x_{(1)}, \dots, x_{(n)}$  unter Verwendung von festen Gewichten  $w_i$ :

$$T_n = \sum_{i=1}^n w_i x_{(i)}, \quad (2.11)$$

worin für die Gewichte  $w_i$  gilt:  $0 \leq w_i \leq 1$  für  $i = 1, \dots, n$  und  $\sum w_i = 1$ . Verwendet man symmetrische Gewichte, d.h.  $w_i = w_{n-i+1}$ ,  $i = 1, \dots, n/2$ , so sind L-Schätzer unverzerrte (unbiased) Schätzer für die Lokalisation symmetrischer Verteilungen.

Da die Zuweisung der Gewichte  $w_i$  auf verschiedene Weise erfolgen kann, gibt es verschiedene L-Schätzer, d.h., es handelt sich bei den L-Schätzern um eine Klasse von Schätzern.

### Median

Als spezieller L-Schätzer ergibt sich der Median  $x_{0,50}$  mit den Gewichten

a) bei geradem Stichprobenumfang n:

$$w_i = \begin{cases} \frac{1}{2} & \text{für } i = n \cdot p, n \cdot p + 1 \\ 0 & \text{sonst,} \end{cases} \quad (2.12)$$

b) bei ungeradem Stichprobenumfang n:

$$w_i = \begin{cases} 1 & \text{für } i = (n+1) \cdot p \\ 0 & \text{sonst} \end{cases} \quad (2.13)$$

mit  $p = 0,5$ . Der Median ist im Gegensatz zum arithmetischen Mittel  $\bar{x}$  ein robuster Schätzer, da er nur den (die) mittleren Wert(e) verwendet, während für die Berechnung von  $\bar{x}$  alle Beobachtungswerte, also auch die Ausreißer, herangezogen werden.

Unter SPSS kann man den Median über das Dialogfeld „Explore: Statistics“ (siehe Abb. 2.3) erhalten, in dem auf Descriptives entschieden wird.

### Der $\alpha$ -getrimmte Mittelwert ( $\alpha$ -gestützter Mittelwert, $\alpha$ -trimmed mean)

Beim  $\alpha$ -getrimmten Mittelwert, als eine spezielle Klasse von L-Schätzern, werden  $\alpha \cdot 100\%$  der kleinsten und  $\alpha \cdot 100\%$  der größten Beobachtungswerte bei der Berechnung des Mittelwertes weggelassen, d.h., der Mittelwert wird somit aus den  $(1 - 2\alpha) \cdot 100\%$  der „mittleren“ Beobachtungswerte berechnet. Ist  $v$  die größte ganze Zahl, für die gilt  $v \leq n \cdot \alpha$ , und  $0 \leq \alpha \leq 0,5$ , so ergibt sich der  $\alpha$ -getrimmte Mittelwert als

$$T_n = \bar{x}_{tr,\alpha} = \frac{1}{n - 2v} \sum_{i=v+1}^{n-v} x_{(i)} \quad (2.14)$$

mit den Gewichten

$$w_i = \begin{cases} 0 & \text{für } i \leq v \quad \text{und} \quad i \geq n - v + 1 \\ \frac{1}{n - 2v} & \text{sonst.} \end{cases} \quad (2.15)$$

Die  $v$  kleinsten und die  $v$  größten Stichprobenwerte haben somit keinen Einfluß auf den Schätzwert der Lokalisation. Der  $\alpha$ -getrimmte Mittelwert kann auch als ein gewogenes arithmetisches Mittel aufgefaßt werden:

$$\bar{x}_{tr,\alpha} = \frac{\sum_{i=1}^n w_i x_{(i)}}{\sum_{i=1}^n w_i}, \quad (2.16)$$

## 2. Entdeckung und Identifikation von Ausreißern

wobei  $w_i = 0$  für die ausgeschlossenen Werte und  $w_i = 1$  für die eingeschlossenen Werte ist.

Der  $\alpha$ -getrimmte Mittelwert ist insoweit ein geeigneterer Schätzer als der Median, da er nicht nur die Rangordnung der Werte und einen einzelnen Wert berücksichtigt, sondern auf einer größeren Anzahl von mittleren Beobachtungswerten beruht.

Der  $\alpha = 0,25$  getrimmte Mittelwert wird als midmean bezeichnet, da er den mittleren Teil der Beobachtungswerte in die Berechnung einbezieht.

Oftmals verwendet man für den  $\alpha$ -getrimmten Mittelwert nicht nur den ganzzahligen Teil von  $\alpha n$ , sondern trimmt exakt, d.h. auch unter Berücksichtigung des gebrochenen Teils  $g = \alpha n - v$ . Das führt zu

$$T_n = \frac{1}{n(1-2\alpha)} \left\{ (1-g)[x_{v+1} + x_{n-v}] + \sum_{i=v+2}^{n-v-1} x_{(i)} \right\}, \quad (2.17)$$

worin die Gewichte gegeben sind mit

$$w_i = \begin{cases} 0 & \text{für } i \leq v \text{ und } i \geq n - v + 1 \\ \frac{v+1-\alpha n}{(1-2\alpha)n} & \text{für } i = v+1, n-v \\ \frac{1}{(1-2\alpha)n} & \text{für } v+1 < i < n-v. \end{cases} \quad (2.18)$$

Unter SPSS wird ein 0,05-getrimmter Mittelwert (im Output bezeichnet mit 5% Trimmed Mean) nach der Formel (2.18) berechnet, den man über die Dialogfelder „Explore“ (siehe Abb. 2.1) und „Explore: Statistics“ (siehe Abb. 2.3) bei Wahl von Descriptives erhalten kann.

### Der $\alpha$ -winsorisierte Mittelwert

Winsorisierung (Winsorisation, nach Ch.P. Winsor) bedeutet, dass in der geordneten Stichprobe die Ausreißer durch ihre benachbarten Werte ersetzt werden, wodurch die Richtung ihrer Abweichung jedoch berücksichtigt wird. Außerdem bleibt der Stichprobenumfang erhalten.

Ist wiederum  $v$  die größte ganze Zahl, für die gilt  $v \leq n \cdot \alpha$ , und  $0 \leq \alpha < 0,5$ , so werden die  $v$  kleinsten Werte gleich  $x_{(v+1)}$  und die  $v$  größten Werte gleich  $x_{(n-v)}$  gesetzt und aus der so korrigierten Reihe der Werte das arithmetische Mittel berechnet:

$$T_n = \bar{x}_{w,\alpha} = \frac{1}{n} \left[ \sum_{i=v+1}^{n-v} x_{(i)} + v(x_{(v+1)} + x_{(n-v)}) \right]. \quad (2.19)$$

Die Gewichte des  $\alpha$ -winsorisierten Mittelwertes, als eine weitere spezielle Klasse der L-Schätzer, sind:

$$w_i = \begin{cases} \frac{1}{n}v & \text{für } i < v+1 \text{ und } i > n-v \\ \frac{1}{n} & \text{für } v+1 \leq i \leq n-v. \end{cases} \quad (2.20)$$

Auch eine Winsorisierung nur auf einer Seite ist möglich.

Der  $\alpha$ -winsorisierte Mittelwert ist unter SPSS nicht implementiert. Man kann ihn nur berechnen, wenn man die Datei in der angegebenen Weise manipuliert.

### Der $\alpha$ -Gastwirth-Cohen-Mittelwert

Der Schätzer dieser speziellen Klasse von L-Schätzern ist definiert als

$$T_n = \lambda(x_{(v+1)} + x_{(n-v)}) + (1 - 2\lambda)x_{0,5} \quad (2.21)$$

Hierin sind  $x_{0,5}$  der Median,  $v$  die größte ganze Zahl mit  $v \leq n \cdot \alpha$  und vorgegebenes  $\alpha$  ( $0 \leq \alpha < 0,5$ ) und  $\lambda$  ein frei wählbarer Parameter mit  $0 \leq \lambda \leq 0,5$ . Die Gewichte des Schätzers (2.21) sind bei geradem  $n$

$$w_i = \begin{cases} \lambda & \text{für } i = v + 1, n - v \\ \frac{1}{2} & \text{für } i = \frac{n}{2}, \frac{n}{2} + 1 \\ 0 & \text{sonst} \end{cases} \quad (2.22)$$

und bei ungeradem  $n$

$$w_i = \begin{cases} \lambda & \text{für } i = v + 1, n - v \\ 1 - 2\lambda & \text{für } i = \frac{n+1}{2} \\ 0 & \text{sonst.} \end{cases} \quad (2.23)$$

Insbesondere erhält man mit  $\lambda = 0,25$  und  $\alpha = 0,25$  (und somit  $v = n/4$ ) das sogenannte trimean, in dessen Berechnung die drei Quartile eingehen;

$$T_n = \frac{1}{4}(x_{(v+1)} + x_{(n-v)}) + \frac{1}{2}x_{0,5} = \frac{1}{4}(x_{0,25} + x_{0,75}) + \frac{1}{2}x_{0,5} \quad (2.24)$$

Der  $\alpha$ -Gastwirth-Cohen-Mittelwert ist ebenfalls nicht in SPSS implementiert. Das trimean lässt sich jedoch leicht nach der Ausgabe der Quartile ermitteln.

- Beispiel 2.5 (Fortsetzung):

Für die Variable Barrel der Datei erdöl.sav werden das arithmetische Mittel (mean), der Median und der 0,05-getrimmte Mittelwert berechnet. Der dafür relevante Teil des SPSS-Outputs ist nachfolgend angegeben.

In die Berechnung des arithmetischen Mittels gehen die beiden Ausreißer Bohrloch 20 mit einem Beobachtungswert von 1328 Mio. Barrel und Bohrloch 8 mit einem Beobachtungswert von

## 2. Entdeckung und Identifikation von Ausreißern

### SPSS-Output 2.5-4: L-Schätzer für das Beispiel 2.5

**Descriptives**

	Statistic	Std. Error
Mio. Barrel Mean	120,721	28,582
5% Trimmed Mean	85,172	
Median	41,500	

775 Mio. Barrel ein, wodurch dieser Mittelwert nach oben gezogen wird.

Für den 0,05-getrimmten Mittelwert ergibt sich

$$\alpha = 0,05, n\alpha = 58 \cdot 0,05 = 2,9, v = 2, g = 0,9$$

und gemäß (2.17):

$$\begin{aligned} T_{n=58} &= \bar{x}_{tr,\alpha=0,05} = \frac{1}{n(1-2\alpha)} \left\{ (1-g)[x_{v+1} + x_{n-v}] + \sum_{i=v+2}^{n-v-1} x_{(i)} \right\} \\ &= \frac{1}{58(1-2 \cdot 0,05)} \{(1-0,9)[6,9 + 482] + 4397,1\} = 85,172 \end{aligned}$$

Durch den Wegfall von 5% der kleinsten und 5% der größten Beobachtungswerte und damit der beiden Ausreißer resultiert für die „mittleren“ 90% der Werte ein arithmetisches Mittel  $\bar{x}_{tr, 0,05}$ , das deutlich kleiner ist als  $\bar{x}$ .

Bei Median wird von der Rangordnung der Beobachtungswerte ausgegangen. Da  $n = 58$  eine gerade Anzahl ist, werden nur die Werte  $x_{(29)} = 41$  und  $x_{(30)} = 42$  mit einem Gewicht von jeweils 0,5 einbezogen:

$$x_{0,5} = (41 + 42)/2 = 41,5$$

50 % der Beobachtungswerte weisen einen Wert auf, der kleiner als  $x_{0,5}$  ist, und 50% haben einen Wert größer als  $x_{0,5}$ .

Für den  $\alpha$ -winsorisierten Mittelwert wird  $v = 2$  und damit  $\alpha = 0,0345$  gewählt. Die zwei kleinsten Beobachtungswerte werden durch den drittkleinsten Wert und die beiden größten Beobachtungswerte durch den drittgrößten Wert ersetzt. Für die so veränderte Stichprobe, die damit weiterhin vom Umfang  $n = 58$  ist, wird das arithmetische Mittel berechnet. Es ergibt sich zu

$$\begin{aligned} T_{n=58} &= \bar{x}_{w,\alpha=0,0345} = \frac{1}{n} \left[ \sum_{i=v+1}^{n-v} x_{(i)} + v(x_{(v+1)} + x_{(n-v)}) \right] \\ &= \frac{1}{58} [4886 + 2(6,9 + 482)] = 101,1 \end{aligned}$$

Durch die Ersetzung der beiden großen Ausreißer durch den drittgrößten Wert bleibt die Richtung der extremen Werte erhalten, was sich in dem winsorisierten Mittelwert (im Vergleich zu den anderen L-Schätzern) niederschlägt.

Läßt man sich die Quartile ausgeben, so resultiert für das trimean gemäß (2.24):

$$T_{n=58} = (16,5 + 116,75)/4 + 41,5/2 = 54,0625.$$

Da der Abstand des dritten Quartils zum Median erheblich größer ist als der Abstand des ersten Quartils zum Median (siehe auch Abb. 2.25), ist das trimean größer als der Median.

## M-Schätzer

Es sei  $X$  eine Zufallsvariable in der Grundgesamtheit mit der Verteilungsfunktion  $F(x)$  und  $\omega$  ein unbekannter Lokalisationsparameter von  $F(x)$ . Aus dieser Grundgesamtheit wird eine Zufallsstichprobe  $X_1, \dots, X_n$  vom Umfang  $n$  gezogen.

Ein M-Schätzer ergibt sich generell als Lösung eines Minimierungsproblems. Ein-M-Schätzer, als eine Schätzfunktion  $T_n$  des Parameters  $\omega$ , minimiert die Funktion

$$\sum_{i=1}^n \theta(X_i, T_n). \quad (2.25)$$

Dabei ist  $\theta$  eine geeignet gewählte Funktion, von der im weiteren angenommen wird, dass sie eine Ableitung bezüglich  $T_n$  über alle reellen Zahlen  $T_n = t$  hat.

Man unterscheidet:

a) nicht-skaleninvariante M-Schätzer

Grundlage dieser M-Schätzer bilden die Abweichungen  $u_i = x_i - T_n$  ( $i = 1, \dots, n$ ). Ein nicht-skaleninvarianter M-Schätzer  $T_n$  minimiert die Summenfunktion

$$\sum_{i=1}^n \theta(x_i - T_n) = \sum_{i=1}^n \theta(u_i) \quad (2.26)$$

für die konkrete Stichprobe  $x_1, \dots, x_n$ . Es ist also derjenige Wert  $T_n = t$  zu finden, für den gilt

$$\sum_{i=1}^n \xi(x_i - T_n) = \sum_{i=1}^n \xi(u_i) = 0, \quad (2.27)$$

mit  $\xi(u) = \partial\theta(u)/\partial u$  als Ableitung von  $\theta(u)$ .

Beispiel:

Wählt man  $\theta(u) = u^2/2$  und somit  $\xi(u) = u$ , so ist wegen  $\sum(x_i - T_n) = 0$  der Schätzwert  $t$  das arithmetische Mittel  $\bar{x}$  und somit der Schätzer  $T_n$  der Stichprobenmittelwert. Dieses Beispiel verdeutlicht gleichzeitig, dass nicht jede Funktion  $\theta(u)$  zu einem robusten Schätzer führt.

## 2. Entdeckung und Identifikation von Ausreißern

### b) skaleninvariante M-Schätzer

Wählt man die relative Abweichung  $z_i = (x_i - T_n)/cS_n$  als Grundlage, so erhält man einen skaleninvarianten M-Schätzer  $T_n$  für die Lokalisation, der die Summenfunktion

$$\sum_{i=1}^n \theta\left(\frac{x_i - T_n}{cS_n}\right) = \sum_{i=1}^n \theta(z_i) \quad (2.28)$$

für die konkrete Stichprobe  $x_1, \dots, x_n$  minimiert. Darin sind

- $S_n$  ein Schätzer für die Streuung aufgrund dieser Stichprobe, für den im allgemeinen der Median der absoluten Abweichungen vom Median (MAD, siehe weiter unten) verwendet wird,
- $c$  eine positive Konstante, die oftmals eine Relation zur Standardabweichung der Standardnormalverteilung darstellt.

In diesem Fall gilt für die Ableitung  $\xi(z) = \partial\theta(z)/\partial z$

$$\sum_{i=1}^n \xi\left(\frac{x_i - T_n}{cS_n}\right) = \sum_{i=1}^n \xi(z_i) = 0. \quad (2.29)$$

Bei M-Schätzern werden die Ausreißer nicht grundsätzlich ausgeschlossen, ihnen werden aber geringere Gewichte beigemessen als den übrigen Werten. Da die Festlegung der Funktion  $\theta$  und damit die Zuweisung von Gewichten auf verschiedene Art und Weise erfolgen kann, gibt es verschiedene M-Schätzer. Allgemein erfolgt die Gewichtszuweisung jedoch so, dass die Gewichte kleiner werden, je größer die Abweichung von einem Lokalisationsparameter der Verteilung wird. Die Gewichtsfunktion lautet allgemein:

$$w(z) = \xi(z)/z. \quad (2.30)$$

Es sollen nun die für SPSS relevanten und wohl auch bekanntesten M-Schätzer diskutiert werden. Sie sind skaleninvariante Schätzer, die die Summenfunktion (2.28) minimieren.

#### Huber-k-Schätzer

Für den Huber-k-Schätzer gilt:

$$\theta(z) = \begin{cases} \frac{z^2}{2} & \text{für } |z| \leq k \\ k|z| - \frac{k^2}{2} & \text{für } |z| > k \end{cases} \quad (2.31)$$

und

$$\xi(z) = \begin{cases} z & \text{für } |z| \leq k \\ k \operatorname{sgn}(z) & \text{für } |z| > k. \end{cases} \quad (2.32)$$

$\text{sgn}$  ist die Vorzeichenfunktion Signum mit

$$\text{sgn}(z) = \begin{cases} 1 & \text{für } z > 0 \\ 0 & \text{für } z = 0 \\ -1 & \text{für } z < 0. \end{cases} \quad (2.33)$$

$k$  ist eine positive, reelle Konstante als Punkt, von dem an sich die Gestalt der Funktion verändert. Im Bereich  $-k \leq z \leq k$  ist  $\theta(z)$  quadratisch und  $\xi(z)$  linear, im Bereich  $|z| > k$  ist  $\theta(z)$  linear und  $\xi(z)$  konstant. Für  $k = 0$  ist  $\theta(z)$  insgesamt linear und für  $k = \infty$  quadratisch. Die Wahl von  $k$  hängt vom Grad der „Verschmutzung“ der Beobachtungsdaten ab (d.h. dem Teil, der nicht einer Normalverteilung entspricht). Für große Stichproben und einem Verschmutzungsgrad zwischen 0,1 und 0,01 der Beobachtungswerte sollte  $1,140 < k < 1,945$  gewählt werden. Der voreingestellte Wert unter SPSS ist  $k = 1,339$ . Die Abbildungen 2.26 und 2.27 zeigen diese Funktion für  $k=1$ .

Abbildung 2.26.:  $\theta(z)$  für den Huber-1-Schätzer

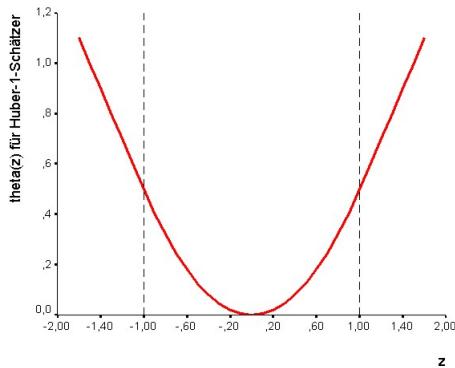
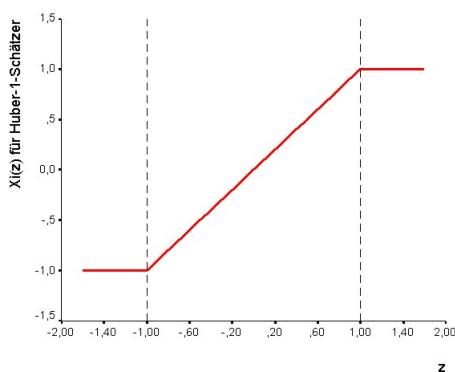


Abbildung 2.27.:  $\xi(z)$  für den Huber-1-Schätzer



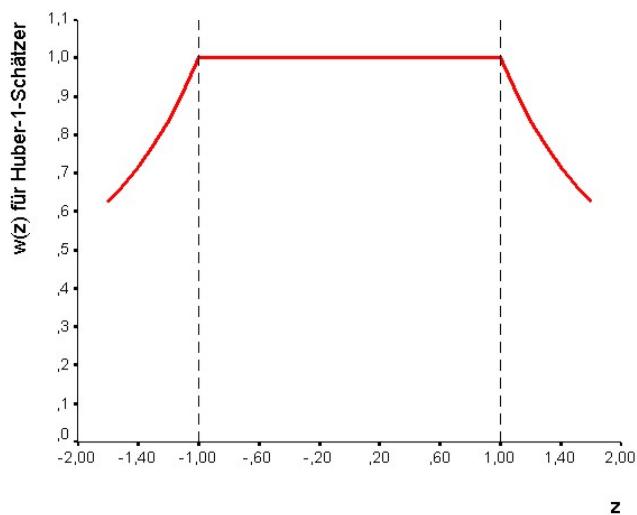
## 2. Entdeckung und Identifikation von Ausreißern

Die Gewichtsfunktion  $w(z)$  des Huber-k-Schätzers ergibt sich nach (2.30) wie folgt:

$$w(z) = \begin{cases} \frac{z}{|z|} = 1 & \text{für } |z| \leq k \\ \frac{k \operatorname{sgn}(z)}{|z|} & \text{für } |z| > k. \end{cases} \quad (2.34)$$

Die Abb. (2.28) zeigt diese Gewichtsfunktion  $w(z)$  für den Huber-1-Schätzer.

Abbildung 2.28.: Gewichtsfunktion  $w(z)$  des Huber-1-Schätzers



Je nach der Wahl von  $k$  ist der Bereich größer oder kleiner, der eine konstante Gewichtung von 1 aufweist. Da die Gewichte auch bei großem Abstand vom Schätzwert nicht auf Null abfallen, bleibt der Einfluß extremer Beobachtungswerte in gewissem Maße erhalten.

### Hampel-Schätzer

Die für den Hampel-Schätzer gewählte Funktion  $\theta(z)$  ist in Formel (2.35) und ihre Ableitung  $\xi(z)$  in Formel (2.36) angegeben. Die darin enthaltenen Konstanten haben unter SPSS die voreingestellten Werten  $a = 1,7$ ,  $b = 3,4$  und  $c = 8,5$ . Die Abbildungen 2.29 und 2.30 zeigen diese beiden Funktionen.

Bei diesem Schätzer erhalten die  $z$ -Werte, für die  $-a \leq z \leq a$  gilt, ein Gewicht von 1. Für  $|z|$ , die zwischen  $a$  und  $b$  liegen, resultiert ein Gewicht von  $a/z$ . Für  $|z|$ , die zwischen  $b$  und  $c$  liegen, ergibt sich ein Gewicht von  $[a(c-z)]/[z(c-b)]$  und für  $|z|$ , die größer sind als  $c$ , ein Gewicht von 0 (d.h., sie werden weggelassen). Die Gewichtsfunktion  $w(z)$  ist in Formel (2.37)

und Abb. 2.31 gegeben.

$$\theta(z) = \begin{cases} \frac{z^2}{2} & \text{für } |z| \leq a \\ a|z| - \frac{a^2}{2} & \text{für } a < |z| \leq b \\ ab - \frac{a^2}{2} + (c-b)\frac{a}{2} \left[ 1 - \left( \frac{c-|z|}{c-b} \right)^2 \right] & \text{für } b < |z| \leq c \\ ab - \frac{a^2}{2} + (c-b)\frac{a}{2} & \text{für } |z| > c \end{cases} \quad (2.35)$$

$$\xi(z) = \begin{cases} z & \text{für } |z| \leq a \\ a \operatorname{sgn}(z) & \text{für } a < |z| \leq b \\ a \frac{c-|z|}{c-b} \operatorname{sgn}(z) & \text{für } b < |z| \leq c \\ 0 & \text{für } |z| > c \end{cases} \quad (2.36)$$

$$w(z) = \begin{cases} 1 & \text{für } |z| \leq a \\ \frac{a \operatorname{sgn}(z)}{|z|} & \text{für } a < |z| \leq b \\ a \frac{c-|z|}{c-b} \frac{\operatorname{sgn}(z)}{|z|} & \text{für } b < |z| \leq c \\ 0 & \text{für } |z| > c \end{cases} \quad (2.37)$$

Die Konstante  $a$  wird im allgemeinen wie Hubers  $k$  gewählt, d.h. entsprechend dem Verschmutzungsgrad der Beobachtungswerte.  $b$  sollte so bestimmt werden, dass im Mittel nur wenige Beobachtungen außerhalb des Bereiches  $(-b, b)$  für  $z$  liegen, und für  $c$  sollte gelten:  $c - b \geq 2a$ . Für Stichprobenverteilungen mit stärker besetzten Enden sollten die Konstanten entsprechend kleiner gewählt werden.

## 2. Entdeckung und Identifikation von Ausreißern

Abbildung 2.29.:  $\theta(z)$  des Hampel-Schätzers für  $a = 1,7$ ,  $b = 3,4$  und  $c = 8,5$

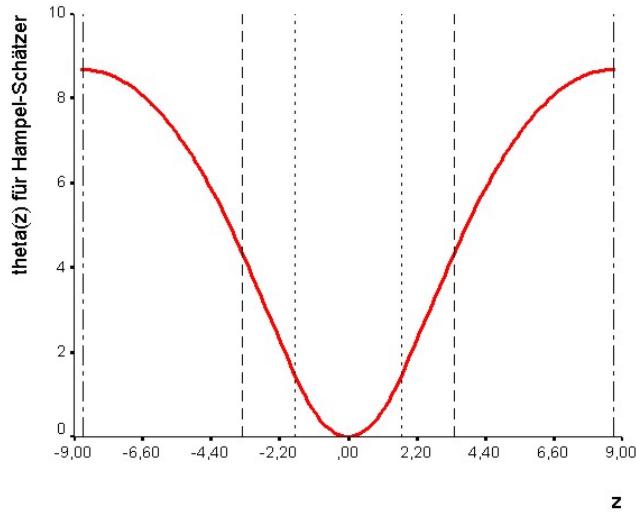


Abbildung 2.30.:  $\xi(z)$  des Hampel-Schätzers für  $a = 1,7$ ,  $b = 3,4$  und  $c = 8,5$

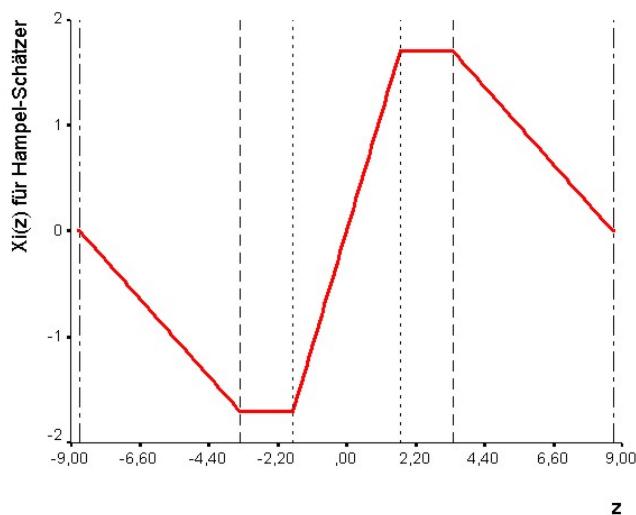
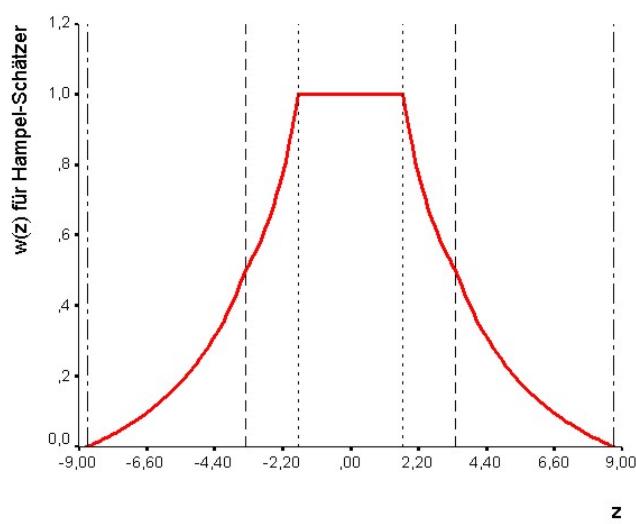


Abbildung 2.31.: Gewichtsfunktion  $w(z)$  des Hampel-Schätzers für  $a = 1,7$ ,  $b = 3,4$  und  $c = 8,5$



Andrews wave

Für diesen M-Schätzer gilt

$$\theta(z) = \begin{cases} \frac{a^2}{\pi^2} \cdot (1 - \cos \frac{\pi z}{a}) & \text{für } |z| \leq a \\ \frac{2a^2}{\pi^2} & \text{für } |z| > a \end{cases} \quad (2.38)$$

und

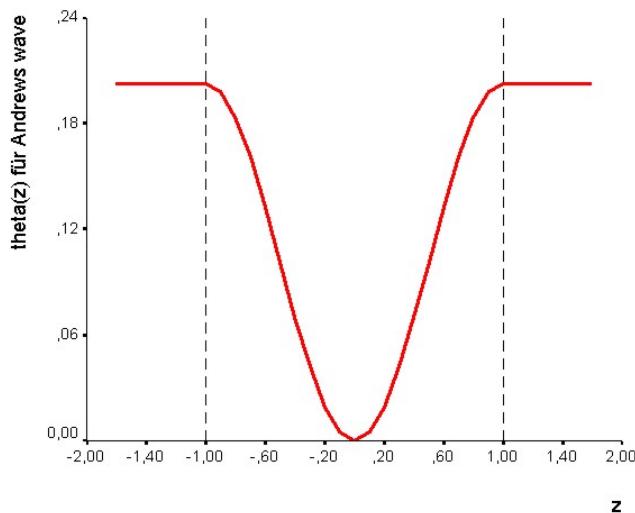
$$\xi(z) = \begin{cases} \frac{a}{\pi} \cdot \sin \frac{\pi z}{a} & \text{für } |z| \leq a \\ 0 & \text{für } |z| > a. \end{cases} \quad (2.39)$$

a ist eine Konstante mit  $a > 0$ . Unter SPSS ist sie mit  $a = 1,34\pi$  voreingestellt. Die Gewichtsfunktion  $w(z)$  des Andrew-Schätzers ist wie folgt:

$$w(z) = \begin{cases} \frac{a}{\pi z} \cdot \sin \frac{\pi z}{a} & \text{für } |z| \leq a \\ 0 & \text{für } |z| > a. \end{cases} \quad (2.40)$$

$w(z)$  ist für  $z = 0$  nicht definiert.

Abbildung 2.32.:  $\theta(z)$  des Andrews-Schätzers für  $a = 1$



## 2. Entdeckung und Identifikation von Ausreißern

Abbildung 2.33.:  $\xi(z)$  des Andrews-Schätzers für  $a = 1$

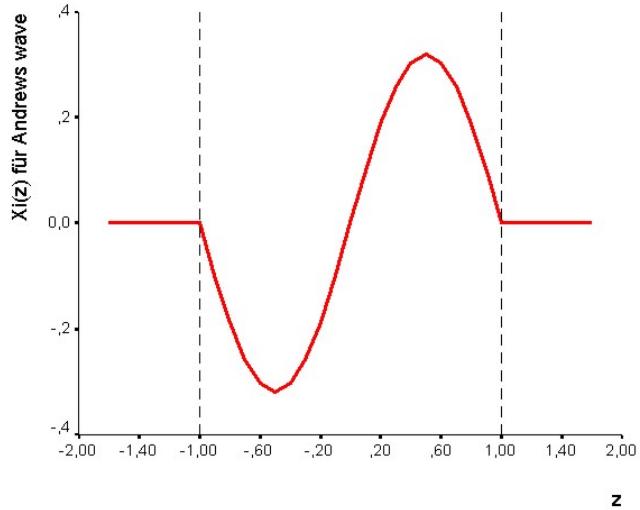
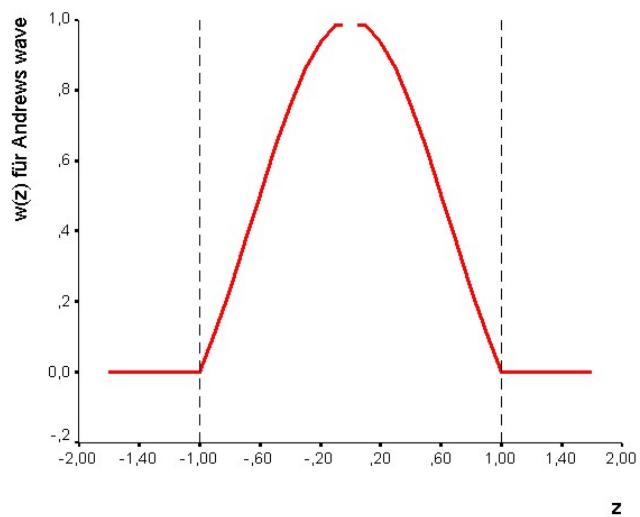


Abbildung 2.34.: Gewichtsfunktion  $w(z)$  des Andrews-Schätzers für  $a = 1$



### Tukey's biweight

Für diesen Schätzer gilt

$$\theta(z) = \begin{cases} \frac{a^2}{6} \cdot \left(1 - \left(1 - \frac{z^2}{a^2}\right)^3\right) & \text{für } |z| \leq a \\ \frac{a^2}{6} & \text{für } |z| > a \end{cases} \quad (2.41)$$

und

$$\xi(z) = \begin{cases} z \left(1 - \frac{z^2}{a^2}\right)^2 & \text{für } |z| \leq a \\ 0 & \text{für } |z| > a. \end{cases} \quad (2.42)$$

$a$  ist eine Konstante mit  $a > 0$ . Unter SPSS ist sie mit  $a = 4,685$  voreingestellt. Die Abbildungen (2.35) und (2.36) zeigen diese Funktion für  $a = 1$ .

Ausreißer beeinflussen diesen Schätzer nicht, da  $\xi(z) = 0$  gilt, wenn  $|z|$  genügend groß ist. Die Gewichtsfunktion  $w(z)$  enthält Formel (2.43) und Abb. 2.37.

$$w(z) = \begin{cases} \left(1 - \frac{z^2}{a^2}\right)^2 & \text{für } |z| \leq a \\ 0 & \text{für } |z| > a. \end{cases} \quad (2.43)$$

Abbildung 2.35.: Funktion  $\theta(z)$  des Tukey's biweight für  $a = 1$

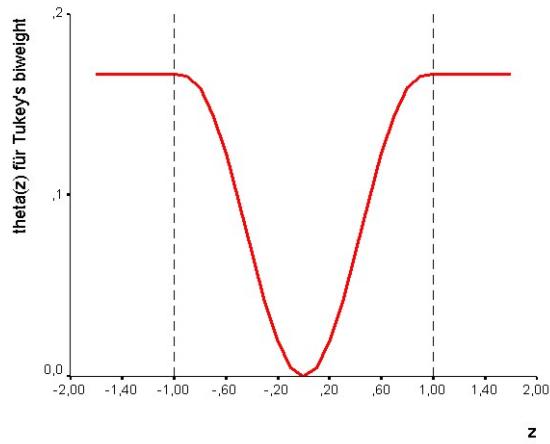
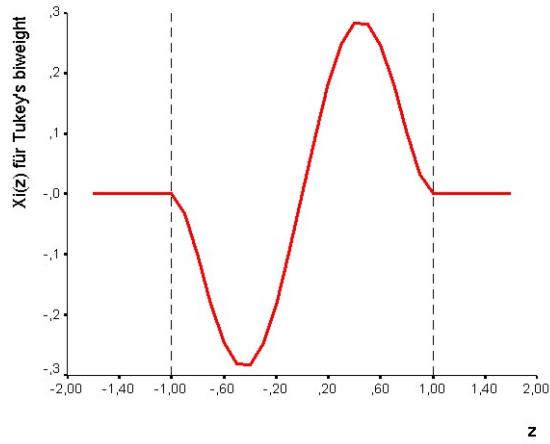
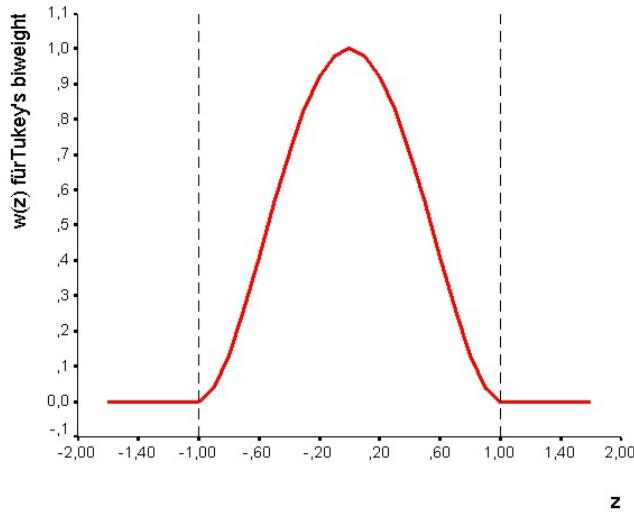


Abbildung 2.36.: Funktion  $\xi(z)$  des Tukey's biweight für  $a = 1$



## 2. Entdeckung und Identifikation von Ausreißern

Abbildung 2.37.: Gewichtsfunktion  $w(z)$  des Tukey-Schätzers für  $a = 1$



Hampel-, Andrews- und Tukey-Schätzer sind dadurch gekennzeichnet, daß die potentiellen Ausreißer (extremen Werte) keinen Einfluß mehr ausüben, da sie weggelassen werden, denn die Funktion  $\xi(z)$  verschwindet außerhalb des endlichen Bereiches. Solche Schätzer werden redescending M-Estimators genannt.

Die Berechnung der M-Schätzer kann nur in einem Iterationsprozeß erfolgen, da der unbekannte Schätzer  $T_n$  in der Funktion  $\theta(z)$  enthalten ist. Der Iterationsprozeß beginnt mit einem vorgegebenen Startwert  $T_n(0)$ , der im allgemeinen (auch in SPSS) der Median  $x_{0,5}$  der Ausgangswerte  $x_i$  ist, also  $T_n(0) = x_{0,5}$ .

Bei der gewichteten Schätzung (W-estimation, iteratively reweighted least-squares, IRLS) werden die z-Werte in jedem Iterationsschritt  $r$  ( $r = 0, 1, 2, \dots$ ) nach

$$z_i(r) = \frac{x_i - T_n(r)}{cS_n}, \quad i = 1, \dots, n \quad (2.44)$$

und die Gewichte  $w(z)$  nach  $w(z) = \xi(z)/z$  neu bestimmt. Der Schätzwert  $T_n(r)$  im r-ten Schritt ( $r = 1, 2, \dots$ ) berechnet sich als gewogenes Mittel gemäß

$$T_n(r) = \frac{\sum_{i=1}^n w(z_i(r-1))x_i}{\sum_{i=1}^n w(z_i(r-1))}. \quad (2.45)$$

Das Verfahren wird abgebrochen, wenn die Differenz zwischen  $T_n(r-1)$  und  $T_n(r)$  kleiner ist als eine vorgegebene kleine Zahl  $\varepsilon > 0$ .

Diese gewichtete Schätzung wird auch in SPSS verwendet. Der Iterationsprozeß wird abgebrochen, wenn  $|T(r+1) - T(r)| \leq \varepsilon[(T(r+1) + T(r))/2]$  mit  $\varepsilon = 0,005$  erfüllt ist oder wenn die Anzahl der Iterationen größer als 30 wird.

Werden die M-Schätzer dagegen direkt berechnet, wird das Newton-Raphson-Verfahren verwendet, bei dem im r-ten Iterationsschritt

$$T_n(r+1) = T_n(r) + cS_n \frac{\sum_{i=1}^n \xi(z_i(r))}{\sum_{i=1}^n \xi'(z_i(r))} \quad (2.46)$$

berechnet wird.  $z_i(r)$  wird entsprechend (2.44) ermittelt. Auch hierbei wird das Verfahren abgebrochen, wenn  $|T_n(r) - T_n(r+1)| < \varepsilon$  ist.

Unter SPSS können M-Schätzer unter

- Analyze
  - Descriptive Statistics
  - Explore...

angefordert werden. Nach Betätigung der Schaltfläche „Statistics“ im Dialogfeld „Explore“ (siehe Abb. 2.1) wird im Dialogfeld „Explore: Statistics“ (siehe Abb. 2.3) auf M-estimators entschieden.

- Beispiel 2.5 (Fortsetzung):

Für die Variable Barrel der Datei erdöl.sav werden die unter SPSS verfügbaren M-Schätzer berechnet. Der dafür relevante Teil des SPSS-Outputs ist nachfolgend angegeben.

**SPSS-Output 2.5-5: M-Schätzer für das Beispiel 2.5**

**M-Estimators**

	Huber's M-Estimator <sup>a</sup>	Tukey's Biweight <sup>b</sup>	Hampel's M-Estimator <sup>c</sup>	Andrews' Wave <sup>d</sup>
Mio. Barrel	50,000	36,221	44,400	35,843

*a.* The weighting constant is 1,339.

*b.* The weighting constant is 4,685.

*c.* The weighting constants are 1,700, 3,400, and 8,50.

*d.* The weighting constant is 1,340\*pi.

Da beim Huber-k-Schätzer die Gewichte nicht auf Null abfallen, werden auch die beiden Ausreißer (Bohrlöcher 20 und 8) mit gewissen Gewichten versehen, so dass dieser M-Schätzer im Vergleich zu den anderen am größten ist. Die in Etappen erfolgende Gewichtung beim Hampel-Schätzer schlägt sich ebenfalls in seiner Größe nieder. Andrews wave und Tukey's biweight liefern sehr ähnliche Schätzwerte.

## 2. Entdeckung und Identifikation von Ausreißern

- Beispiel 2.2 (Fortsetzung):

Für die Variable einkomp1 (monatliches Nettoeinkommen in DM) der Datei allbus.sav werden neben dem arithmetischen Mittel robuste Schätzer unter SPSS berechnet. Es werden nur die dafür relevanten Teile des Outputs wiedergegeben.

Die Variable einkomp1 (monatliches Nettoeinkommen in DM) weist einen extremen Wert auf: Die befragte Person mit der ID-Nummer 245 hat ein monatliches Nettoeinkommen von 15000 DM angegeben (siehe auch Abb. 2.7). Der nächste Beobachtungswert liegt bei 7000 DM.

Da nur für den David-Hartley-Pearson-Test für die Anzahl der gültigen Fälle von 716 näherungsweise kritische Werte zu finden sind, wird dieser Ausreißertest auf dem Signifikanzniveau von  $\alpha = 0,05$  durchgeführt, wobei die Nullhypothese formuliert als:

$H_0 : x_{(716)}$  ist kein Ausreißer. Für den Wert der Teststatistik resultiert:

$$T = 14880/1150,01 = 12,938.$$

**SPSS-Output 2.2-3:** Robuste Schätzer für das Beispiel 2.2

Descriptives

		Statistic	Std. Error
Monatl. Nettoeinkommen in DM	Mean	1881,40	42,98
	5% Trimmed Mean	1799,88	
	Median	1800,00	
	Std. Deviation	1150,01	
	Minimum	120	
	Maximum	15000	
	Range	14880	

M-Estimators

	Huber's M-Estimator <sup>a</sup>	Tukey's Biweight <sup>b</sup>	Hampel's M-Estimator <sup>c</sup>	Andrews' Wave <sup>d</sup>
Monatl. Nettoeinkommen in DM	1763,68	1720,20	1758,68	1720,18

a. The weighting constant is 1,339.

b. The weighting constant is 4,685.

c. The weighting constants are 1,700, 3,400, and 8,50.

d. The weighting constant is 1,340\*pi.

Bei David/Hartley/Pearson (1954), S. 491, findet man für den Stichprobenumfang  $n = 716$  keine kritischen Werte. Es werden deshalb diejenigen für  $n = 500$  verwendet:

$$Q_{L;500;0,05} = 5,37 \text{ und } Q_{U;500;0,05} = 6,94.$$

Die  $H_0$  wird auf dem 5%-Niveau abgelehnt. Der größte Wert  $x_{(716)}$  kann als ein Ausreißer angesehen werden. Die Interpretation der robusten Schätzer ist analog zum Beispiel 2.5 zu führen. Es wäre also zu hinterfragen, ob es sich möglicherweise bei dem Beobachtungswert von 15000 DM um einen Eingabefehler handelt.

## **Robuste Schätzer für die Streuung**

**Mittlere absolute Abweichung vom arithmetischen Mittel:**

$$d(\bar{x}) = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| \quad (2.47)$$

Für die Stichproben aus einer normalverteilten Grundgesamtheit und  $n \rightarrow \infty$  ist  $E(d(\bar{x})) = \sigma(\pi/2)^{0,5}$ .

**Mittlere absolute Abweichung vom Median:**

$$d(x_{0,5}) = \frac{1}{n} \sum_{i=1}^n |x_i - x_{0,5}| \quad (2.48)$$

**Median der absoluten Abweichungen vom Median** (median absolute deviation, MAD):

$$MAD = \text{median}(|x_i - x_{0,5}|) \quad (2.49)$$

Der Erwartungswert  $E(MAD)$  ist bei einer  $N(\mu; \sigma^2)$ -verteilten Grundgesamtheit und großen Stichproben approximativ gleich  $0,6745\sigma$ , so dass  $\hat{\sigma} = MAD/0,6745 = 1,4826 \cdot MAD$  als Schätzung für die Standardabweichung  $\sigma$  einer normalverteilten Grundgesamtheit verwendet wird. Da MAD vielfach bei den M-Schätzern für  $S_n$  verwendet wird, kann eine Schätzung der Gewichtungsbereiche erfolgen:

$$c \cdot S_n = c \cdot MAD = c \cdot 0,6745 \cdot \hat{\sigma}.$$

Wählt man z.B. für Tukey's biweight  $S_n = MAD$  und  $c = 4,685$  in (2.28), so bedeutet dies, dass Beobachtungswerte, die mehr als  $4,685 \cdot 0,6745 \cdot \hat{\sigma} \approx 3$  Standardabweichungen vom Median entfernt liegen, ein Gewicht von Null erhalten.

**Interquartilsabstand** (interquartile range, IQR):

$$IQR = x_{0,75} - x_{0,25} \quad (2.50)$$

Der IQR ist bei einer  $N(\mu; \sigma^2)$ -verteilten Grundgesamtheit gleich  $1,35\sigma$ , so dass

$$IQR/1,35 = 0,741 \cdot IQR$$

als Schätzung für die Standardabweichung  $\sigma$  einer normalverteilten Grundgesamtheit verwendet wird.

## 2. Entdeckung und Identifikation von Ausreißern

### $\alpha$ -getrimmte Varianz bzw. $\alpha$ -getrimmte absolute Abweichung:

Unter Verwendung des  $\alpha$ -getrimmten Mittelwertes  $\bar{x}_{tr;\alpha}$  aus (2.14) erhält man

$$s_{tr;\alpha}^2 = \frac{1}{n - 2v} \sum_{i=v+1}^{n-v} (x_{(i)} - \bar{x}_{tr;\alpha})^2 \quad (2.51)$$

bzw.

$$d_{tr;\alpha} = \frac{1}{n - 2v} \sum_{i=v+1}^{n-v} |x_{(i)} - \bar{x}_{tr;\alpha}| \quad (2.52)$$

Unter SPSS ist von diesen robusten Schätzern für die Streuung nur der Interquartilsabstand IQR verfügbar. Der MAD wird zwar für die M-Schätzer verwendet, aber nicht ausgegeben.

Wählt man im Dialogfeld „Explore: Statistics“ (siehe Abb. 2.3) die Option Descriptives, so wird darunter auch der IQR ausgegeben.

**Ausgewählte Literatur zum Ausreißerproblem und zur Robustheit:**

- Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W. (1972)
- Barnett, V., Lewis, T. (1994)
- Büning, H., Trenkler, G. (1978), S. 296 ff.
- Büning, H. (1991)
- David, H.A., Hartley, H.O., Pearson, E.S. (1954)
- Dixon, W.J. (1951)
- du Toit, S.H.C., Steyn, A.G.W., Stumpf, R.H. (1986)
- Ferguson, Th.s. (1961)
- Gastwirth, J.L., Cohen, M.L. (1970)
- Grubbs, F.E. (1950)
- Grubbs, F.E., Beck, G. (1972)
- Hartung, Elpelt, Klösener (1993), S. 343 - 349, 861 - 886
- Hawkins, D.M. (1980)
- Heiler, S., Michels, P. (1994), S. 100 ff., 110 ff.
- Hoaglin, D.C., Mosteller, F., Tukey, J.W. (1983)
- Huber, P.J. (1981)
- Jobson, J.D. (1991), S. 56 ff.
- Kinnison (1985)
- Launer, R.L., Wilkinson, G.N. (Hrsg.) (1979)
- Läuter, H., Pincus, R. (1989), S. 349 - 357
- Pearson, E.S., Hartley, H.O. (1970, 1972)
- Sachs, L. (1992), S. 363 ff.
- Staudte, R.G., Sheather, S.J. (1990)

## *2. Entdeckung und Identifikation von Ausreißern*

### **3. Prüfung der Verteilungform von Variablen**

Die Frage nach der Form der empirischen Häufigkeitsverteilung einer Variablen ist ein weiterer Schritt sorgfältiger Datenanalyse auf dem Weg zur statistischen Modellbildung und beinhaltet vor allem zwei Aspekte:

- Zum einen soll sie eine effektive Präsentation der Daten ermöglichen, die eine sachgerechte Interpretation erleichtert und gleichzeitig die Sensibilität gegenüber unerwarteten und ungewöhnlichen Beobachtungen erhöht.

Das folgende Beispiel zeigt, dass eine sinnvolle Präsentation der Häufigkeitsverteilung langatmige Erläuterungen ersetzen kann.

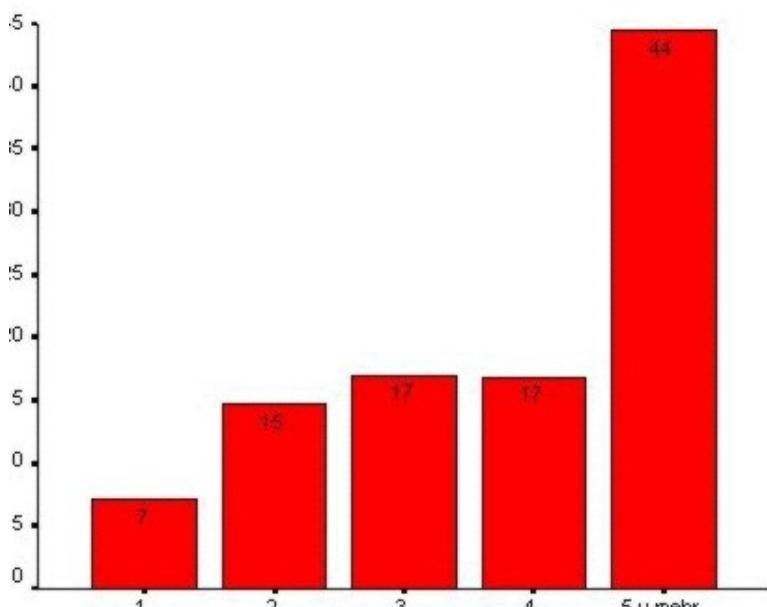
- Zum anderen setzt die Mehrzahl statistischer Verfahren bestimmte (theoretische) Verteilungen voraus, so dass vor ihrer Anwendung überprüft werden muss, ob diese Verteilung für die betrachtete Variable auch gegeben ist. In diesem Zusammenhang spielt bei metrisch skalierten Variablen die Normalverteilung eine ganz entscheidende Rolle, da viele statistische Verfahren zumindest annähernd eine normalverteilte Grundgesamtheit voraussetzen. Viele sozialwissenschaftliche Variablen weichen jedoch mehr oder weniger stark von einer Normalverteilung ab, was die Relevanz der Prüfung der Verteilung unterstreicht.

Informationen über die Verteilungsform einer Variablen können mittels der explorativen, deskriptiven oder konfirmatorischen Datenanalyse gewonnen werden. Wenn diese drei Möglichkeiten nachfolgend getrennt dargestellt werden, so wird man sie bei praktischen Untersuchungen jedoch oftmals komplex anwenden.

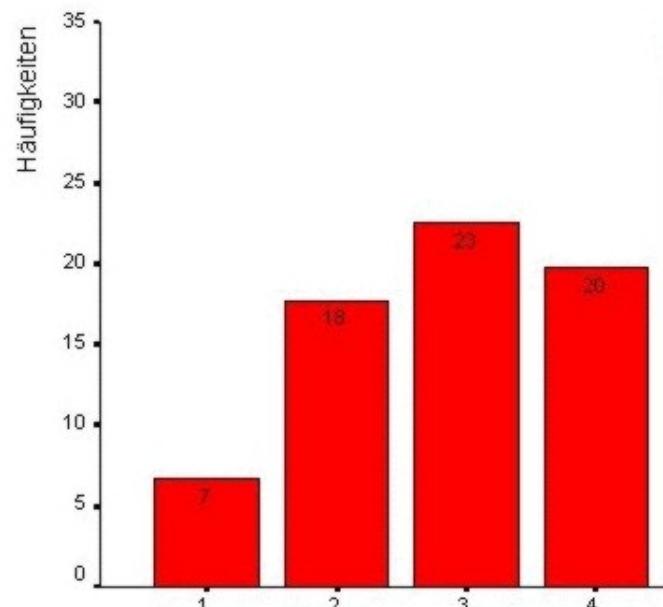
Ein entscheidendes Kriterium für die Auswahl von statistischen Verfahren zur Prüfung der Verteilungsform ist das jeweilige Skalenniveau der Variablen.

### 3. Prüfung der Verteilungform von Variablen

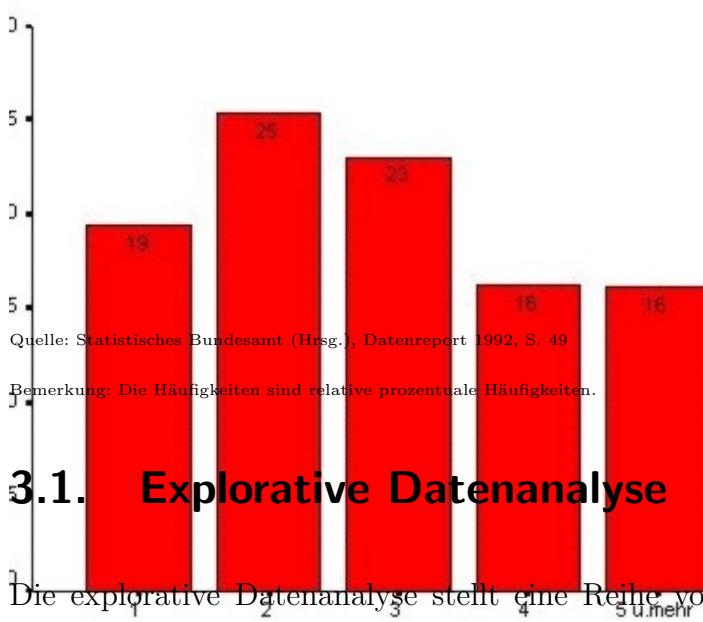
Abbildung 3.1.: Haushaltsgrößen im früheren Bundesgebiet



Haushaltsgröße 1900

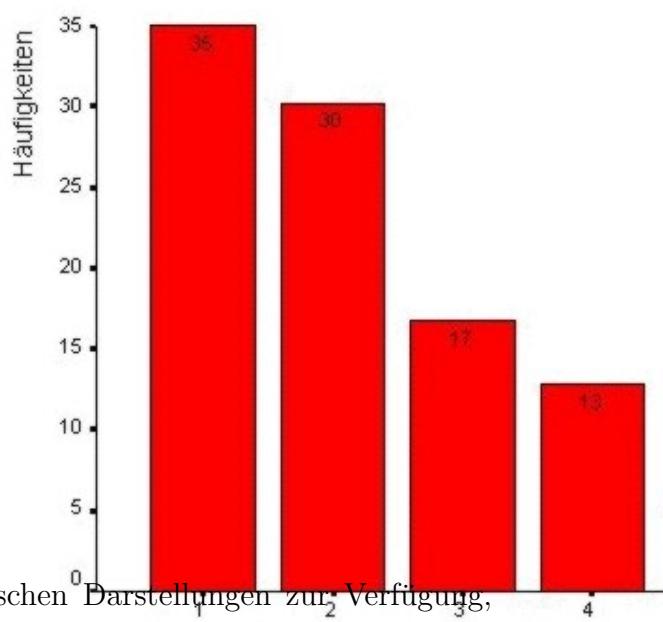


Haushaltsgröße 1925



## 3.1. Explorative Datenanalyse

Die explorative Datenanalyse stellt eine Reihe von grafischen Darstellungen zur Verfügung, über die ein erster visueller Eindruck über die Verteilungsform einer Variablen gewonnen und davon ausgehend Hypothesen über die Verteilungsform formuliert werden können.



## Balkendiagramme (bar charts)

Balkendiagramme eignen sich zur Darstellung der empirischen Häufigkeitsverteilung vor allem von nominalskalierten und ordinalskalierten Variablen, aber auch von metrisch skalierten diskreten Variablen mit wenigen Variablenausprägungen (Faustregel:  $n < 10$ ) bzw. klassierten Variablen.

### Einfaches Balkendiagramm (simple bar chart)

Ein einfaches Balkendiagramm stellt die Häufigkeit jeder Variablenausprägung einer Variablen X als separaten Balken dar. Es kann unter SPSS über zwei Möglichkeiten angefordert werden, in denen wiederum zwei Versionen existieren.

#### 1. Möglichkeit

##### ■ Graphs

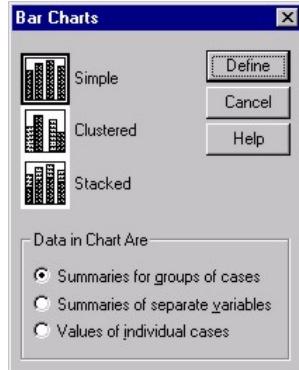
###### ■ Bar...

Version 1:

Man gelangt in das Dialogfeld „Bar Charts“ (siehe Abb. 3.2), in dem die Voreinstellung für „Simple“ (einfaches Balkendiagramm) und „Summaries for groups of cases“ belassen werden.

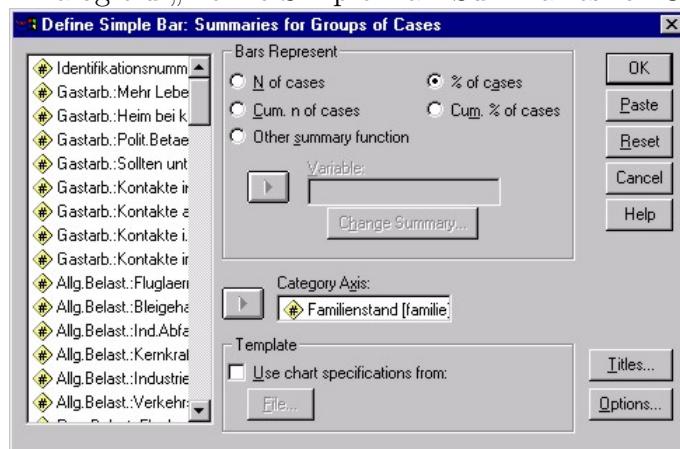
### 3. Prüfung der Verteilungform von Variablen

Abbildung 3.2.: Dialogfeld „Bar Charts“



Summaries for groups of cases erzeugt einen Balken für jede Kategorie einer kategorialen Variablen, d.h., in diesem Falle werden die Häufigkeiten erst ausgezählt. Über die Schaltfläche „Define“ gelangt man in das Dialogfeld „Define Simple Bar: Summaries for Groups of Cases“.

Abbildung 3.3.: Dialogfeld „Define Simple Bar: Summaries for Groups of Cases“



Zunächst wird die gewünschte Variable aus der linken Variablenliste in das Feld „Category Axis:“ (Abszissenachse) gebracht. Im Feld „Bars Represent“ kann man entscheiden, ob auf der Ordinatenachse (in SPSS als Skalenachse bezeichnet) N of cases (Anzahl der Fälle, absolute Häufigkeiten) oder % of cases (relative prozentuale Häufigkeiten) abgetragen werden sollen.

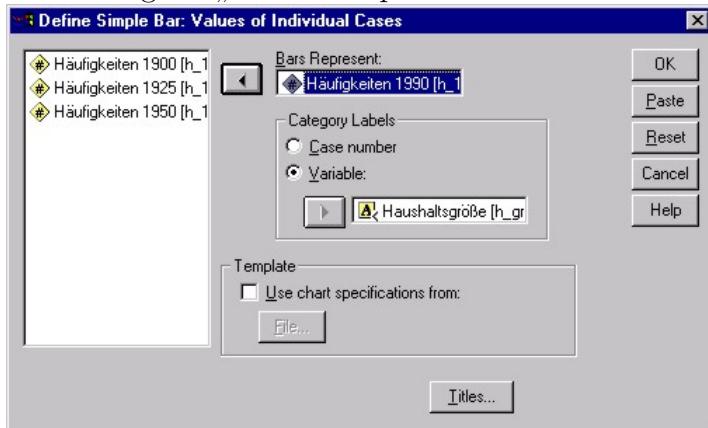
Da Missings in dem Balkendiagramm nicht als gesonderte Kategorie erscheinen sollen, geht man über die Schaltfläche „Options...“ in das Dialogfeld „Options“ und entfernt durch Anklicken das Kreuz vor „Display groups defined by missing values“.

Version 2:

Sind für eine Variable die absoluten oder relativen Häufigkeiten der einzelnen Variablenausprägungen schon in einer anderen Variablen der gleichen Datei enthalten, dann muß in dem

Dialogfeld „Bar Charts“ (Abb. 3.2) unter „Data in Chart Are“ auf „Values of individual cases“ entschieden werden. In dem über „Define“ folgenden Dialogfeld (Abb. 3.4) ist „Variable“ unter Category Labels anzuklicken und die kategoriale Variable in das entsprechende Feld zu bringen. Die Variable, die die Häufigkeiten enthält, kommt nach „Bar Represent“.

Abbildung 3.4.: Dialogfeld „Define Simple Bar: Values of Individual Cases“



## 2. Möglichkeit

Version 1:

Bei Variablen, für die die Beobachtungswerte jedes einzelnen Falles in der Datei enthalten sind und deren (absoluten bzw. relativen) Häufigkeiten erst bei der Erstellung des Balkendiagramms ermittelt werden, kann der Aufruf sofort erfolgen.

Version 2:

Sind für eine Variable die absoluten oder relativen Häufigkeiten der einzelnen Variablenausprägungen in einer anderen Variablen der gleichen Datei enthalten, dann muss vor dem Aufruf für die Erstellung eines Balkendiagramms eine Gewichtung der Fälle mit der Gewichtsvariablen erfolgen (siehe 1. Kapitel).

Für beide Versionen wird zur Erstellung des Balkendiagramms wie folgt weiter verfahren: Aufruf von

■ Analyze

■ Descriptive Statistics

■ Frequencies...

In dem Dialogfeld „Frequencies“ (Abb. 3.5) wird die gewünschte Variable ausgewählt und in das Dialogfeld „Variable(s):“ gebracht.

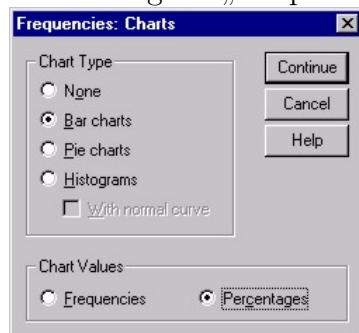
Über die Schaltfläche „Charts...“ kann man in dem Dialogfeld „Frequencies: Charts“ (Abb. 3.6) auf die Ausgabe eines Bar Chart mit absoluten Häufigkeiten (Frequencies) oder mit relativen Häufigkeiten (Percentages) entscheiden.

### 3. Prüfung der Verteilungform von Variablen

Abbildung 3.5.: Dialogfeld „Frequencies“



Abbildung 3.6.: Dialogfeld „Frequencies: Charts“



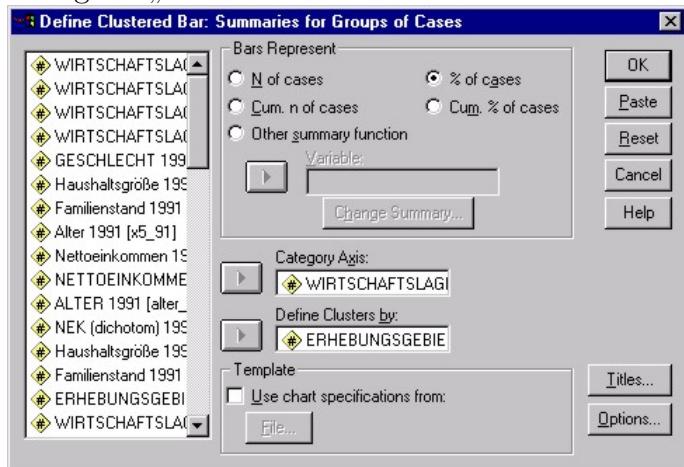
Ein Beispiel für einfache Balkendiagramme mit relativen prozentualen Häufigkeiten zeigt Abb. 3.1.

### Gruppiertes Balkendiagramm (clustered bar chart)

Bei vielen praktischen statistischen Auswertungen ist von Interesse, die Häufigkeitsverteilung einer Variablen Y separiert nach den Ausprägungen einer Gruppierungs- oder Faktorvariablen X zu untersuchen. Es ergeben sich in diesem Fall bedingte Häufigkeitsverteilungen. Die relative Häufigkeit dafür, daß sich die Variable Y zu einer bestimmten Ausprägung  $y_i$  realisiert, unter der Bedingung, dass die Variable X die Ausprägung  $x_j$  angenommen hat, ist  $f(Y = y_i | X = x_j)$ .

Zur graphischen Darstellung bedingter Häufigkeitsverteilungen kann ein gruppiertes Balkendiagramm verwendet werden. Es erzeugt für jede Ausprägung der Variablen Y eine Gruppe von Balken entsprechend den Ausprägungen der Gruppierungsvariablen X. Das gruppierte Balkendiagramm erhält man, indem man im Dialogfeld „Bar Charts“ (siehe Abb. 3.2) auf „Clustered“ entscheidet, die Voreinstellung „Summaries for groups of cases“ beläßt und die Schaltfläche „Define“ betätigt.

Abbildung 3.7.: Dialogfeld „Define Clustered Bar: Summaries for Groups of Cases“



Im Dialogfeld „Define Clustered Bar: Summaries for Groups of Cases“ wird diejenige Variable (Y), für deren Ausprägungen die Gruppen von Balken erzeugt werden sollen, aus der linken Quellliste in das Feld „Category Axis:“ und die Gruppierungsvariable (X) in das Feld „Define Clusters by:“ gebracht. Im Feld „Bars Represent“ kann wieder eine Entscheidung darüber getroffen werden, ob die Balken die absolute Häufigkeit (N of cases) oder die relative Häufigkeit (% of cases) repräsentieren sollen.

Achtung: Bei der Interpretation des Clustered Bar Chart ist darauf zu achten, dass die bedingte Verteilung der Variablen im Feld „Category Axis:“, gegeben die Ausprägungen der Variablen im Feld „Define Clusters by:“ dargestellt wird.

- Beispiel 3.1:

Bei der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS<sup>9</sup>) wurde u.a. folgende Frage gestellt: Wie beurteilen Sie ganz allgemein die heutige wirtschaftliche Lage in Deutschland? Diese Response-Variable weist die möglichen Ausprägungen 1 - sehr gut, 2 - gut, 3 - teils, teils, 4 - schlecht und 5 - sehr schlecht auf. Darauf hinaus wurde aus dem ALLBUS die Gruppierungsvariable Erhebungsgebiet mit 1 - alte Bundesländer und 2 - neue Bundesländer entnommen. Die Beobachtungsdaten dieser Variablen sind in der Datei percept\_91\_96.sav unter

<sup>9</sup>Der ALLBUS wird vom ZUMA (Zentrum für Umfragen, Methoden und Analysen e.V., Mannheim) und vom Zentralarchiv für empirische Sozialforschung (Köln) in Zusammenarbeit mit den Mitgliedern des ALLBUS-Ausschusses realisiert. Die vorgenannten Institutionen tragen keine Verantwortung für die Verwendung der Daten in diesem Skript. Alle inhaltlichen Ausführungen zum ALLBUS beziehen sich auf: „Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, ALLBUS 1980-94“, Codebuch, ZA-Nr. 1795, Zentralarchiv für empirische Sozialforschung an der Universität Köln, Zentrum für Umfragen, Methoden und Analysen Mannheim.

### 3. Prüfung der Verteilungform von Variablen

anderem für das Jahr 1991 enthalten.

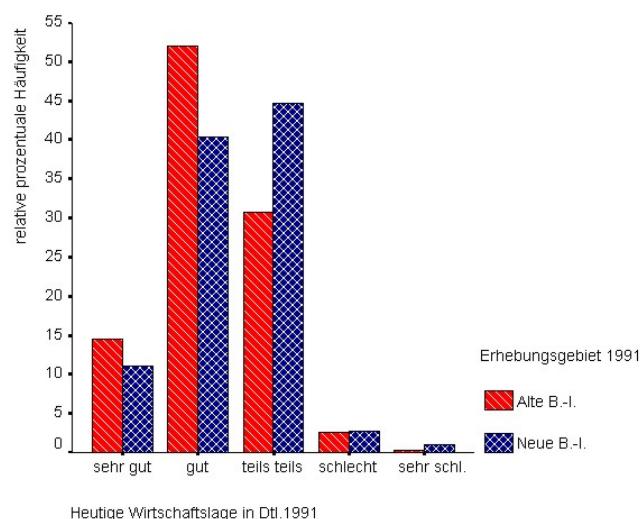
Es soll die Häufigkeitsverteilung der Variablen „Heutige Wirtschaftslage in Deutschland 1991“ ( $y_{1\_91}$ ) separiert nach den beiden Ausprägungen der Variablen Erhebungsgebiet ( $x_{1\_91}$ ) mittels eines gruppierten Balkendiagramms dargestellt werden. Es ergeben sich die bedingten Häufigkeitsverteilungen:

- heutige Wirtschaftslage in Deutschland 1991, gegeben die Ausprägung „alte Bundesländer“ der Variablen Erhebungsgebiet  $f(y_{1\_91}|x_{1\_91} = 1)$  und
- heutige Wirtschaftslage in Deutschland 1991, gegeben die Ausprägung „neue Bundesländer“ der Variablen Erhebungsgebiet  $f(y_{1\_91}|x_{1\_91} = 2)$ .

Im Dialogfeld „Define Clustered Bar: Summaries for Groups of Cases“ (siehe Abb. 3.7) wird die Variable  $y_{1\_91}$  in das Feld „Category Axis:“ und die Variable  $x_{1\_91}$  in das Feld „Define Clusters by:“ gebracht. Als „Bars Represent“ wird % of cases gewählt.

In dem gruppierten Balkendiagramm (Abb. 3.8) repräsentieren die einfach schraffierten (roten) Balken die bedingte Häufigkeitsverteilung heutige Wirtschaftslage in Deutschland 1991, gegeben die Ausprägung „alte Bundesländer“ der Variablen Erhebungsgebiet  $f(y_{1\_91}|x_{1\_91} = 1)$ , und die kreuzschraffierten (blauen) Balken die bedingte Verteilung heutige Wirtschaftslage in Deutschland 1991, gegeben die Ausprägung „neue Bundesländer“ der Variablen Erhebungsgebiet  $f(y_{1\_91}|x_{1\_91} = 2)$ .

Abbildung 3.8.: Gruppiertes Balkendiagramm für Beispiel 3.1



## Stem-and-Leaf Plot und Boxplot

Beide Plots eignen sich zur Darstellung der empirischen Häufigkeitsverteilung metrisch skalierter Variablen, der Boxplot jedoch auch für ordinalskalierte Variablen. Der Stem-and-Leaf Plot

wurde bereits im Abschnitt 2.1 und der Boxplot im Abschnitt 2.2 behandelt. Während dort das Augenmerk nur auf potentielle Ausreißer gerichtet war, interessiert nunmehr die gesamte Häufigkeitsverteilung und ihre Form.

Stem-and-Leaf Plot und Boxplot können ebenfalls separiert nach den Ausprägungen einer Gruppierungsvariablen ausgegeben werden. Dazu ist in dem Dialogfeld „Explore“ (siehe Abb. 2.1) die Gruppierungsvariable in das Feld „Factor List:“ zu bringen. Bei der Erstellung der Boxplots über Graphs - Boxplot ist im Dialogfeld „Boxplot“ (siehe Abb. 2.5) die Option „Summaries for groups of cases“ zu wählen und im Dialogfeld „Define Simple Boxplot: Summaries of Groups of Cases“ die Gruppierungsvariable in das Feld „Category Axis:“ zu bringen.

Beim Vergleich der Stem-and-Leaf Plots für die Ausprägungen der Gruppierungsvariablen ist darauf zu achten, dass nicht zwangsläufig Stem width und die Anzahl der Fälle je Blatt (Each Leaf) übereinstimmen müssen.

- Beispiel 3.2:

Betrachtet man den Stem-and-Leaf Plot der Variablen monatliches persönliches Nettoeinkommen des Beispiels 2.2 aus Abschnitt 2.1 (SPSS-Output 2.2-1), so ist unter Berücksichtigung der auftretenden extremen hohen Einkommenswerte eine rechtsschiefe (linkssteile) Häufigkeitsverteilung zu diagnostizieren. Zu einer analogen Schlußfolgerung gelangt man, wenn man den Boxplot dieser Variablen (Abb. 2.7) aus dem Beispiel 2.2 des Abschnittes 2.1 heranzieht.

Die Häufigkeitsverteilungen des monatlichen persönlichen Nettoeinkommens, getrennt nach den Ausprägungen des Geschlechts, sind als Stem-and-Leaf Plots im SPSS-Output 3.2-1 und als Boxplots in der Abb. 3.9 enthalten.

### 3. Prüfung der Verteilungform von Variablen

**SPSS-Output 3.2-1:** Stem-and-Leaf Plots des monatlichen persönlichen Nettoeinkommens nach dem Geschlecht

Moantl. Nettoeinkommen in DM, offen Stem-and-Leaf Plot for

SEX= Mann

Frequency	Stem	Leaf
,00	0	.
6,00	0	. 233
7,00	0	. 455
12,00	0	. 666777
8,00	0	. 889
11,00	1	. 00011
13,00	1	. 222233
35,00	1	. 4444445555555555
28,00	1	. 66666677777777
60,00	1	. 888888888888888888888999999999
63,00	2	. 0000000000000000000000000000000011111
40,00	2	. 222222222223333333
27,00	2	. 444555555555
28,00	2	. 66666667777777
20,00	2	. 888888999
25,00	3	. 00000000011
10,00	3	. 2233
15,00	3	. 4455555
7,00	3	. 667
5,00	3	. 88
12,00	4	. 00000&
3,00	4	. 2&
16,00	Extremes	(>=4400)

Stem width: 1000

Each leaf: 2 case(s)

& denotes fractional leaves.

### 3.1. Explorative Datenanalyse

Monatliches Nettoeinkommen in DM, offen Stem-and-Leaf Plot for

SEX= Frau

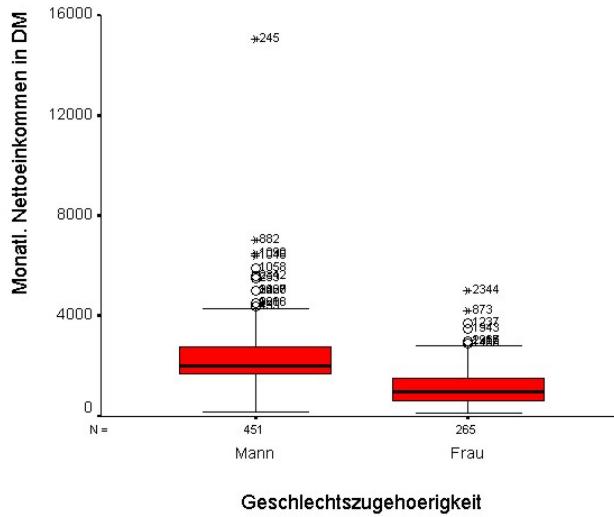
Frequency	Stem	Leaf
2,00	1	. 25
3,00	2	. 049
12,00	3	. 000000225567
15,00	4	. 000000000223689
13,00	5	. 000000455778
27,00	6	. 0000000000000000000023345557
8,00	7	. 00001555
19,00	8	. 0000000000000245555
14,00	9	. 00000000022679
26,00	10	. 0000000000000000000000000000155
8,00	11	. 00000000
19,00	12	. 0000000000000000000000000000
6,00	13	. 000005
10,00	14	. 000055568
18,00	15	. 0000000000000000000000000000
11,00	16	. 0000000055
6,00	17	. 000000
8,00	18	. 00000000
4,00	19	. 0000
11,00	20	. 000000000000
4,00	21	. 0000
4,00	22	. 0000
2,00	23	. 00
1,00	24	. 0
4,00	25	. 0000
,00	26	.
,00	27	.
2,00	28	. 00
8,00	Extremes	(>=2900)

Stem width: 100

Each leaf: 1 case(s)

### 3. Prüfung der Verteilungform von Variablen

Abbildung 3.9.: Boxplot des monatlichen persönlichen Nettoeinkommens nach dem Geschlecht



Die Einkommensverteilung der Männer und der Frauen weisen eine deutliche rechtsschiefe Form auf. Beim Vergleich der beiden Stem-and-Leaf Plots ist die unterschiedliche Stem width und die unterschiedliche Anzahl der Fälle je Blatt (Each leaf) zu beachten.

## Histogramm

Das Histogramm eignet sich zur grafischen Darstellung der empirischen Häufigkeitsverteilung stetiger Variablen, jedoch auch für diskrete Variablen mit sehr vielen Variablenwerten, da solche Variablen vielfach als (quasi-) stetige Variable behandelt werden. Ausgehend von der Stichprobe der Variablen X mit den geordneten Beobachtungswerten  $x_{(1)}, \dots, x_{(n)}$  werden k Klassen mit den Klassengrenzen  $x_0, x_1, \dots, x_k$  und den Klassenbreiten  $b_j = x_j - x_{j-1}$  ( $j = 1, \dots, k$ ) gebildet. Der Beobachtungswert  $x_{(i)}$  ( $i = 1, \dots, n$ ) fällt in die Klasse j, wenn  $x_{j-1} < x_{(i)} \leq x_j$  gilt. Die Klassengrenzen oder Klassenmitten werden auf der Abszissenachse abgetragen. Über den Klassen werden Rechtecke in Höhe der Häufigkeitsdichten eingezeichnet. Die Häufigkeitsdichte ist der Quotient aus relativer Häufigkeit  $f_j$  und Klassenbreite  $b_j$  einer Klasse:

$$\hat{f}_H(x) = \begin{cases} \frac{f_j}{b_j} = \frac{h_j}{nb_j} & \text{für } x_{j-1} < x < x_j \\ 0 & \text{sonst ,} \end{cases} \quad (3.1)$$

worin  $h_j$  die absolute Häufigkeit der Klasse j bezeichnet.  $\hat{f}_H(x)$  ist nicht negativ und wird als Histogramm-Dichte-Schätzer bezeichnet. Unter Verwendung der empirischen Verteilungsfunk-

tion  $F_n(x)$  kann er auch wie folgt geschrieben werden:

$$\hat{f}_H(x) = \frac{F_n(x_j) - F_n(x_{j-1})}{b_j} \quad \text{für } x_{j-1} < x \leq x_j \quad (3.2)$$

Damit entspricht die Fläche der Rechtecke den relativen (Klassen-) Häufigkeiten (flächenproportionale Darstellung). Die Flächen der Rechtecke über den Klassen summieren sich zu Eins:

$$\sum_{j=1}^k b_j \hat{f}_H(x) = \sum_{j=1}^k b_j \frac{h_j}{nb_j} = \frac{1}{n} \sum_{j=1}^k h_j = \frac{n}{n} = 1. \quad (3.3)$$

Speziell für  $b_j = b$  für alle  $j = 1, \dots, k$  (gleiche Klassenbreite für alle  $k$  Klassen) folgt:

$$\hat{f}_H(x) = \begin{cases} \frac{f_j}{b} = \frac{h_j}{nb} & \text{für } x_{j-1} < x < x_j \\ 0 & \text{sonst .} \end{cases} \quad (3.4)$$

In den Statistik-Softwarepaketen wird im allgemeinen eine gleiche Klassenbreite angenommen.  $1/(nb)$  ist ein konstanter Skalierungsfaktor, so dass in diesem Fall für die grafische Darstellung der Nenner ignoriert und die Höhe der Rechtecke mit den absoluten Klassenhäufigkeiten  $h_j$  dargestellt werden kann. Die Klassengrenzen ergeben sich in diesem Fall als  $x_0, x_1 = x_0 + b, x_2 = x_0 + 2b, \dots, x_k = x_0 + kb$ .

Bei der Erstellung eines Histogramms ergeben sich zwei Probleme: die Wahl des Startpunktes  $x_0$  und die Wahl der Klassenbreiten. Beide Faktoren beeinflussen das Bild des Histogramms und damit auch die Auswertung, wodurch ein subjektiver Einfluß nicht zu vermeiden ist.

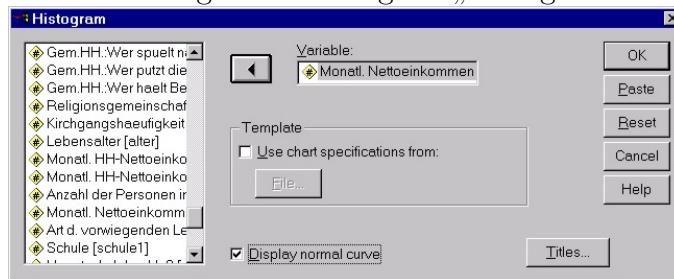
Angefordert werden kann ein Histogramm in SPSS über drei verschiedene Möglichkeiten.

a) ■ Graphs

■ Histogramm...

In dem sich öffnenden Dialogfeld „Histogram“ ist die gewünschte Variable aus der linken Variablenliste in das Feld „Variable:“ zu bringen.

Abbildung 3.10.: Dialogfeld „Histogram“



Bei der Entscheidung für „Display normal curve“ wird in das Histogramm eine Normalverteilungskurve eingezeichnet.

### 3. Prüfung der Verteilungform von Variablen

#### b) ■ Analyze

##### ■ Descriptive Statistics

##### ■ Frequencies...

In dem Dialogfeld „Frequencies“ (siehe Abb. 3.5) wird die darzustellende Variable aus der linken Variablenliste in das Feld „Variable(s)“ gebracht und anschließend die Schaltfläche „Charts...“ betätigt. In dem dann erscheinenden Dialogfeld „Frequencies: Charts“ (siehe Abb. 3.6) wird auf Histograms entschieden.

Auch bei dieser Variante kann die Normalverteilung in das Histogramm eingefügt werden.

#### c) ■ Analyze

##### ■ Descriptive Statistics

##### ■ Explore...

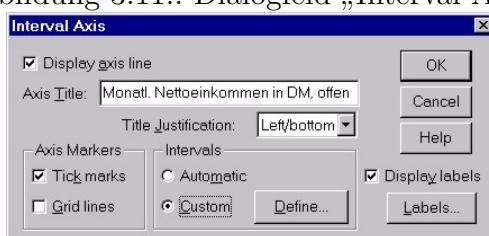
In dem Dialogfeld „Explore“ (siehe Abb. 2.1) wird die zu untersuchende Variable in das Feld „Dependent List:“ gebracht, im Feld „Display“ Plots angeklickt und anschließend die Schaltfläche „Plots...“ betätigt. Im Dialogfeld „Explore: Plots“ (siehe Abb. 2.2) wird nur auf Histogram entschieden.

Bei dieser Variante kann die Normalverteilung in das Histogramm nicht eingefügt werden.

Unter SPSS werden für die darzustellende Variable Klassen mit gleich großer Klassenbreite berechnet. Wegen der gleich großen Klassenbreiten können im Histogramm die Rechtecke über den Klassen in Höhe der Häufigkeiten eingezeichnet werden. Als Häufigkeiten werden stets die absoluten Häufigkeiten verwendet.

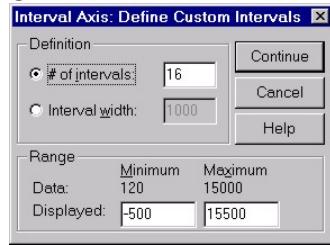
Wird eine andere Klassenbreite oder eine andere Anzahl von Klassen gewünscht, so ist die Grafik (durch Doppelklick auf diese Grafik) in den SPSS Chart Editor zu bringen. Über „Chart“ in der Menüleiste des Chart Editors kann das Histogramm in vielfältiger Weise bearbeitet werden. Durch die Wahl von „Axis...“ in dem Pull-Down-Menü öffnet sich das Dialogfeld „Axis Selection“, in dem Interval zu wählen ist. In dem sich öffnenden Dialogfeld „Interval Axis“ (siehe Abb. 3.11), die identisch mit der Abszissenachse ist, wird in dem Feld „Intervals“ die automatische Intervallbestimmung auf „Custom“ gesetzt und anschließend die Schaltfläche „Define“ betätigt.

Abbildung 3.11.: Dialogfeld „Interval Axis“



Es erscheint das Dialogfeld „Interval Axis: Define Custom Intervals“ (siehe Abb. 3.12).

Abbildung 3.12.: Dialogfeld „Interval Axis: Define Custom Intervals“



Hier kann nun entweder die Anzahl der Intervalle (# of intervals) oder die Klassenbreite (interval width) verändert werden. Dabei bleibt aber die Klassenbreite stets für alle Klassen gleich. Außerdem kann auch der Bereich der Beobachtungswerte (Range), für den das Histogramm gezeichnet werden soll, über Minimum und/oder Maximum variiert werden, wodurch

- Extremwerte ausgeschaltet werden können,
- der Startwert  $x_0$  verändert werden kann,
- interessierende Ausschnitte des Histogramms selektiert werden können.

Bei der Darstellung des Histogramms werden voreingestellt auf der Abszissenachse die Klassenmitten angegeben. Das kann jedoch verändert werden, indem im Dialogfeld „Interval Axis“ (siehe Abb. 3.11) die Schaltfläche „Labels...“ betätigt wird und im nächsten Dialogfeld „Interval Axis: Labels“ im Feld „Type“ Range (Klassengrenzen) ausgewählt wird.

Ein großer Vorteil der Histogramm-Darstellung ist, dass über die Möglichkeiten a) und b) auch die Dichtefunktion der Normalverteilung in die Grafik eingezeichnet werden kann, was den visuellen Vergleich sehr erleichtert. Diese Normalverteilung wird aufgrund der aus den Daten geschätzten Werte von Mittelwert und Varianz berechnet.

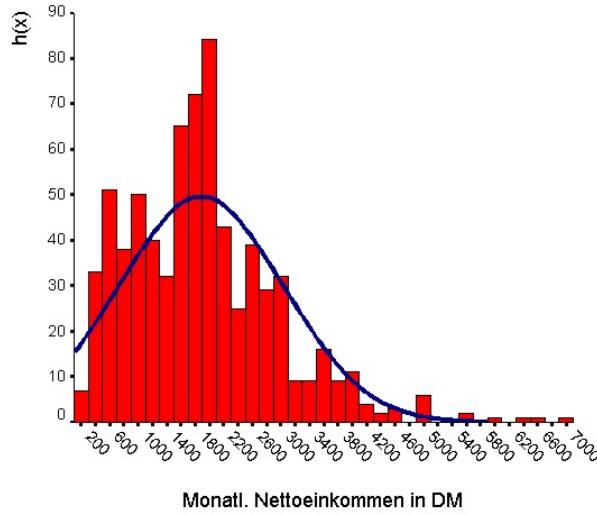
- Beispiel 3.2 (Fortsetzung):

Für die Variable monatliches persönliches Nettoeinkommen der Datei allbus.sav wird ein Histogramm mit Normalverteilung über die Version a) erzeugt. Das in der folgenden Abbildung enthaltene Histogramm wurde in der Weise derart verändert, dass die Intervallbreite auf 200 gesetzt und der darzustellende Wertebereich von 100 bis 7100 begrenzt wurde, um den Ausreißer nicht einzubeziehen.

Das explorative Experimentieren mit veränderter Klassenbreite oder Klassenzahl erweist sich als außerordentlich nützlich, da sich das Bild des Histogramms und somit die daraus folgende Interpretation und Hypothesenformulierung durchaus verändern können. Mit dem Variieren

### 3. Prüfung der Verteilungform von Variablen

Abbildung 3.13.: Histogramm mit Normalverteilung für Beispiel 3.2, Klassenbreite 200



der Klassenbreite sollte erreicht werden, dass das Histogramm möglichst „glatt“ (glatt nicht im Sinne von eben, sondern im Sinne eines gedachten Kurvenverlaufes durch die Höhe der Rechtecke über den Klassenmitten) wird, ohne jedoch die Besonderheiten der Daten aus den Augen zu verlieren. Eine zu kleine Klassenbreite hat zur Folge, dass noch zu viel von der Stichprobenvariabilität (auch der unwesentlichen) im Histogramm sichtbar bleibt, es weist sehr viele Modalwerte auf (siehe Abb. 3.13). Eine zu große Klassenbreite überglättet dagegen das Histogramm, so dass möglicherweise wichtige Eigenschaften der unterliegenden Verteilung verwischt werden. Letztendlich bleibt bei dieser Vorgehensweise die Entscheidung über die „passende“ Klassenbreite immer subjektiv und birgt auch eine gewisse Gefahr in sich. In der Literatur<sup>10</sup> werden verschiedene Regeln zur rechnerischen Ermittlung der Klassenanzahl bzw. -breite vorgeschlagen, die auf unterschiedlichen Konzepten und bestimmten Verteilungsannahmen basieren. Einige dieser Regeln sollen hier genannt werden, ohne auf ihre Herleitung einzugehen:

$$\begin{aligned} k_1 &= \sqrt{n} & k_2 &= 2\sqrt{n} & k_3 &= 10 \cdot \log_{10} n \\ b_1 &= \frac{3,49 \cdot s}{\sqrt[3]{n}} & b_2 &= \frac{2 \cdot IQR}{\sqrt[3]{n}}, \end{aligned} \tag{3.5}$$

worin  $n$  die Anzahl der gültigen Fälle (Stichprobenumfang),  $s$  die aus den Daten geschätzte Standardabweichung und  $IQR$  der aus den Daten nach (2.50) berechnete Interquartilsabstand sind.

- Beispiel 3.2 (Fortsetzung):

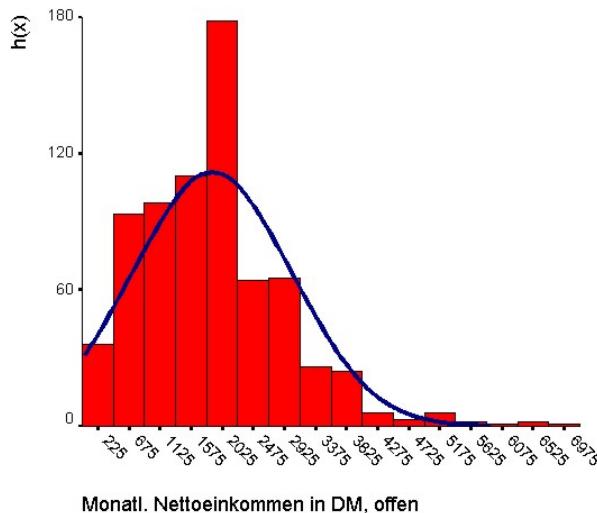
Für die Variable monatliches persönliches Nettoeinkommen sind  $n = 716$ ,  $s = 1150,01$  und

<sup>10</sup>vgl. u.a. Heiler, S., Michels, P. (1994), S. 73 ff.

$IQR = 1307,5$ . Damit ergeben sich (gerundete Werte)  $k_1 \approx 27$ ;  $k_2 \approx 54$ ;  $k_3 \approx 29$ ;  $b_1 \approx 449$  und  $b_2 \approx 292$ . Die Regeln  $k_1$  bis  $k_3$  unterschätzen damit deutlich die Klassenbreite ( $k_1$  unterstellt eine Gleichverteilung,  $k_2$  eine Dreiecksverteilung und  $k_3$  stellt einen Kompromiß zwischen verschiedenen Verteilungen dar), während  $b_1$  (unterstellt Normalverteilung) und auch  $b_2$  hier angemessenere Schätzungen liefern.

Die Abb. 3.14 zeigt das Histogramm mit einer Klassenbreite von 450 und einem Wertebereich von 0 bis 7200.

Abbildung 3.14.: Histogramm mit Normalverteilung für Beispiel 3.2, Klassenbreite 450



Mit der Klassenbreite von 450 werden die Anforderungen an ein Histogramm im wesentlichen erfüllt. Im Vergleich zum Histogramm mit einer Klassenbreite von 200 (Abb. 3.13) ist eine deutliche Glättung des Histogramms erreicht worden.

Das Histogramm ist eine recht grobe Schätzung der „wahren“ unterliegenden Dichtefunktion und weist einige Nachteile auf, z.B. der willkürlich gewählte Startpunkt  $x_0$ , die willkürlich gewählte (feste) Klassenbreite und damit die mehr oder weniger willkürlich gewählten Klassengrenzen, die Darstellung der unterliegenden stetigen Häufigkeitsverteilung durch eine Treppenfunktion (Sprünge bei den Klassengrenzen).

## Kerndichteschätzung

Eine Möglichkeit, die einige der Nachteile des Histogramms vermeiden, die den Beobachtungsdaten unterliegende Gesamtstruktur grafisch verdeutlichen und die „wahre“ Dichtefunktion approximieren soll, ist die sogenannte Kerndichteschätzung. Die Grundidee ist in einfacherster Weise, dass zentrale Intervalle um  $x$ ,  $(x - w; x + w)$ , über die X-Achse gleiten und die Wahrscheinlichkeit, dass Beobachtungswerte  $x_{(i)}$  in ein solches Intervall fallen,  $P(x-w < X < x+w)$ ,

### 3. Prüfung der Verteilungform von Variablen

durch die relative Häufigkeit approximiert wird:

$$\hat{f}_S(x) = \frac{F_n(x+w) - F_n(x-w)}{2w}, \quad (3.6)$$

wobei eine Maßstabsänderung erfolgte, so dass die Gesamtfläche unter  $\hat{f}_S(x)$  zu eins integriert. Der Parameter  $w(w > 0)$  wird als Bandbreite (band width, window width) oder Glättungsparameter bezeichnet und bestimmt den Streubereich um  $x$ .<sup>11</sup> Dieser Schätzer ist ähnlich einem Histogramm mit Rechtecken der konstanten Breite  $2w$ , aber der Anfangspunkt  $x_0$  ist nicht fest. Die Wahrscheinlichkeitsmasse (das Gewicht) der Beobachtungswerte wird dabei auf die Umgebung der Breite  $2w$  verteilt. Eine äquivalente Darstellung von (3.6) ist

$$\hat{f}_S(x) = \frac{1}{nw} \sum_{i=1}^n S_w\left(\frac{x-x_i}{w}\right) \quad (3.7)$$

mit

$$S_w(u) = \begin{cases} \frac{1}{2} & \text{wenn } |u| < 1, \\ 0 & \text{sonst.,} \end{cases} \quad u = \frac{x-x_i}{w} \quad (3.8)$$

Die geschätzte Dichtefunktion wird also so konstruiert, dass für jedes gegebene  $x$  die Höhen der um jeden Beobachtungspunkt  $x_{(i)}$  zentrierten Rechtecke summiert werden und der Durchschnitt ermittelt wird.

Dieser einfache Schätzer beseitigt noch nicht die Diskontinuität, die auch für das Histogramm typisch ist. Gewöhnlich wird deshalb eine allgemeinere, symmetrische Dichtefunktion  $K_w$  verwendet, die eine Glättung bewirkt. Die Funktion  $K_w$  wird als Kernfunktion (kernel function) bezeichnet und bestimmt die Gestalt der Verteilung der Wahrscheinlichkeitsmasse über den Intervallen. Ein Kerndichteschätzer ist damit allgemein definiert als

$$\hat{f}_K(x) = \frac{1}{nw} \sum_{i=1}^n K_w\left(\frac{x-x_i}{w}\right). \quad (3.9)$$

- Beispiel 3.3:

Ein Beispiel<sup>12</sup> soll die Wirkungsweise der Kernschätzung verdeutlichen. Es liege eine (künstliche) Stichprobe vom Umfang  $n = 10$  mit den geordneten Werten 6,9; 7,2; 7,3; 8,0; 9,0; 11,2; 11,4; 11,9; 12,0; 12,2 vor. Für eine Stichprobe solchen Umfangs wird man in der Regel keine Kerndichteschätzung anwenden; es soll auch nur zur Demonstration dienen. Die Abb. 3.15 zeigt in Form eines Stabdiagramms die empirische Wahrscheinlichkeitsfunktion auf der Basis der relativen Häufigkeiten.

---

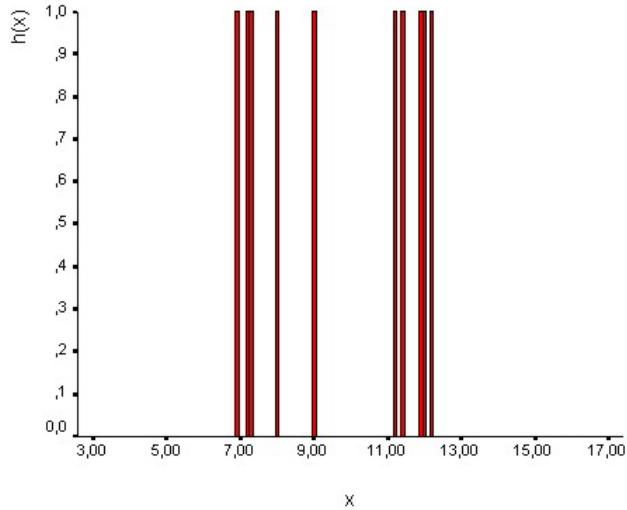
<sup>11</sup>Die Bandbreite wird in der Literatur im allgemeinen mit  $h$  symbolisiert. Da  $h$  in diesen Ausführungen jedoch schon für die absolute Häufigkeit vergeben ist, wurde hier das Symbol  $w$  (von width) gewählt.

<sup>12</sup>Dieses Beispiel wurde aus Fox, J., Long, J.S. (1990) entnommen.

Als Kerndichtefunktion  $K_w(u)$  wird die Dichtefunktion der Normalverteilung, der sogenannte Normalkern, mit einer Bandbreite von  $w = 0,8$  gewählt:

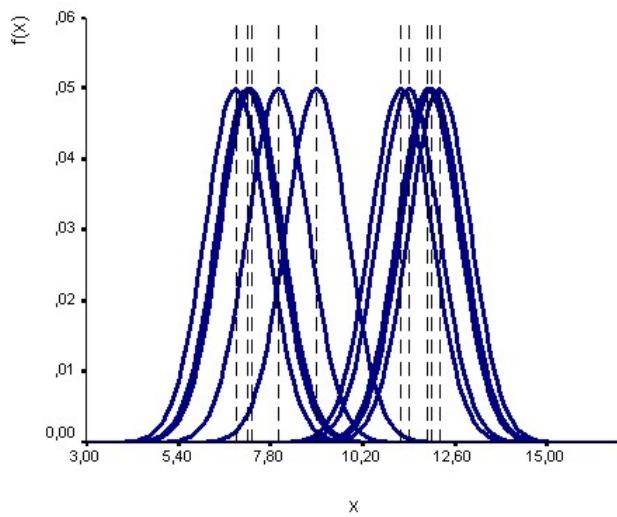
$$K_w(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right). \quad (3.10)$$

Abbildung 3.15.: Stabdiagramm der relativen Häufigkeiten



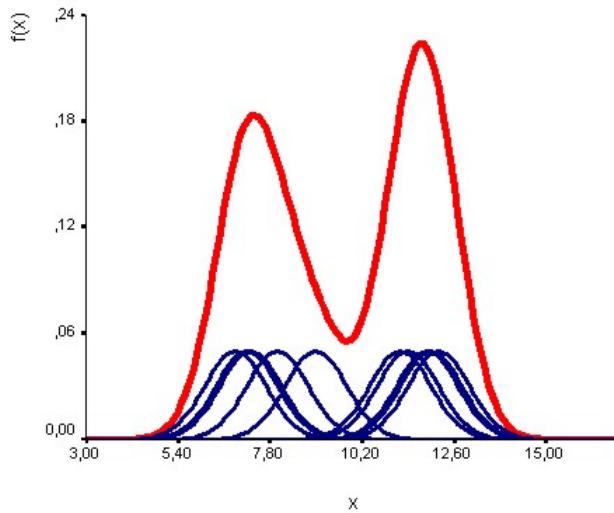
Die Normalkerne über jeder Beobachtung  $x_{(i)}$  zeigt Abb. 3.16 und den Kerndichteschätzer die Abb. 3.17, wobei zu beachten ist, dass der Maßstab der Ordinatenachse in beiden Abbildungen unterschiedlich ist.

Abbildung 3.16.: Normalkerne über den Beobachtungen



### 3. Prüfung der Verteilungform von Variablen

Abbildung 3.17.: Kerndichteschätzer mit Normalkern und Bandbreite  $w = 0,8$



Die mit der Kerndichteschätzung verbundenen Probleme sind vor allem die Wahl der Kernfunktion und der Bandbreite  $w$ , so dass eine optimale Approximation an die den Daten unterliegende „wahre“ Dichtefunktion erreicht wird, wobei verschiedene Kriterien zugrundegelegt werden können. Diese Probleme sollen hier nicht weiter vertieft werden, da Kerndichteschätzungen Gegenstand der Lehrveranstaltung „Semiparametrische Modelle“ sind.<sup>13</sup>

Kerndichteschätzer sind nicht in SPSS implementiert. Eine Statistik-Software, mit der die Kerndichteschätzungen realisiert werden können, ist z.B. XploRe.

## Wahrscheinlichkeitsplots

Wahrscheinlichkeitsplots werden im allgemeinen verwendet, um grafisch zu prüfen, ob die empirische Verteilung einer stetigen Variablen einer angenommenen Testverteilung entspricht. SPSS bietet folgende Testverteilung an: beta, chi-square, exponential, gamma, half-normal, Laplace, Logistic, Lognormal, normal, pareto, Student's t, Weibull und uniform. Falls notwendig können zu der Testverteilung die Anzahl der Freiheitsgrade und bestimmte Parameter spezifiziert werden. Am häufigsten wird jedoch der Vergleich zur Normalverteilung geführt, der auch hier demonstriert werden soll.

Unter SPSS sind vier Wahrscheinlichkeitsplots verfügbar:

- der Q-Q Plot,
- der trendbereinigte (detrended) Q-Q Plot,

---

<sup>13</sup>Siehe auch Fox, J., Long, J.S. (1990), S. 88 ff.; Härdle, W. (1991); Härdle, W., Klinke, S., Turlach, B.A. (1995); Heiler, S., Michels, P. (1994), S. 54 ff.

- der P-P Plot,
- der trendbereinigte (detrended) P-P Plot,

wobei Q-Q für Quantil-Quantil und P-P für Probability-Probability steht.

#### Normal Q-Q Plot:

Jedem (geordneten) Beobachtungswert  $x_{(i)}$  der Variablen X werden Rangzahlen  $R(x_{(i)})$  zugeordnet, auf deren Basis die empirischen Anteilswerte  $f(x_{(i)})$  (proportion estimation) geschätzt werden. Unter SPSS stehen vier verschiedene Möglichkeiten zur Bestimmung der  $f(x_{(i)})$  zur Verfügung:

$$f(x_{(i)}) = \begin{cases} \frac{R(x_{(i)}) - \frac{3}{8}}{\frac{1}{n + \frac{1}{4}}} & \text{Blom} \\ \frac{R(x_{(i)}) - \frac{1}{2}}{\frac{1}{n}} & \text{Rankit} \\ \frac{R(x_{(i)}) - \frac{1}{3}}{\frac{1}{n + \frac{1}{3}}} & \text{Tukey} \\ \frac{R(x_{(i)})}{\frac{n+1}{n+1}} & \text{Van der Waerden} \end{cases} \quad (3.11)$$

Dabei bietet SPSS verschiedene Versionen für die Behandlung von Bindungen (Ties) an (Rank Assigned to Ties). Mit  $f(x_{(i)})$  können für jedes  $x_{(i)}$  die Werte  $F(x_{(i)})$  der empirischen Verteilungsfunktion bestimmt werden.

Für die Werte  $F(x_{(i)})$  werden die zugehörigen  $z_{F(i)}$ -Werte aus der Verteilungsfunktion der Standardnormalverteilung ermittelt.

Der Normal Q-Q Plot ergibt sich, indem die empirischen Quantile  $x_{F(i)}$  gegen die entsprechenden (unter der Normalverteilung) erwarteten Quantile  $z_{F(i)}$  in ein Koordinatensystem eingetragen, wobei die Abszissenachse die empirischen Quantile und die Ordinatenachse die erwarteten Quantile aufnimmt. Wenn die Variable aus einer normalverteilten Grundgesamtheit stammt, liegen die Punkte  $(x_{F(i)}, z_{F(i)})$  mehr oder weniger auf einer Winkelhalbierenden. Bei systematischen Abweichungen von der Geraden liegt eine andere Verteilung zugrunde. Starke Abweichungen von der Geraden nur an den Enden signalisieren Ausreißer in den Beobachtungswerten.

#### Trendbereinigter Normal Q-Q Plot:

Bei diesem Wahrscheinlichkeitsplot werden die geordneten Beobachtungswerte  $x_{(i)}$  gegen die

### 3. Prüfung der Verteilungform von Variablen

Abweichungen  $d_{(i)}$  abgetragen. Die Abweichungen  $d_{(i)}$  ergeben sich als Differenz der beobachteten standardisierten  $x_{(i)}$ -Werte von den (unter der Normalverteilung) erwarteten Werten. Diese Abweichungen werden gemäß

$$d_{(i)} = \frac{x_{(i)} - \bar{x}}{s} - z_{(i)}, \quad (3.12)$$

ermittelt, worin  $\bar{x}$  und  $s$  die aus den Daten geschätzten Werte von Mittelwert und Standardabweichung sind. Wenn die Variable aus einer normalverteilten Grundgesamtheit stammt, liegen die Punkte  $(x_{(i)}, d_{(i)})$  mehr oder weniger auf einer horizontalen, durch den Nullpunkt verlaufenden Geraden.

#### Normal P-P Plot:

Ausgegangen wird wiederum von den geordneten Beobachtungswerten  $x_{(i)}$ , die einer Standardisierung unterzogen werden.  $F(x_{(i)})$  bezeichne die empirische Verteilungsfunktion, d.h. die kumulierten Häufigkeiten für jeden Wert  $x_{(i)}$ , und  $\Phi(z_{(i)})$  den Wert der Verteilungsfunktion der Standardnormalverteilung der zugehörigen  $z_{(i)}$ -Werte. Im Normal P-P Plot wird  $\Phi(z_{(i)})$  in Abhängigkeit von  $F(x_{(i)})$  dargestellt. Wurde die Variable aus einer normalverteilten Grundgesamtheit entnommen, so sollten die Punkte  $(F(x_{(i)}), \Phi(z_{(i)}))$  wiederum annähernd auf einer Winkelhalbierenden liegen.

#### Trendbereinigter Normal P-P Plot:

Bei diesem Plot werden die Abweichungen  $D_{(i)} = \Phi(z_{(i)}) - F(x_{(i)})$  in Abhängigkeit von  $F(x_{(i)})$  im Koordinatensystem abgetragen.

#### Unter SPSS können beide Normal Q-Q Plots über

##### ■ Analyze

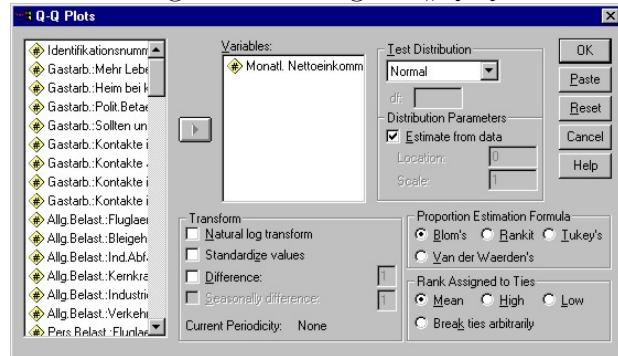
##### ■ Descriptive Statistics

##### ■ Explore

angefordert werden. Im Dialogfeld „Explore“ (siehe Abb. 2.1) wird die zu untersuchende Variable in das Feld „Dependent List:“ gebracht, im Feld „Display“ nur Plots aktiviert und anschließend die Schaltfläche „Plots...“ betätigt. Im Dialogfeld „Explore: Plots“ (siehe Abb. 2.2) wird nur auf „Normality plots with tests“ entschieden. Eine Auswahl der Berechnung von  $f(x_{(i)})$  entsprechend (3.11) besteht nicht; es wird die Berechnung nach van der Waerden verwendet.

Alle vier Wahrscheinlichkeitsplots lassen sich unter SPSS über das Menü „Graphs“ erstellen, und zwar die Q-Q Plots durch die Wahl von „Q-Q...“ und die P-P Plots durch die Wahl von „P-P...“. In beiden Fällen öffnet sich ein identisches Dialogfeld, das als Beispiel für die Q-Q Plots in der folgenden Abbildung enthalten ist.

Abbildung 3.18.: Dialogfeld „Q-Q Plots“

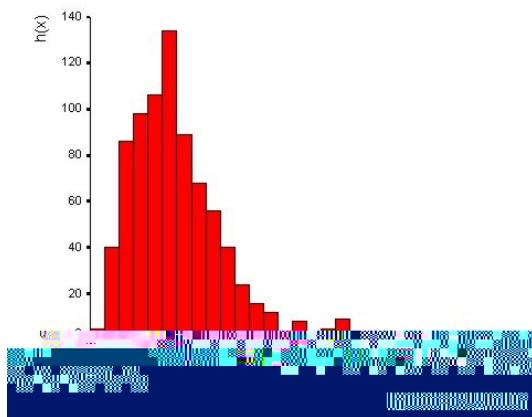


In diesem Dialogfeld wird als erstes die zu analysierende Variable aus der linken Variablenliste in das Feld „Variables:“ gebracht. Unter „Test Distribution“, „Distribution Parameters“, „Proportion Estimation Formula“, „Rank Assigned to Ties“ und „Transform“ können die entsprechenden Entscheidungen getroffen werden.

- Beispiel 3.4:

Als Datenbasis für dieses Beispiel dient die Datei mieten.sav<sup>14</sup>, in der u.a. die Variable Höhe der Monatsmiete (in DM) für 815 Berliner Mietwohnungen enthalten ist. Für diese Variable werden das Histogramm und die Wahrscheinlichkeitsplots über das Menü „Graphs“ erzeugt, wobei ein Vergleich mit der Normalverteilung erfolgen soll und alle Voreinstellungen belassen werden.

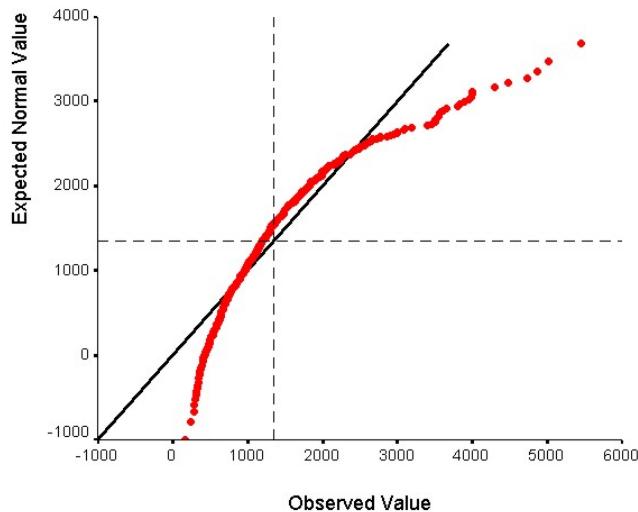
Abbildung 3.19.: Histogramm für Beispiel 3.4



<sup>14</sup>Die Datei mieten.sav wurde von Herrn Prof. Dr. P. P. Eckstein, Fachhochschule für Technik und Wirtschaft Berlin, im Internet zur Verfügung gestellt und ist beschrieben in: Eckstein, P. (1997), S. 43 f. (<http://www.f3.fhtw-berlin.de/Professoren/Eckstein/download.html>).

### 3. Prüfung der Verteilungform von Variablen

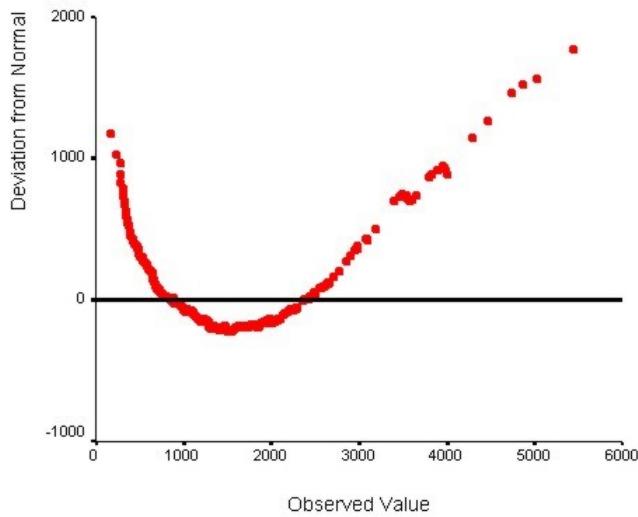
Abbildung 3.20.: Normal Q-Q Plot für Beispiel 3.4



Die gestrichelten Referenzlinien bezeichnen den Mittelwert  $\bar{x} = 1343$ . Die konkave Gestalt des Punkteverlaufs im Normal Q-Q Plot lässt auf eine rechtsschiefe Verteilung der Monatsmiete schließen, wie auch das Histogramm verdeutlicht.

Der zugehörige trendbereinigter Normal Q-Q Plot ist als zweite Grafik im SPSS Viewer enthalten.

Abbildung 3.21.: Trendbereinigter Normal Q-Q Plot für Beispiel 3.4



Die beiden Normal P-P Plots für die Variable Monatsmiete geben die zwei folgenden Abbildungen wieder.

Abbildung 3.22.: Normal P-P Plot für Beispiel 3.4

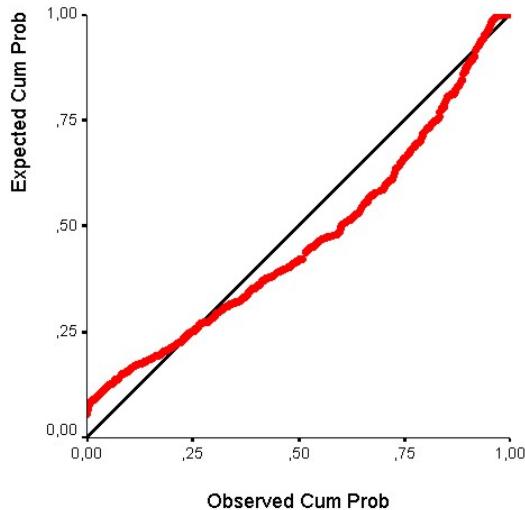
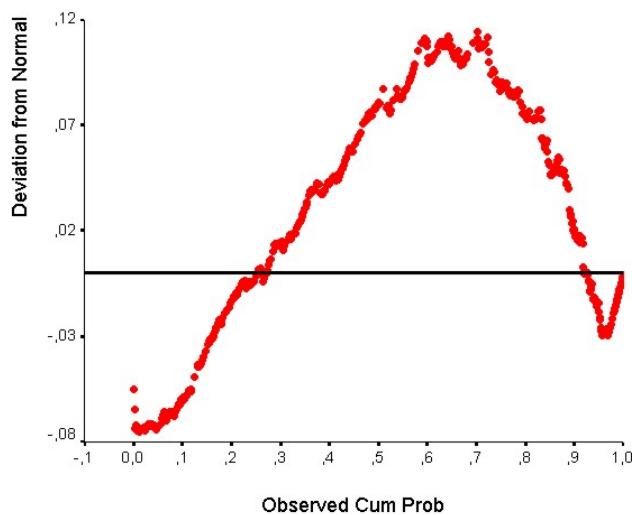


Abbildung 3.23.: Trendbereinigter Normal P-P Plot für Beispiel 3.4



Eine weitere Option, die die Darstellung der Wahrscheinlichkeitsplots über das Menü „Graphs“ bietet, ist die Transformation der Beobachtungswerte  $x_i$ . Sie bietet u.a. die Möglichkeit, die standardisierten Werte  $z_{(i)}$  oder den natürlichen Logarithmus der  $x_{(i)}$ -Werte den Wahrscheinlichkeitsplots zugrunde zu legen. Geeignete transformierte Beobachtungswerte folgen oftmals eher einer Normalverteilung als die Ausgangswerte (zur Transformation der Daten siehe Kapitel 1).

Für die Variable Monatsmiete wird eine ln-Transformation durchgeführt und wiederum das Histogramm sowie die beiden Q-Q Plots erstellt.

### 3. Prüfung der Verteilungform von Variablen

Abbildung 3.24.: Histogramm für Beispiel 3.4 mit den ln-transformierten Werten

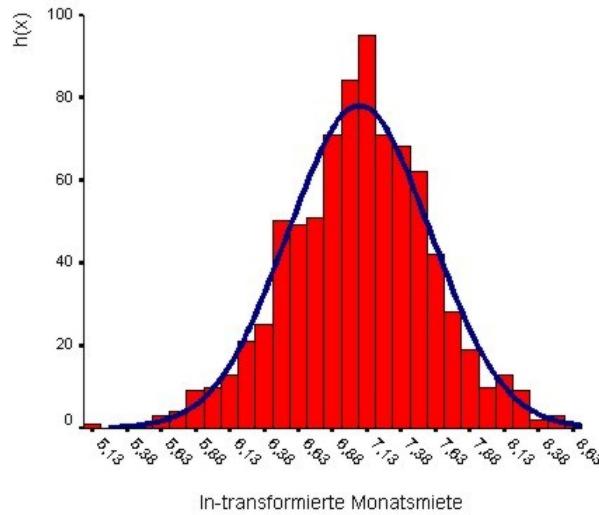


Abbildung 3.25.: Normal Q-Q Plot für Beispiel 3.4 mit den ln-transformierten Werten

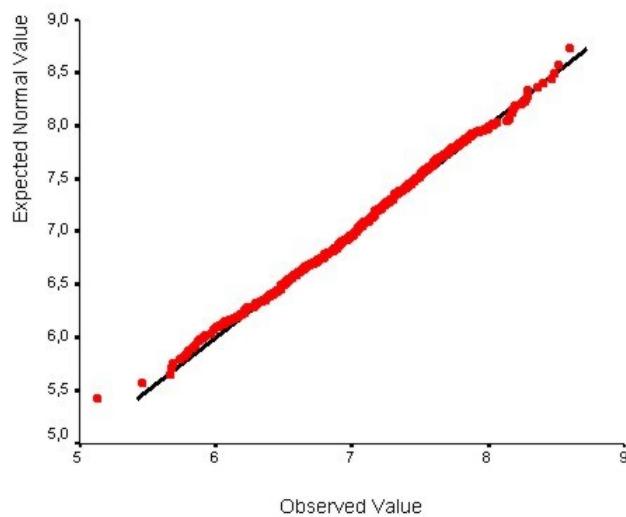
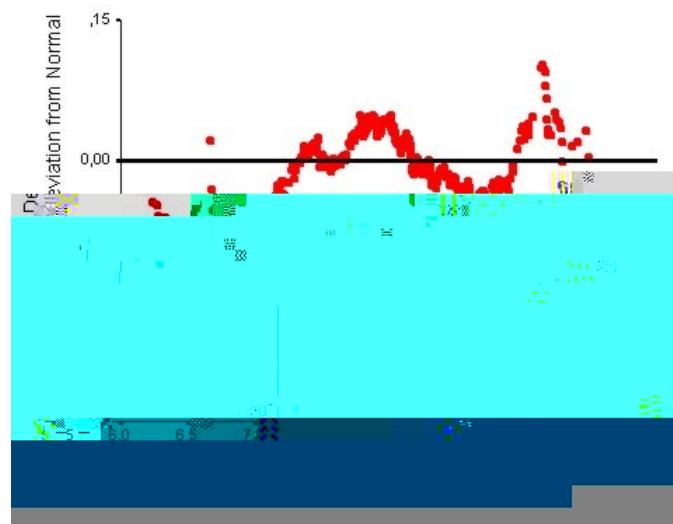


Abbildung 3.26.: Trendbereinigter Normal Q-Q Plot für Beispiel 3.4 mit den ln-transformierten Werten



Während die Monatsmiete in ihren Ausgangswerten eine rechtsschiefe Verteilung aufweist, konnte durch die ln-Transformation eine sehr gute Anpassung an die Normalverteilung erreicht werden.

## 3.2. Beschreibende Datenanalyse

Mittels beschreibender Maßzahlen können Informationen über das Zentrum und die Ausdehnung von ordinal- und metrisch skalierten Variablen sowie über die Symmetrie bzw. Schiefe einer empirischen Verteilung von metrisch skalierten Variablen gewonnen werden.

### Vergleich von Mittelwert, Median und Modus:

Während das arithmetische Mittel metrisches Skalenniveau erfordert, kann der Median für wenigstens ordinalskalierte Variablen und der Modus für Variablen jeden Skalenniveaus ermittelt werden.

Ist eine metrisch skalierte Variable gegeben, so fallen diese drei Mittelwerte zusammen, wenn eine symmetrische Häufigkeitsverteilung vorliegt. Für eine rechtsschiefe Verteilung gilt in der Regel die Größenbeziehung  $\text{Modus} < \text{Median} < \text{Mittelwert}$  und für eine linksschiefe Verteilung  $\text{Mittelwert} < \text{Median} < \text{Modus}$ .

### Verwendung der Quartile und des Interquartilsabstandes:

Diese Maßzahlen erfordern wenigstens ordinale Skalenniveau. Wie bereits im Abschnitt 2.2 bei der Behandlung des Boxplots, in dessen Darstellung die Quartile  $x_{0,25}$ ,  $x_{0,5}$ ,  $x_{0,75}$  und der Interquartilsabstand (IQR) eingehen, gezeigt wurde, können sie verwendet werden, um Aussagen über Symmetrie oder Schiefe einer Verteilung zu treffen. Es gilt für eine

- symmetrische Verteilung:

$$(x_{0,25} + x_{0,75})/2 = x_{0,5}$$

$$x_{0,25} + IQR/2 = x_{0,5} \quad x_{0,75} - IQR/2 = x_{0,5} \quad x_{0,5} - x_{0,25} = x_{0,75} - x_{0,5}$$

- rechtsschiefe Verteilung:

$$(x_{0,25} + x_{0,75})/2 > x_{0,5}$$

$$x_{0,25} + IQR/2 > x_{0,5} \quad x_{0,75} - IQR/2 > x_{0,5} \quad x_{0,5} - x_{0,25} < x_{0,75} - x_{0,5}$$

- linksschiefe Verteilung:

$$(x_{0,25} + x_{0,75})/2 < x_{0,5}$$

$$x_{0,25} + IQR/2 < x_{0,5} \quad x_{0,75} - IQR/2 < x_{0,5} \quad x_{0,5} - x_{0,25} > x_{0,75} - x_{0,5}$$

### 3. Prüfung der Verteilungsform von Variablen

#### Schiefe und Exzeß<sup>15</sup>:

Die Schiefe (Skewness) bezeichnet die Asymmetrie einer unimodalen Häufigkeitsverteilung eines wenigstens ordinalskalierten Merkmals oder einer Wahrscheinlichkeitsfunktion bzw. Dichtefunktion einer Zufallsvariablen. Sind metrisch skalierte Merkmale bzw. Zufallsvariable gegeben, kann die Schiefe mittels Maßzahlen gemessen werden, die die Größenordnung und die Richtung der Abweichung von der Symmetrie angeben. Diese Maßzahlen sind so definiert, dass sie dimensionslos und somit maßstabsunabhängig sind und bei symmetrischen Verteilungen den Wert Null, bei Rechtsschiefe der Verteilung einen positiven Wert und bei Linksschiefe einen negativen Wert annehmen. Für ein metrisch skaliertes Merkmal X mit den Beobachtungswerten  $x_j$  ( $j = 1, \dots, k$ ), den zugehörigen absoluten Häufigkeiten  $h(x_j)$ , dem arithmetischen Mittel  $\bar{x}$ , der Standardabweichung s und dem dritten zentralen Moment  $m_3(\bar{x})$  ist u.a. der Momentkoeffizient der Schiefe (Bowle-Fishersches Schiefemaß) gebräuchlich:

$$g = \frac{m_3(\bar{x})}{s^3} = \frac{\frac{1}{n} \sum_{j=1}^k (x_j - \bar{x})^3 h(x_j)}{s^3}. \quad (3.13)$$

Der Exzeß (Kurtosis, Wölbung) ist eine Maßzahl für eine unimodale Häufigkeitsverteilung eines metrisch skalierten Merkmals oder der Dichtefunktion bzw. Wahrscheinlichkeitsfunktion einer Zufallsvariablen, die die Abweichung der Steilheit der Verteilung von der Steilheit der Dichtefunktion der Normalverteilung angibt (bei gleicher Varianz). Der Exzeß ist definiert unter Verwendung des 4. zentralen Moments  $m_4(\bar{x})$  der Verteilung, normiert auf die vierte Potenz der Standardabweichung. Der Exzeß berechnet sich für eine Häufigkeitsverteilung mit den Beobachtungswerten  $x_j$  ( $j = 1, \dots, k$ ), den zugehörigen absoluten Häufigkeiten  $h(x_j)$ , dem arithmetischen Mittel  $\bar{x}$  und der Standardabweichung s bzw. Varianz  $s^2$  als

$$e = \frac{m_4(\bar{x})}{s^4} - 3 = \frac{\frac{1}{n} \sum_{j=1}^k (x_j - \bar{x})^4 h(x_j)}{\left( \frac{1}{n} \sum_{j=1}^k (x_j - \bar{x})^2 h(x_j) \right)^2} - 3 \quad (3.14)$$

Der Exzeß der Dichtefunktion der Normalverteilung ist gleich Null. Ist eine Verteilung steiler als die Normalverteilung, d.h., ist das absolute Maximum größer als das der Normalverteilung (hochgipflig, leptokurtisch), ist e positiv. Ist e negativ, so ist die Verteilung flacher (platykurtisch).

---

<sup>15</sup>Vgl. u.a. Rönnz, B., Strohe, H.G. (Hrsg.) (1994), S. 323 f., S. 115

Die bisher genannten Maßzahlen können unter SPSS auf verschiedene Weise angefordert werden.

Mittelwert, Median, Modus (Modalwert), Schiefe, Exzeß und die Quartile sind erhältlich über

■ Analyze

■ Descriptive Statistics

■ Frequencies...

Nach der Anwahl der Schaltfläche „Statistics...“ im Dialogfeld „Frequencies“ (siehe Abb. 3.5) können im folgenden Dialogfeld „Frequencies: Statistics“ die gewünschten Maßzahlen ausgewählt werden. Der Modus wird nur über diese Variante ausgegeben. Gibt es mehrere Variablenwerte mit der gleichen maximalen beobachteten Häufigkeit, wird der kleinste Variablenwert ausgegeben.

Mittelwert, Schiefe und Exzeß sind weiterhin über

■ Analyze

■ Descriptive Statistics

■ Descriptives...

erhältlich. Nach Anklicken der Schaltfläche „Options...“ im Dialogfeld „Descriptives“ (siehe Abb. 2.23) können diese drei Maßzahlen im folgenden Dialogfeld „Descriptives: Options“ (siehe Abb. 2.24) angekreuzt werden.

Mittelwert, Median, Schiefe, Exzeß, Quartile und Interquartilsabstand können über

■ Analyze

■ Descriptive Statistics

■ Explore...

ausgegeben werden. Nach Betätigung der Schaltfläche „Statistics...“ im Dialogfeld „Explore“ (Abb. 2.1) erscheint als nächstes das Dialogfeld „Explore: Statistics“ (Abb. 2.3). Dort ist Descriptives bereits voreingestellt, die Mittelwert, Median, Schiefe, Exzeß und Interquartilsabstand liefern. Durch Ankreuzen von Percentile erhält man die Werte für die Perzentile 5%, 10%, 25%, 50%, 75%, 90% und 95%. Die Perzentilwerte werden nach der Methode Heverage bestimmt. Zur Erläuterung dieser Methode wird folgende Symbolik vereinbart:

$p$  - der geforderte Anteil

$x_p$  - Perzentil

$g_i$  - das Gewicht des Wertes  $x_{(i)}$ ,

$w_j$  - die kumulierte Summe der Gewichte  $g_i$  bis einschließlich  $x_{(j)}$ :

$$w_j = \sum_{i=1}^j g_i,$$

### 3. Prüfung der Verteilungform von Variablen

W - die Summe der Gewichte aller Variablenwerte:

$$W = \sum_{i=1}^n g_i.$$

Aus den Bedingungen

$$w_{j1} \leq W \cdot p < w_{j1+1} \text{ und } w_{j2} \leq (W + 1) \cdot p < w_{j2+1}$$

sind j1 und j2 zu bestimmen.

Die Festlegung der Perzentile  $x_p$  erfolgt nach folgenden Regeln:

- (a) Wenn  $(W + 1)p - w_{j2} \geq 1$ , dann ist

$$x_p = x_{j2+1}. \quad (3.15)$$

- (b) Wenn  $(W + 1)p - w_{j2} < 1$  und  $g_{(j2+1)} \geq 1$ , dann ist

$$x_p = (1 - (W + 1)p + w_{j2}) \cdot x_{j2} + [(W + 1)p - w_{j2}] \cdot x_{(j2+1)}. \quad (3.16)$$

- (c) Wenn  $(W + 1)p - w_{j2} < 1$  und  $g_{(j2+1)} < 1$ , dann ist

$$x_p = \left(1 - \frac{(W + 1)p - w_{j2}}{g_{j2+1}}\right) \cdot x_{(j2)} + \frac{(W + 1)p - w_{j2}}{g_{j2+1}}. \quad (3.17)$$

Durch diese Regeln wird berücksichtigt, ob p genau mit einem  $x_{(i)}$  identisch ist oder zwischen zwei  $x_{(i)}$ -Werten fällt, wobei die unterschiedlichen Abstände von  $x_{(p)}$  zu  $x_{(i)}$  bzw. von  $x_{(p)}$  zu  $x_{(i+1)}$  einbezogen werden.

- Beispiel 3.4 (Fortsetzung):

Für die Variable Monatsmiete der Datei mieten.sav werden nach der ersten Variante die notwendigen Maßzahlen berechnet. Es ergibt sich nachstehender Output.

#### SPSS-Output 3.4-1: Deskriptive Maßzahlen für Beispiel 3.4

##### Statistics

Monatsmiete in DM

N	Valid	815
	Missing	0
Mean		1343,4567
Median		1194,3000
Mode		1500,00
Skewness		1,727
Std. Error of Skewness		,086
Kurtosis		4,482
Std. Error of Kurtosis		,171
Percentiles	25	850,0000
	50	1194,3000
	75	1650,6000

Die Ermittlung des Modus ist für diese Verteilung nicht sehr sinnvoll, da mehrere Modi auftreten, d.h., die Verteilung nicht unimodal ist.

Wegen

- Median = 1194,3 < Mittelwert = 1343,4567,
- $x_{0,5} - x_{0,25} = 344,3 < x_{0,75} - x_{0,5} = 456,3$ ,
- einer positiven Schiefe,
- eines positiven Exzeß,

ist die empirische Häufigkeitsverteilung rechtsschief und ist im Vergleich zu einer Normalverteilung gleicher Varianz steiler.

Anhand dieser Daten soll die Berechnung der Perzentile nach Heverage für das erste Quartil demonstriert werden. Es ist somit  $p = 0,25$ . Da die 815 gültigen Fälle einzeln in der Datei aufgelistet sind, hat jeder Variablenwert das gleiche Gewicht  $g_i = 1$  für  $i = 1, \dots, 815$ . Daraus folgt:

$$\begin{aligned} W &= n = 815, \\ W \cdot p &= 815 \cdot 0,25 = 203,75 \\ (W + 1) \cdot p &= 816 \cdot 0,25 = 204. \end{aligned}$$

Die Erfüllung der notwendigen Bedingungen führt zu

$$\begin{aligned} w_{j1} \leq W \cdot p &< w_{j1+1} & 203 < 203,75 < 204 \\ w_{j2} \leq (W + 1) \cdot p &< w_{j2+1} & 204 = 204 < 205 \end{aligned}$$

mit  $j1 = 203$  und  $j2 = 204$ , so dass aufgrund von  $(W + 1)p - w_{j2} = 0 < 1$  und  $g_{(j2+1)} = 1$  die Regel (b) zur Anwendung kommt. Es resultiert:

$$x_{0,25} = (1 - 204 + 204) \cdot x_{(j2)} + [204 - 204] \cdot x_{(j2+1)} = x_{(j2)} = 850.$$

Letter-values, midsummaries, spreads<sup>16</sup>:

Ausgehend von einer geordneten Datenreihe (ohne Missing-Werte) weist jeder Beobachtungswert eine bestimmte Tiefe (depth) auf. Die Tiefe eines Wertes ist definiert als seine Position, bezogen auf das nächstliegende Ende der Datenreihe. Der kleinste und der größte beobachtete Wert weisen somit jeweils eine Tiefe 1 auf.

---

<sup>16</sup>Es wird hier die englische Bezeichnung beibehalten, da sich eine deutsche Übersetzung nicht eingebürgert hat. Siehe u.a. Tukey, J.W. (1977); Fox, J., Long, J.S. (1990); Heiler, S., Michels, P. (1994).

### 3. Prüfung der Verteilungform von Variablen

- Beispiel 3.5:

Für die Variable Arbeitslosenquote der Datei beisp1.sav sind nachfolgend die Beobachtungswerte  $x_{(i)}$  und ihre Tiefen angegeben.

NR.	LAND	ALQ	DEPTH
1	Baden-Wü	3,7	1
2	Bayern	4,4	2
3	Hessen	5,1	3
4	Rheinl-P	5,4	4
5	Schl-Hol	7,3	5
6	Nordr-We	7,9	6
7	Niedersa	8,1	7
8	Saarland	8,6	8
9	Hamburg	8,7	9
10	S-Anhalt	9,1	8
11	Berlin-W	9,4	7
12	Thüring	10,2	6
13	Brandenb	10,3	5
14	Sachsen	10,3	4
15	Bremen	10,7	3
16	Berlin-O	12,2	2
17	M.-Vorp	12,5	1

Nun stelle man sich vor, die Datenreihe wird so übereinander gefaltet, dass der kleinste und der größte Beobachtungswert übereinanderliegen. Dann wird wieder gefaltet, nochmals gefaltet usw. Jedesmal entstehen Knickpunkte in der Datenreihe. Diese werden mit Buchstaben bezeichnet und die zugehörigen Variablenwerte sind die **letter-values**.

Die Tiefe des Medians ist  $d(M) = (n+1)/2$ . Der Median teilt die Beobachtungsreihe in zwei Hälften, in denen jeweils gleichviele Beobachtungswerte liegen. Bei den weiteren letter-values mit geringerer Tiefe treten jeweils zwei Werte auf, da die gleiche Tiefe zweimal auftritt: ein unterer Wert (lower value) und ein oberer Wert (upper value). letter-values sind somit

<u>letter value</u>	<u>zugehörige Tiefe</u>
Median	$d(M) = (n+1)/2$
hinges (Angelpunkte, fourths)	$d(H) = ([d(M)] + 1)/2$
eighths (Achtel)	$d(E) = ([d(H)] + 1)/2$
sixtenths (Sechzehntel)	$d(D) = ([d(E)] + 1)/2$
thirtyseconds (Zweiunddreißigstel)	$d(C) = ([d(D)] + 1)/2$ usw.

Dabei bedeutet  $[d(.)]$ , dass die Tiefe auf eine ganze Zahl gerundet wurde. In Abhängigkeit vom

Stichprobenumfang (d.h., wenn  $n$  genügend groß ist) können weitere letter-values bestimmt werden, die mit den Buchstaben B, A, Z, Y, X, ... bezeichnet werden. Jeder letter-value liegt in der Mitte zwischen dem vorangegangenen letter-value und dem näheren Extremwert. Wenn die Tiefe  $d(\cdot)$  eine gebrochene Zahl ist, wird der Durchschnitt aus den beiden benachbarten Anordnungswerten gebildet. Diese letter-values stimmen nicht notwendig genau mit den Perzentilen überein.

- Beispiel 3.5 (Fortsetzung):

Für das Beispiel sind:

<u>Tiefe</u>	<u>letter-value</u>	
$d(M) = (n+1)/2 = 9$		$x_{0,5} = 8,7$
$d(H) = ([d(M)] + 1)/2 = 5$	$H_L = 7,3$	$H_U = 10,3$
$d(E) = ([d(H)] + 1)/2 = 3$	$E_L = 5,1$	$E_U = 10,7$
$d(D) = ([d(H)] + 1)/2 = 2$	$D_L = 4,4$	$D_U = 12,2$

Unter SPSS kann man die hinges über

■ Analyze

■ Descriptive Statistics

■ Explore

erhalten. Nach Betätigung der Schaltfläche „Statistics...“ im Dialogfeld „Explore“ (Abb. 2.1) erscheint als nächstes das Dialogfeld „Explore: Statistics“ (Abb. 2.3). Durch Ankreuzen von Percentiles erhält man neben den Werten für die Perzentile auch die hinges.

Mittels der letter-values können nun Aussagen über das Zentrum und die Schiefe der Verteilung gewonnen werden, indem von den letter-value-Paaren der Durchschnitt berechnet wird. Diese werden als **midsummaries** bezeichnet. Man erhält also

$$\begin{aligned} \text{mid-hinge : } & \quad \text{mid}(H) = (H_L + H_U)/2 \\ \text{mid-eighth : } & \quad \text{mid}(E) = (E_L + E_U)/2 \text{ usw.} \end{aligned}$$

Wenn die beobachtete Häufigkeitsverteilung symmetrisch ist, dann fallen die midsummaries mit dem Median zusammen, da jeweils der untere und der obere letter-value gleich weit vom Median entfernt liegt. Steigen die midsummaries jedoch tendenziell an, liegt eine rechtsschiefe Verteilung vor (positive Schiefe). Analog ist bei einer abwärts gerichteten Tendenz der midsummaries eine linksschiefe Verteilung (negative Schiefe) gegeben.

### 3. Prüfung der Verteilungform von Variablen

- Beispiel 3.5 (Fortsetzung):

Die midsummaries sind:

$$\begin{aligned} \text{mid-hinge : } & \quad \text{mid}(H) = 8,8 \\ \text{mid-eighth : } & \quad \text{mid}(E) = 7,9 \\ \text{mid-sixteenth : } & \quad \text{mid}(D) = 8,3 \end{aligned}$$

Bildet man die Differenz korrespondierender letter-values, ergeben sich die **spreads**, z.B.

$$\begin{aligned} \text{hinge-spread oder H-spread : } & \quad H_U - H_L = s_H \\ \text{eighth-spread oder E-spread : } & \quad E_U - E_L = S_E \text{ usw.} \end{aligned}$$

Diese spreads können nun verwendet werden, um die empirische Verteilung z.B. mit einer Normalverteilung  $N(\mu; \sigma)$  zu vergleichen und eine Schätzung für  $\sigma$  zu erhalten. Dazu ermittelt man zunächst die theoretischen spreads der Standardnormalverteilung  $N(0; 1)$ , Spalte 3 der folgenden Tabelle.

Für die Hinges ist z.B. die einseitige Wahrscheinlichkeit  $\Phi(-z) = 0,25$  und  $\Phi(z) = 0,75$ . Für diese Wahrscheinlichkeiten findet man in der Tabelle der Verteilungsfunktion der Standardnormalverteilung die Hinges-Werte:  $H_L = -0,675$  und  $H_U = 0,675$ , womit ein hinge-spread von  $s_H = 1,35$  resultiert. Die theoretischen spreads einer  $N(\mu; \sigma)$ -Verteilung ergeben sich durch die Multiplikation der spreads der  $N(0; 1)$  mit  $\sigma$  (Spalte 4 der Tabelle 3.1). Im Umkehrschluß gilt nun für große Stichproben: Wenn die Stichprobe aus einer  $N(\mu; \sigma)$ -verteilten Grundgesamtheit stammt, müssen die spreads der empirischen Verteilung in etwa den theoretischen spreads der  $N(\mu; \sigma)$  entsprechen.

Tabelle 3.1.: Theoretische spreads der  $N(0; 1)$

letter-value	einseitige Wahrscheinlichkeit	spread der $N(0; 1)$	spread der $N(\mu; \sigma)$	Schätzung für $\sigma$
H	1/4	1,349	$1,349 \cdot \sigma$	$\hat{\sigma} = s_H / 1,349$
E	1/8	2,301	$2,301 \cdot \sigma$	$\hat{\sigma} = s_E / 2,301$
D	1/16	3,068	$3,068 \cdot \sigma$	$\hat{\sigma} = s_D / 3,068$
C	1/32	3,726	$3,726 \cdot \sigma$	$\hat{\sigma} = s_C / 3,726$
B	1/64	4,308	$4,308 \cdot \sigma$	$\hat{\sigma} = s_B / 4,308$
A	1/128	4,835	$4,835 \cdot \sigma$	$\hat{\sigma} = s_A / 4,835$
Z	1/256	5,320	$5,320 \cdot \sigma$	$\hat{\sigma} = s_Z / 5,320$
Y	1/512	5,771	$5,771 \cdot \sigma$	$\hat{\sigma} = s_Y / 5,771$
X	1/1024	6,195	$6,195 \cdot \sigma$	$\hat{\sigma} = s_X / 6,195$

Eine Schätzung von  $\sigma$  erhält man, indem die empirischen spreads durch die theoretischen

spreads der  $N(0; 1)$  dividiert werden (Spalte 5 der Tabelle 3.1). Die  $\hat{\sigma}$ -Werte werden auch als Pseudosigmas bezeichnet. Es müssen sich annähernd gleiche Werte für  $\hat{\sigma}$  ergeben, wenn die (große) Stichprobe aus einer Normalverteilung stammt. Ist die empirische Verteilung im wesentlichen symmetrisch, aber an den Enden stärker besetzt bzw. treten Ausreißer auf, so ist z.B.  $\hat{\sigma} = s_H/1,349$  ein robuster Schätzer für  $\sigma$ . Diese Aussagen gelten nur für große Stichproben, für mittlere und kleinere Stichproben müssen Korrekturen der Werte vorgenommen werden.

- Beispiel 3.6:

Für die Variable monatliches Haushaltsnettoeinkommen der Datei allbus.sav werden in der Tabelle 3.2 die letter-values (von denen aber nur die hinges unter SPSS verfügbar sind), die midsummaries, die spreads und die Pseudosigmas angegeben. Die Anzahl der gültigen Werte dieser Variablen ist  $n = 1351$ .

Für die nach der Variablen monatliches Nettoeinkommen geordneten Datei allbus.sav ergeben sich die Werte in der Zeile der H (für die Hinges) der Tabelle 3.2 z.B. wie folgt:

$$\begin{aligned}
 (d(M) + 1)/2 &= 677/2 = 338,5 \\
 x_{(338)} &= 1602, \quad x_{(339)} = 1630 \text{ und somit } H_L = (1602 + 1630)/2 = 1616, \\
 x_{(1351-339)} &= x_{(1012)} = 3400, \quad x_{(1351-338)} = x_{(1013)} = 3400 \text{ und somit } H_U = 3400, \\
 mid(H) &= (1616 + 3400)/2 = 2508, \\
 s_H &= 3400 - 1616 = 1748, \quad \hat{\sigma} = 1784/1,349 = 1322,46.
 \end{aligned}$$

Tabelle 3.2.: letter-values, midsummaries, spreads und Pseudosigmas

letter-value	depth	lower value	upper value	midsummary	spread	Pseudo-sigma
M	676	2400	2400	2400		
H	338,5	1616	3400	2508	1784	1322,46
E	169,5	1200	4200	2700	3000	1303,78
D	85	920	5000	2960	4080	1329,86
C	43	800	5500	3150	4700	1261,41
B	22	600	6000	3300	5400	1253,48
A	11,5	518,5	7000	3759,25	6481,5	1340,54
Z	6	480	8000	4240	7520	1413,53

Deutlich zu erkennen ist der ansteigende Trend der midsummaries, der auf eine rechtsschiefe Verteilung des monatlichen Haushaltsnettoeinkommens hinweist. Nur im Bereich  $D_L$  bis  $D_U$  sind die Pseudosigmas annähernd gleich, d.h. nur in diesem Bereich können die Werte als grob

### 3. Prüfung der Verteilungform von Variablen

normalverteilt angesehen werden. Zur Veranschaulichung kann man sich noch den Stem-and-Leaf Plot und den Boxplot ausgeben lassen.

Der im Abschnitt 2.2 behandelte Boxplot wird oft unter Verwendung der hinges und des hinge-spread gezeichnet. Der untere und der obere Rand der Box entspricht  $H_L$  und  $H_U$ . In der Box wird der Median durch eine Linie markiert. Ungewöhnliche Beobachtungswerte werden dadurch sichtbar gemacht, dass sogenannte Zäune (fences) gezogen werden: die inner fences (innere Zäune) beim 1,5-fachen des hinges-spread von den hinges entfernt:

$$f_L = H_L - 1,5 \cdot s_H \text{ und } f_U = H_U + 1,5 \cdot s_H,$$

sowie die outer fences (äußersten Zäune) beim 3-fachen des hinges-spread von den hinges entfernt:

$$F_L = H_L - 3 \cdot s_H \text{ und } F_U = H_U + 3 \cdot s_H.$$

Werte die außerhalb der inner fences liegen, werden als outside (außerhalb) und Werte, die außerhalb der outer fences liegen als far outside (weit außerhalb) bezeichnet.

## 3.3. Statistische Tests

Die bisher angegebenen Möglichkeiten zur Prüfung der Verteilungsform einer Variablen geben nur erste Anhaltspunkte über die Gestalt der Häufigkeitsverteilung der betrachteten Variablen. Sie sind jedoch in ihren Ergebnissen recht subjektiv und treffen keine Aussage darüber, wie gut die empirische Verteilung mit einem theoretischen Verteilungsmodell übereinstimmt. Um diesbezüglich einen Schritt voranzukommen und eine statistisch abgesicherte Antwort zu erhalten, werden statistische Tests, sogenannte Anpassungstests (goodness-of-fit-tests), durchgeführt. Anpassungstests sind Tests zur Überprüfung der Hypothese, ob die Verteilung  $F_n(x)$  der Stichprobe  $(X_1, \dots, X_n)$  aus einer Grundgesamtheit mit der Verteilung  $F_0(x)$  stammt. Sie gehören zu den nichtparametrischen Tests.

Die generelle Vorgehensweise bei Anpassungstests ist im Prinzip wie bei Parametertests. Es wird eine Teststatistik konstruiert, die die Informationen über die hypothetische Verteilung sowie die Verteilung in der Zufallsstichprobe enthält und auf deren Basis eine Aussage über die Nullhypothese möglich ist. Die Verteilung der Teststatistik muss unter der Nullhypothese (zumindest approximativ) bekannt sein. Auch bei Anpassungstests wird stets die Nullhypothese statistisch geprüft und in Abhängigkeit von der Testentscheidung besteht die Möglichkeit, einen Fehler 1. Art mit der Wahrscheinlichkeit  $\alpha$  bzw. einen Fehler 2. Art mit der Wahrscheinlichkeit  $\beta$  zu begehen. Mit dem vorgegebenen Signifikanzniveau  $\alpha$  kann die Wahrscheinlichkeit eines Fehlers 1. Art niedrig gehalten werden; die Wahrscheinlichkeit eines Fehlers 2. Art ist dagegen in der Regel nicht bekannt. Man wird deshalb bestrebt sein, die Nullhypothese abzulehnen, da dann die statistische Sicherheit einer Fehlentscheidung bekannt ist.

Wenn die hypothetische Verteilung die wahre Verteilung in der Grundgesamtheit ist, dann ist zu erwarten, dass diese Verteilung im Prinzip auch in der Stichprobe zu beobachten ist. Im Prinzip bedeutet dabei, dass Abweichungen zwischen der beobachteten Verteilung in der Stichprobe und der unter der Verteilungsannahme erwarteten Verteilung in der Stichprobe in der Regel immer auftreten werden. Zu entscheiden ist, ob die Abweichungen noch zufallsbedingt sind, oder ob es sich um signifikante Abweichungen handelt. Um die erwartete Verteilung in der Stichprobe ermitteln zu können, muss unter der Nullhypothese angenommen werden, dass genau die hypothetische Verteilung die wahre Verteilung in der Grundgesamtheit ist. Damit lautet das Hypothesenpaar stets:

$H_0$ : Die Zufallsvariable X in der Grundgesamtheit weist die hypothetische Verteilung auf.

$H_1$ : Die Zufallsvariable X in der Grundgesamtheit weist eine andere als die hypothetische Verteilung auf.

Große Abweichungen zwischen der beobachteten Verteilung und der erwarteten Verteilung in der Stichprobe deuten tendenziell auf eine falsche Verteilungsannahme hin, d.h., man wird die Nullhypothese ablehnen.

Es sei darauf hingewiesen, dass das Signifikanzniveau  $\alpha$  stets **vor** der Testdurchführung festzulegen ist. Allgemeinere Ausführungen zur Testentscheidung unter Verwendung statistischer Software sind im Anhang A enthalten.

### 3.3.1. Kolmogorov-Smirnov-Test

Die Voraussetzungen des Kolmogorov-Smirnov-Tests<sup>17</sup> sind:

- Die Variable X muß metrisches Skalenniveau aufweisen.
- Die Variable X darf nicht klassiert vorliegen.
- Die theoretische Verteilung  $F_0(x)$  muß stetig sein. Ist dies nicht gegeben, so wird die Nullhypothese länger als notwendig beibehalten (konservativer Test).
- Die Parameter der hypothetischen Verteilung  $F_0(x)$  müssen vollständig bekannt sein.

Geprüft wird die Nullhypothese

$$H_0 : F_n(x) = F_0(x) \text{ für alle } x$$

gegen die Alternativhypothese

$$H_1 : F_n(x) \neq F_0(x) \text{ für mindestens ein } x.$$

Wird die empirische Verteilung beispielsweise gegen die Normalverteilung  $N(\mu; \sigma)$  geprüft, so

---

<sup>17</sup>Siehe u.a. Hartung, Elpelt, Klösener (1993), S. 183 ff.; Büning, Trenkler (1978), S. 85 ff.; Rönnz, Strohe (Hrsg.) (1994), S. 185 ff.

### 3. Prüfung der Verteilungform von Variablen

ist  $F_0(x) = \Phi([x - \mu]/\sigma)$  mit spezifizierten Werten von  $\mu$  und  $\sigma$ .

Als Teststatistik wird

$$D_n = \max_x |F_n(x) - F_0(x)| \quad (3.18)$$

verwendet.  $D_n$  beinhaltet den größten absoluten vertikalen Abstand zwischen empirischer und hypothetischer Verteilungsfunktion. Unter der Nullhypothese hängt die Verteilungsfunktion von  $D_n$  nur von  $n$ , jedoch nicht von  $F_0$  ab.

Bei der praktischen Bestimmung der Teststatistik  $D_n$  ist zu berücksichtigen, dass die empirische diskrete Verteilungsfunktion eine Treppenfunktion ist. Die Abweichungen zwischen  $F_n(x)$  und  $F_0(x)$  sind deshalb sowohl an der unteren als auch an der oberen Sprungstelle der geordneten  $x_{(i)}$  zu berechnen:

$$\begin{aligned} D_n^1 &= \max |F_n(x_{(i-1)}) - F_0(x_{(i)})| \\ D_n^2 &= \max |F_n(x_{(i)}) - F_0(x_{(i)})|. \end{aligned}$$

Die maximale Abweichung ergibt sich zu:  $D_n = \max(D_n^1, D_n^2)$ .

Entspricht die beobachtete Verteilung der hypothetischen Verteilung, so werden unter  $H_0$  die Abweichungen zwischen  $F_n(x)$  und  $F_0(x)$  nur gering und vom Zufall bestimmt sein. Ist  $D_n \geq d_{n;1-\alpha}$ , so wird  $H_0$  abgelehnt. Dabei ist  $d_{n;1-\alpha}$  das Quantil der Ordnung  $1 - \alpha$  der Verteilung von  $D_n$ ,  $n$  der Stichprobenumfang und  $\alpha$  das vorgegebene Signifikanzniveau. Eine Tabelle mit ausgewählten Quantilen ist im Anhang B enthalten.<sup>18</sup>

Für  $n \rightarrow \infty$  konvergiert die Verteilungsfunktion von  $Z_n = D_n \cdot n^{1/2}$  bei Gültigkeit von  $H_0$  gegen die Kolmogorov-Verteilung und es gilt die asymptotische Beziehung

$$d_{n;1-\alpha} \sqrt{n} \approx k_{1-\alpha}, \quad (3.19)$$

wobei  $k_{1-\alpha}$  ein Quantil der Kolmogorov-Verteilung ist. Asymptotische kritische Werte der Kolmogorov-Verteilung sind für großes  $n$  und ausgewählte  $\alpha$ <sup>19</sup>:

$\alpha$	0,20	0,10	0,05	0,02	0,01
$k_{1-\alpha}$	1,073	1,224	1,358	1,517	1,628

Folgendes einfaches Beispiel soll die Bestimmung der Kolmogorov-Smirnov-Teststatistik verdeutlichen.

- Beispiel 3.7:

Aus einer Grundgesamtheit mit  $\mu = 22,3125$  und  $\sigma = 18,65947$  wird eine einfache Zufallsstichprobe vom Umfang  $n = 16$  gezogen. Es soll auf dem 5%-Niveau geprüft werden, ob die Verteilung

<sup>18</sup>Siehe u.a. auch Büning, Trenkler (1978), S. 372; Müller, H., Neumann, P., Storm, R. (1973), S.238 ff.

<sup>19</sup>Entnommen aus: Müller, H., Neumann, P., Storm, R. (1973), S. 244. Hartung, Elpelt, Klösener (1993), S.

184, geben auch kritische Werte für kleine  $n$  ( $n < 40$ ) an.

der Variablen X in dieser Grundgesamtheit eine Normalverteilung ist:  $F_0(x) = N(\mu; \sigma)$ . In der folgenden Tabelle 3.3 sind die Ausgangswerte  $x_{(i)}$  der Variablen X und notwendige Zwischenberechnungen enthalten.

Tabelle 3.3.: Zwischenberechnungen für den Kolmogorov-Smirnov-Test

$x_{(i)}$	$z_{(i)}$	$f(x_{(i)})$	$F_n(x_{(i)})$	$F_n(x_{(i-1)})$	$\Phi(z_{(i)})$	$D_n^1$	$D_n^2$
,4	-1,17434	,0625	,0625	,0000	,12013	<i>-,12013</i>	-,05763
,8	-1,15290	,0625	,1250	,0625	,12448	-,06198	,00052
,9	-1,14754	,0626	,1875	,1250	,12558	-,00058	,06192
2,6	-1,05643	,0625	,2500	,1875	,14538	,04212	,10462
15,7	-,35438	,0625	,3125	,2500	,36153	<i>-,11153</i>	-,04903
16,3	-,32222	,0625	,3750	,3125	,37364	-,06114	,00136
18,3	-,21504	,0625	,4375	,3750	,41487	-,03987	,02263
19,8	-,13465	,0625	,5000	,4375	,44644	-,00894	,05356
20,4	-,10249	,0625	,5626	,5000	,45918	,04082	,10332
21,1	-,06498	,0625	,6250	,5625	,47409	,08841	,15091
23,8	,07972	,0625	,6875	,6250	,53177	,09323	<b>,15573</b>
29,1	,36376	,0625	,7500	,6875	,64198	,04552	,10802
34,1	,63172	,0625	,8125	,7500	,73621	,01379	,07629
35,8	,72282	,0625	,8750	,8125	,76511	,04739	,10989
47,3	1,33913	,0625	,9375	,8750	,90974	-,03474	,02776
70,6	2,58783	,0625	1,0000	,9375	,99517	-,05767	,00483

Die  $z_{(i)}$ -Werte in Spalte 2 ergeben sich durch die Transformation

$$z_{(i)} = (x_{(i)} - \mu)/\sigma = (x_{(i)} - 22,3125)/18,65947.$$

Die empirische Verteilungsfunktion  $F_{16}(x)$  in Spalte 4 entsteht als Treppenfunktion durch Summation der relativen Häufigkeiten der Einzelwerte  $f(x_{(i)}) = 1/16 = 0,0625$  in Spalte 3. Die Werte von  $\Phi(z_{(i)})$  in Spalte 6 erhält man aus der Tabelle der Verteilungsfunktion der Standardnormalverteilung für die Werte  $z_{(i)}$  aus der Spalte 2.

Die größte Differenz  $D_n^1$  wurde in der Spalte 7 der Tabelle kursiv und die größte Differenz  $D_n^2$  in Spalte 8, die auch die größte absolute Differenz ist, fett geschrieben. Der Wert der Teststatistik ist damit  $D_{16} = 0,15573$  und für die Teststatistik  $Z_n = D_n n^{1/2}$  folgt  $Z_{16} = 0,15573 \cdot 4 = 0,62292$ .

Für das Signifikanzniveau  $\alpha = 0,05$  und den Stichprobenfunktion  $n = 16$  findet man den kritischen Wert in der Tabelle im Anhang B:  $d_{16;0,95} = 0,32733$ .

Der Ablehnungsbereich der  $H_0$  ist damit  $\{D_{16}|D_{16} > 0,32733\}$  und der Nichtablehnungsbereich  $\{D_{16}|D_{16} \leq 0,32733\}$ . Da  $D_{16} = 0,15573 < d_{16;0,95} = 0,32733$  ist, besteht keine Veranlassung, die Nullhypothese zu verwerfen.

### 3. Prüfung der Verteilungform von Variablen

Zu dem gleichen Ergebnis gelangt man, wenn die Teststatistik  $Z_n$  mit dem kritischen Wert der Kolmogorov-Verteilung verglichen wird. Aus der Tabelle der kritischen Werte der Kolmogorov-Verteilung findet man für  $1 - \alpha = 0,95$  und  $n = 20$ :  $k_{20;0,95} = 1,31$  (Hartung, Elpelt, Klösener (1993), S. 184), womit ebenfalls der Wert der Teststatistik kleiner als der (näherungsweise) kritische Wert ist und die Nullhypothese nicht abgelehnt wird.

In der Regel wird man nur die Klasse von Verteilungsfunktionen, aber nicht die Parameter der empirischen Verteilung kennen. Es wird z.B. die empirische Verteilungsfunktion einer Variablen X oft gegen die Normalverteilung getestet, wobei aber  $\mu$  und  $\sigma^2$  der Grundgesamtheit unbekannt sind. Diese unbekannten Parameter der Normalverteilung müssen dann durch die Schätzungen  $\bar{x}$  und s aus der Stichprobe ersetzt werden. Die Nullhypothese ist dann wie folgt abzuändern:

$H_0^*$ : Die Stichprobe stammt aus einer Grundgesamtheit mit einer Normalverteilung  $N(\mu; \sigma)$ , mit  $\mu$  und  $\sigma^2$  unbekannt.

Die Teststatistik ist in diesem Fall

$$D_n^{NV} = \max_x \left| F_n(x) - \Phi\left(\frac{x - \bar{x}}{s}\right) \right|, \quad (3.20)$$

wobei  $\bar{x}$  und s die aus der Stichprobe geschätzten Werte für Mittelwert und Standardabweichung sind und das hochgestellte NV die Prüfung gegen die Normalverteilung verdeutlichen soll. Die Verteilung dieser Teststatistik ist nicht mehr die oben angegebene Kolmogorov-Smirnov-Verteilung. Für die Teststatistik  $Z_n^{NV} = D_n^{NV} \cdot n^{1/2}$  gibt Lillefors (1967) Tabellen mit kritischen Quantilen an. Für ausgewählte n und  $1 - \alpha$  sind kritische Werte auch bei Hartung, Elpelt, Klösener (1993), S. 185, zu finden.

Werden diese kritischen Werte mit  $L_{n;1-\alpha}^{NV}$  bezeichnet, so wird  $H_0^*$  zum Signifikanzniveau  $\alpha$  abgelehnt, wenn  $Z_n^{NV} > L_{n;1-\alpha}^{NV}$  ist.

Tabellen kritischer Quantile zum Test gegen die Exponentialverteilung  $F_0(x) = 1 - e^{-\lambda x}$  bei unbekanntem  $\lambda$  sind bei Lillefors (1969) zu finden.

Der Kolmogorov-Smirnov-Test ist im Gegensatz zum  $\chi^2$ -Anpassungstest (siehe weiter unten), der eine Klasseneinteilung mit Mindestzahlen der Besetzung voraussetzt, auch für kleine Stichproben anwendbar. Er reagiert aber empfindlicher auf Abweichungen in der Form der Verteilungsfunktion.

Unter SPSS gibt es zwei Möglichkeiten, den Kolmogorov-Smirnov-Test durchführen zu lassen, die beide unterschiedlich zu interpretieren sind.

#### 1. Variante:

Bei der ersten Möglichkeit wird der Kolmogorov-Smirnov-Test zusammen mit den Wahrscheinlichkeitsplots (siehe Abschnitt 3.1) durchgeführt. Der Weg des Aufrufes ist wie folgt:

## ■ Analyze

### ■ Descriptive Statistics

#### ■ Explore.

Im Dialogfeld „Explore“ (Abb. 2.1) ist die zu testende Variable in das Feld „Dependent List:“ zu bringen, bei Display „Plots“ zu aktivieren, die Schaltfläche „Plots...“ zu betätigen und im erscheinenden Dialogfeld „Explore: Plots“ (Abb. 2.2) nur auf „Normality plots with tests“ zu entscheiden. Neben den Plots werden die Ergebnisse des Kolmogorov-Smirnov-Tests im SPSS-Viewer ausgegeben.

Bei dieser Variante der Testdurchführung in SPSS wird stets die empirische Verteilung gegen die Normalverteilung  $N(\mu; \sigma)$  unter Verwendung der geschätzten Werte von Mittelwert  $\bar{x}$  und Standardabweichung  $s$  geprüft und der Wert der Teststatistik  $Z_n$  mit den kritischen Werten von Lillefors verglichen. Ausgegeben werden der empirische Wert der Teststatistik  $D_n$  (Statistic), die Anzahl der Freiheitsgrade (df) und die Überschreitungswahrscheinlichkeit des Testwertes (Sig.). Diese unter Sig. angegebene Wahrscheinlichkeit ist mit dem vorgegebenen Signifikanzniveau  $\alpha$  zu vergleichen. Die Testentscheidung ist wie folgt:

Wenn  $Sig. \leq \alpha$  ist, wird die Nullhypothese  $H_0$  aufgrund der Stichprobe vom Umfang  $n$  und zum vorgegebenen Signifikanzniveau  $\alpha$  abgelehnt; wenn  $Sig. > \alpha$  ist, besteht keine Veranlassung, die Nullhypothese zu verwerfen.

#### 2. Variante:

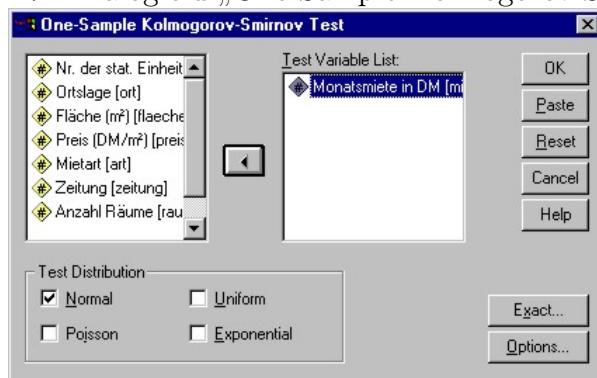
Bei der 2. Möglichkeit wird der Kolmogorov-Smirnov-Test als nichtparametrischer Test aufgerufen:

## ■ Analyze

### ■ Nonparametric Tests

#### ■ 1-Sample K-S...

Abbildung 3.27.: Dialogfeld „One-Sample Kolmogorov-Smirnov-Test“



### 3. Prüfung der Verteilungform von Variablen

In diesem Dialogfeld wird die zu testende Variable aus der linken Variablenliste in das Feld „Test Variable List:“ gebracht. Bei dieser Variante der Durchführung des Kolmogorov-Smirnov-Tests kann die empirische Verteilung gegen eine Normalverteilung (Voreinstellung), gegen eine Gleichverteilung (Uniform), gegen eine Poisson-Verteilung oder gegen eine Exponentialverteilung geprüft werden. Über die Schaltfläche „Options...“ gelangt man in ein weiteres Dialogfeld, in dem zusätzlich Maßzahlen der deskriptiven Statistik und die Quartile angefordert werden können. Im SPSS Viewer erscheint u.a.:

- die Anzahl der gültigen Fälle (N),
- die Verteilung mit ihren Parametern, gegen die geprüft wurde,
- die größte absolute, positive und negative Abweichung, wobei die erste Angabe  $D_n$  entspricht,
- der Wert der Teststatistik (Kolmogorov-Smirnov Z), die bei dieser Testdurchführung  $Z_n = D_n \cdot n^{1/2}$  ist,
- die zweiseitige Überschreitungswahrscheinlichkeit (Asymp.Sig. 2-tailed), deren Interpretation in der gleichen Weise wie bei der o.g. ersten Möglichkeit zu erfolgen hat.

Bei dieser Variante der Testdurchführung in SPSS ist folgendes zu beachten: Die empirische Verteilung wird gegen eine ausgewählte theoretische Verteilung geprüft, deren Parameter per Voreinstellung durch die aus der Stichprobe geschätzten Werte ersetzt werden; also z.B. gegen die Normalverteilung  $N(\mu; \sigma)$  unter Verwendung der geschätzten Werte von Mittelwert  $\bar{x}$  und Standardabweichung s. Der Wert der Teststatistik  $Z_n$  wird jedoch mit den kritischen Werten der Kolmogorov-Verteilung verglichen. Dies wäre nur korrekt, wenn die Verteilung in der Grundgesamtheit auch tatsächlich die Parameterwerte der Stichprobe aufweist.

Dieses Handikap kann jedoch beseitigt werden, indem die Parameter der Verteilung, falls sie für die Grundgesamtheit bekannt sind, über die Syntax gesetzt werden. Dazu ist in folgender Weise vorzugehen:

Zunächst werden, wie bei dieser Variante beschrieben, alle Entscheidungen über die Menüs und Dialogfelder getroffen. Statt zum Schluß jedoch die Schaltfläche „OK“ zu betätigen, wird die Schaltfläche „Paste“ angeklickt. Die hinter den getroffenen Entscheidungen stehende SPSS-Syntax erscheint im Syntax-Editor:

a) wenn Normalverteilung gewählt wurde

NPART TESTS

/K-S(NORMAL)=varlist

/MISSING ANALYSIS.

b) wenn Gleichverteilung gewählt wurde

NPAR TESTS

/K-S(UNIFORM)=varlist

/MISSING ANALYSIS.

c) wenn Poisson-Verteilung gewählt wurde

NPAR TESTS

/K-S(POISSON)=varlist

/MISSING ANALYSIS.

d) wenn Exponentialverteilung gewählt wurde

NPAR TESTS

/K-S(EXPONENTIAL)=varlist

/MISSING ANALYSIS.

Dabei steht im konkreten Fall statt varlist die Variable(n), für die der Test durchgeführt werden soll. Eine Veränderung der voreingestellten Parameter (Schätzwerte aus der Stichprobe) kann vorgenommen werden, indem

a) bei Normalverteilung

NPAR TESTS

/K-S(NORMAL **mean, stddev**)=varlist

/MISSING ANALYSIS.

b) bei Gleichverteilung

NPAR TESTS

/K-S(UNIFORM **min, max**)=varlist

/MISSING ANALYSIS.

c) bei Poisson-Verteilung

NPAR TESTS

/K-S(POISSON **mean**)=varlist

/MISSING ANALYSIS.

d) bei Exponentialverteilung

NPAR TESTS

/K-S(EXPONENTIAL **mean**)=varlist

/MISSING ANALYSIS.

die (hier fett gedruckten) Angaben für mean, stddev bzw. min, max bzw. mean durch die konkreten Parameterwerte der theoretischen Verteilung zu ersetzen sind; z.B.

/K-S(NORMAL 20, 5)=varlist

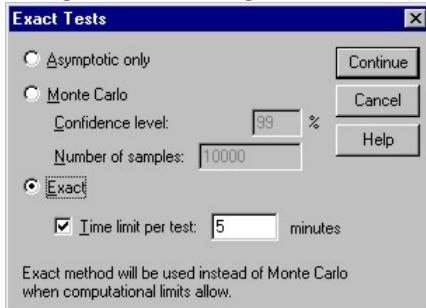
Wurde derartig verfahren, dann ist auch der Vergleich der Teststatistik  $Z_n$  mit den kritischen

### 3. Prüfung der Verteilungform von Variablen

Werten der Kolmogorov-Verteilung in Ordnung.

Bei der 2. Variante gibt es darüber hinaus die Möglichkeit, exakte Ergebnisse bei der Berechnung des Signifikanzniveaus (Überschreitungswahrscheinlichkeit) zu erhalten, wenn die Voraussetzungen der asymptotischen Methode durch die Daten nicht erfüllt sind. Dazu ist die Schaltfläche „Exact...“ im Dialogfeld „One-Sample Kolmogorov-Smirnov-Test“ (Abb. 3.27) zu betätigen und im sich öffnenden Dialogfeld „Exact Tests“ auf Exact zu entscheiden.

Abbildung 3.28.: Dialogfeld „Exact Tests“



- Beispiel 3.4 (Fortsetzung):

Für die Variable Monatsmiete der Datei mieten.sav soll die Prüfung auf Normalverteilung  $N(\mu; \sigma)$  mittels des Kolmogorov-Smirnov-Tests auf dem 5%-Niveau durchgeführt werden. Da  $\mu$  und  $\sigma$  der Grundgesamtheit unbekannt sind und durch die Schätzungen aus der Stichprobe ersetzt werden müssen, wird unter SPSS die 1. Variante des Kolmogorov-Smirnov-Tests mit den kritischen Werten von Lillefors gewählt. Da  $Sig. < \alpha = 0,05$  ist, wird die Nullhypothese verworfen.

**SPSS-Output 3.4-2:** Test auf Normalverteilung der Variablen Monatsmiete nach der

1. Variante

**Test of Normality**

	Kolmogorov-Smirnov <sup>a</sup>		
	Statistic	df	Sig.
Monatsmiete in DM	,117	815	,000

a. Lillefors Significance Correction.

Nur zur Demonstration soll auch der Output der 2. Variante angegeben werden.

**SPSS-Output 3.4-2:** Test auf Normalverteilung der Variablen Monatsmiete nach der

2. Variante

**One-Sample Kolmogorov-Smirnov-Test**

		Monatsmiete in DM
N		815
Normal Parameters <sup>a,b</sup>	Mean	1343,4567
	Std. Deviation	738,1954
Most Extreme	Absolute	,117
Differences	Positive	,117
	Negative	-,077
Kolmogorov-Smirnov Z		3,330
Asymp. Sig. (2-tailed)		,000

a. Test distribution is Normal.

b. Calculated from data.

Es wird zunächst angegeben, dass die theoretische Verteilung eine Normalverteilung mit den Parametern Mean = 1343,4567 und Standard Deviation = 738,1954 ist, die aber die aus den Daten geschätzten Werte sind.  $Z_n = 3,33$  wird hier mit den kritischen Werten der Kolmogorov-Verteilung verglichen.

### 3.3.2. Chi-Quadrat-Anpassungstest

$\chi^2$ -Test ist eine zusammenfassende Bezeichnung für Tests, deren Teststatistik einer  $\chi^2$ -Verteilung<sup>20</sup> genügt. Der zur Prüfung der Verteilungsform einer Variablen in Frage kommende Test ist der  $\chi^2$ -Anpassungstest<sup>21</sup>. Der Test basiert auf einer einfachen Zufallsstichprobe vom vorgegebenen Umfang n. Das Signifikanzniveau  $\alpha$  ist vor der Testdurchführung festzulegen.

Gegeben ist eine Zufallsvariable X in der Grundgesamtheit mit der Verteilung F(x), wobei an das Skalenniveau von X keine Voraussetzungen gestellt werden. Die Verteilung F(x) ist unbekannt. Es existiert jedoch eine Annahme, dass X die hypothetische Verteilung  $F_0(x)$  besitzt.

Ist X eine diskrete Zufallsvariable (darunter werden im weiteren summarisch nominalskalierte, ordinalskalierte sowie diskrete Zufallsvariablen mit sehr wenigen Ausprägungen verstanden), kann sie die Werte  $x_1, \dots, x_k$  annehmen. Es bezeichne:

- $h(x_j) = h_j$  die beobachtete absolute Häufigkeit des Wertes  $x_j$  in der Stichprobe,  $j = 1, \dots, k$ ;
- $P(X = x_j)$  die Wahrscheinlichkeit, dass die Zufallsvariable X den Wert  $x_j$  annimmt,  $j = 1, \dots, k$ .

<sup>20</sup>Vgl. u.a. Rönnz, B., Strohe, H.G. (Hrsg.) (1994), S. 69 - 71

<sup>21</sup>Vgl. u.a. Büning, H., Trenkler, G. (1978), S. 92 ff.; Schlittgen, R. (1990), S. 369 ff.

### 3. Prüfung der Verteilungform von Variablen

Ist X eine stetige Zufallsvariable (darunter werden im weiteren auch die diskreten Zufallsvariablen mit sehr vielen bzw. unendlich vielen Ausprägungen, d.h. die sogenannten quasi-stetigen Zufallsvariablen, gefaßt), muss eine Intervallbildung der beobachteten Werte in disjunkte, aneinander angrenzende Klassen erfolgen. Mit k als Anzahl der Klassen ( $k \geq 2$ ) können die Klassen allgemein wie folgt geschrieben werden:  $(x_0^*, x_1^*], (x_1^*, x_2^*], \dots, (x_{k-1}^*, x_k^*]$  bzw.  $(x_{j-1}^*, x_j^*]$  für  $j = 1, \dots, k$ . Es bezeichne im stetigen Fall:

- $h(x_{j-1}^* < X \leq x_j^*) = h_j$  die beobachtete absolute Häufigkeit der j-ten Klasse in der Stichprobe,  $j = 1, \dots, k$ ;
- $P(x_{j-1}^* < X \leq x_j^*)$  die Wahrscheinlichkeit, dass die Zufallsvariable X einen Wert aus der Klasse  $(x_{j-1}^*, x_j^*]$  annimmt,  $j = 1, \dots, k$ .

Ist X eine diskrete Zufallsvariable, erhält man  $p_j = P(X = x_j | H_0)$  aus der vorgegebenen Wahrscheinlichkeitsfunktion.

Für eine stetige Zufallsvariable X ist die Wahrscheinlichkeit, dass X einen bestimmten Wert x annimmt, jedoch stets Null. Daraus folgt die Notwendigkeit einer Intervallbildung der beobachteten Werte. Die Wahrscheinlichkeit  $p_j = P(x_{j-1}^* < X \leq x_j^* | H_0)$ , dass die stetige Zufallsvariable X einen Wert aus der Klasse  $(x_{j-1}^*, x_j^*]$  annimmt, kann dann mittels der vorgegebenen Verteilungsfunktion bestimmt werden.

Es sei jedoch angemerkt, dass auch für eine diskrete Zufallsvariable eine Klassenbildung vorgenommen werden kann, falls es die Problemstellung erfordert.

Hypothesenformulierung:

Es werden die gleiche Null- und Alternativhypothese wie beim Kolmogorov-Smirnov-Test geprüft:

$$H_0 : F_n(x) = F_0(x) \text{ für alle } x,$$

$$H_1 : F_n(x) \neq F_0(x) \text{ für mindestens ein } x.$$

Das dem Chi - Quadrat - Anpassungstest zugrundeliegende Hypothesenpaar lautet speziell:

- wenn X diskret ist

$$H_0 : P(X = x_j) = p_j \text{ für alle } j = 1, \dots, k$$

$$H_1 : P(X = x_j) \neq p_j \text{ für mindestens ein } j$$

- wenn X stetig ist

$$H_0 : P(x_{j-1}^* < X \leq x_j^*) = p_j \text{ für alle } j = 1, \dots, k$$

$$H_1 : P(x_{j-1}^* < X \leq x_j^*) \neq p_j \text{ für mindestens ein } j.$$

Dabei bezeichnet  $p_j$  ( $j = 1, \dots, k$ ) sowohl im diskreten als auch im stetigen Fall die Wahrscheinlichkeit, dass die Zufallsvariable X den Wert  $x_j$  annimmt bzw. in die j-te Klasse  $(x_{j-1}^*, x_j^*]$  fällt,

wenn die hypothetische Verteilung  $F_0(x)$  zugrundegelegt wird, d.h., wenn die Nullhypothese  $H_0$  gilt:

$$p_j = P(X = x_j | H_0) \text{ bzw. } p_j = P(x_{j-1}^* < X \leq x_j^* | H_0).$$

Die  $p_j$  können bestimmt werden durch die Vorgabe

- einer vollständig spezifizierten theoretischen Verteilung, d.h. Verteilungstyp inklusive sämtlicher Parameter.

Beispiel: Die Annahme besagt, dass die Zufallsvariable  $X$  eine Poisson-Verteilung  $PO(\lambda)$  mit vorgegebenem Parameter  $\lambda$  besitzt.

- einer theoretischen Verteilung mit unbekannten Parametern, d.h., nur der Verteilungstyp ist in der Annahme vorgegeben, die Parameter müssen aus der Stichprobe geschätzt werden.

Beispiel: Die Annahme besagt, dass die Zufallsvariable  $X$  eine Normalverteilung  $N(\mu; \sigma)$  mit unbekanntem Erwartungswert  $\mu$  und unbekannter Standardabweichung  $\sigma$  aufweist, so dass diese beiden Parameter erst aus der Stichprobe zu schätzen sind.

- einer Häufigkeitsverteilung

Beispiel: Die Zufallsvariable  $X$  habe vier mögliche Realisationen. Es wird angenommen, dass diese mit den fest vorgegebenen Wahrscheinlichkeiten bzw. relativen Häufigkeiten  $p_1 = 0,2$ ,  $p_2 = 0,4$ ,  $p_3 = 0,1$  und  $p_4 = 0,3$  auftreten.

Der Test basiert auf dem Vergleich der in der Stichprobe beobachteten Verteilung und der bei Gültigkeit der Nullhypothese in der Stichprobe erwarteten Verteilung. Für die Bestimmung der Teststatistik wird von den absoluten Häufigkeiten ausgegangen. Für die konkrete Stichprobe wird die Anzahl  $h_j$  festgestellt, dass das Ergebnis  $\{X = x_j\}$  bzw.  $\{x_{j-1}^* < X \leq x_j^*\}$  eingetreten ist. Mit den absoluten Häufigkeiten  $h_j$  für alle  $j = 1, \dots, k$  ist die in der Stichprobe beobachtete Verteilung gegeben. Da die absoluten Häufigkeiten  $h_j$  Ergebnis eines Zufallsexperimentes sind, können sie von Stichprobe zu Stichprobe unterschiedliche Werte annehmen, d.h., sie sind Realisationen von Zufallsvariablen  $H_j$ . Wenn die Nullhypothese gilt, sind die in der Stichprobe erwarteten relativen Häufigkeiten durch die Wahrscheinlichkeiten  $p_j$  gegeben. Für die erwarteten absoluten Häufigkeiten folgt:  $np_j$ .

Die Tatsache, dass die beobachteten absoluten Häufigkeiten Zufallsvariablen  $H_j$  sind, lässt sich wie folgt zeigen, wobei es keine Rolle spielt, ob  $X$  diskret oder stetig ist, so dass nur auf eine diskrete Zufallsvariable  $X$  Bezug genommen wird.

Aus der Grundgesamtheit wird ein Element zufällig gezogen und festgestellt, ob der Wert  $x_j$  aufgetreten ist, d.h., ob das Ereignis  $\{X = x_j\}$  eingetreten ist oder nicht. Es gibt somit nur zwei mögliche Ergebnisse des Zufallsexperimentes. Die Wahrscheinlichkeit für das Eintreten des

### 3. Prüfung der Verteilungform von Variablen

Ereignisses  $\{X = x_j\}$  beträgt bei Gültigkeit der Nullhypothese  $p_j$  und die Wahrscheinlichkeit für das Nichteintreten  $1 - p_j$ . Das Zufallsexperiment wird n-mal wiederholt, wobei die einzelnen Versuche unabhängig voneinander (da eine einfache Zufallsstichprobe vorausgesetzt wird) und die Wahrscheinlichkeiten konstant sind. Es liegt somit ein Bernoulli-Experiment vor.

Bei n-maliger Durchführung der Versuche interessiert die Gesamtzahl des Eintretens von  $\{X = x_j\}$ , d.h. die absolute Häufigkeit von  $x_j$  in der Stichprobe. Diese Häufigkeit kann von Stichprobe zu Stichprobe unterschiedlich sein, so dass  $H_j = \{\text{Anzahl des Auftretens von } X = x_j \text{ in einer einfachen Zufallsstichprobe vom Umfang } n\}$  eine diskrete Zufallsvariable ist, die die Werte  $0, \dots, n$  annehmen kann. Die Zufallsvariable  $H_j$  ist binomialverteilt und zwar bei Gültigkeit von  $H_0$  mit den Parametern  $n$  und  $p_j : H_j \sim B(n; p_j)$ . Der Erwartungswert von  $H_j$  ist  $E(H_j) = np_j$  und damit die bei Gültigkeit der  $H_0$  erwartete Häufigkeit des Wertes  $x_j$  in der Stichprobe. Die Variation der absoluten Häufigkeiten für  $x_j$  wird durch die Varianz  $Var(H_j) = np_j(1 - p_j)$  erfaßt. Diese Herleitung gilt für alle  $j = 1, \dots, k$  gleichermaßen.

Für die Konstruktion der Teststatistik wird die Abweichung der Zufallsvariablen von ihrem Erwartungswert gebildet:  $H_j - np_j$ . Große Differenzen sprechen tendenziell gegen die Nullhypothese und deuten auf eine falsche Verteilungsannahme hin.

Zur Vermeidung, dass sich positive und negative Abweichungen aufheben, erfolgt eine Quadratur:  $(H_j - np_j)^2$ . Die quadrierte Abweichung wird durch die unter  $H_0$  erwartete Häufigkeit  $np_j$  dividiert, um den Einfluß des Stichprobenumfanges  $n$  und der Wahrscheinlichkeit  $p_j$  zu berücksichtigen und um der unterschiedlichen Bedeutung der Abweichungen Rechnung zu tragen. Eine Differenz  $h_j - np_j = 5$  fällt bei  $np_j = 10$  stärker ins Gewicht als bei  $np_j = 100$ .

Durch die Summation der normierten Abweichungen über alle  $j$  ergibt sich eine summarische Größe für die in der gesamten Stichprobe enthaltene Abweichung der beobachteten von den erwarteten Häufigkeiten. Die adäquate Teststatistik lautet somit:

$$V = \sum_{j=1}^k \frac{(H_j - np_j)^2}{np_j}. \quad (3.21)$$

Da die  $H_j$  Zufallsvariablen sind, ist auch  $V$  eine Zufallsvariable. Bei Gültigkeit der Nullhypothese, hinreichend großem Stichprobenumfang  $n$  und Einhaltung der Approximationsbedingungen ist die Teststatistik  $V$  approximativ chi - quadrat - verteilt mit der Anzahl der Freiheitsgrade  $f = k - m - 1$ . Dies gilt unabhängig davon, welche Verteilung unter  $H_0$  angenommen wurde. Die Approximation an die Chi - Quadrat - Verteilung ist hinreichend, wenn nachstehende Approximationsbedingungen erfüllt sind:

- $np_j \geq 1$  für alle  $j$  und
- $np_j \geq 5$  für mindestens 80% der erwarteten absoluten Häufigkeiten.

Sind diese Bedingungen nicht erfüllt, müssen vor der Anwendung des Tests benachbarte Werte bzw. Klassen zusammengefaßt werden. Da die  $p_j$  ( $j = 1, \dots, k$ ) unter  $H_0$  vorgegeben sind, folgt außerdem aus den Approximationsvoraussetzungen, dass die Approximation um so besser ist, je größer der Stichprobenumfang  $n$  ist.

Bei der Bestimmung der Anzahl der Freiheitsgrade ist zu berücksichtigen, dass

- $k$  die Anzahl der verbliebenen Werte bzw. Klassen nach einer eventuell notwendigen Zusammenfassung ist.
- ein Freiheitsgrad grundsätzlich verloren geht, weil die beobachteten absoluten Häufigkeiten nicht unabhängig voneinander sind. Für vorgegebenen Stichprobenumfang  $n$  und aufgrund der Bedingung  $\sum_j h_j = n$  folgt, dass jede Häufigkeit  $h_j$  durch die anderen  $k - 1$  Häufigkeiten bestimmt ist.
- weitere Freiheitsgrade verlorengehen, wenn die hypothetische Verteilung  $F_0(x)$  nicht mit all ihren Parametern bekannt ist, sondern diese Parameter aus der Stichprobe geschätzt werden müssen.  $m$  bezeichnet diese Anzahl der zu schätzenden Parameter der hypothetischen Verteilung (wenn unter  $H_0$  eine vollständig spezifizierte Verteilung vorgegeben wurde, ist  $m = 0$ ).

Da in der Teststatistik die Terme  $(H_j - np_j)^2 / np_j$  nur positive Werte annehmen können, nimmt die Teststatistik  $V$  ebenfalls nur positive Werte an. Große Abweichungen  $H_j - np_j$  zwischen beobachteter und erwarteter Verteilung führen zu großen Werten von  $V$ . Somit führen nur große Werte von  $V$  zur Ablehnung der  $H_0$ , während kleine Werte von  $V$  nicht gegen die Nullhypothese sprechen, sondern auf eine gute Übereinstimmung hindeuten. Der Chi - Quadrat - Anpassungstest ist somit ein rechtsseitiger Test. Der kritische Wert  $\chi^2_{1-\alpha;f}$  wird für  $P(V \leq \chi^2_{1-\alpha;f}) = 1 - \alpha$  und die Anzahl der Freiheitsgrade  $f$  aus der Tabelle der Verteilungsfunktion der Chi - Quadrat - Verteilung entnommen (siehe Anhang C). Die Entscheidungsbereiche sind damit:

Ablehnungsbereich der  $H_0$ :

$$\{v | v > \chi^2_{1-\alpha;f}\}$$

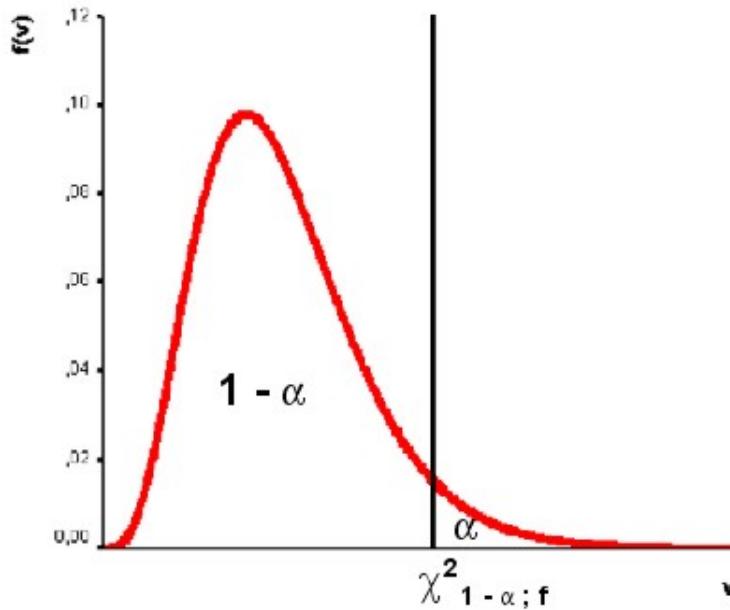
Nichtablehnungsbereich der  $H_0$ :

$$\{v | v \leq \chi^2_{1-\alpha;f}\}.$$

Die Wahrscheinlichkeit, dass die Teststatistik  $V$  eine Realisation aus dem Ablehnungsbereich der Nullhypothese  $H_0$  annimmt, entspricht dem vorgegebenen Signifikanzniveau  $\alpha = P(V > \chi^2_{1-\alpha;f} | H_0)$ . Die Wahrscheinlichkeit, dass die Teststatistik  $V$  eine Realisation aus dem Nichtablehnungsbereich der  $H_0$  annimmt, ist  $P(V \leq \chi^2_{1-\alpha;f} | H_0) = 1 - \alpha$ .

### 3. Prüfung der Verteilungform von Variablen

Abbildung 3.29.: Nichtablehnungsbereich und Ablehnungsbereich der  $H_0$  des Chi-Quadrat-Anpassungstests



Nichtabl.-bereich der  $H_0$  | Ablehnungsbereich der  $H_0$

Wenn die Zufallsstichprobe vom Umfang  $n$  gezogen wurde, können die absoluten Häufigkeiten  $h_j$  ermittelt, gegebenenfalls unbekannte Parameter der hypothetischen Verteilung geschätzt und die erwarteten Häufigkeiten  $np_j$  berechnet werden. Einsetzen in die Teststatistik führt zu einem Prüfwert  $v$ . Wenn  $v$  in den Ablehnungsbereich der  $H_0$  fällt, wird die Nullhypothese auf dem Signifikanzniveau  $\alpha$  und basierend auf der Zufallsstichprobe vom Umfang  $n$  abgelehnt. Es konnte statistisch gezeigt werden, dass die Verteilung der Zufallsvariablen  $X$  in der Grundgesamtheit nicht der hypothetischen Verteilung  $F_0(x)$  entspricht. Bei dieser Entscheidung besteht die Möglichkeit, einen Fehler 1. Art zu begehen, wenn in Wirklichkeit die Nullhypothese richtig ist. Die Wahrscheinlichkeit für einen Fehler 1. Art entspricht dem vorgegebenen Signifikanzniveau  $\alpha$ .

Wenn  $v$  in den Nichtablehnungsbereich der  $H_0$  fällt, wird die Nullhypothese basierend auf der Zufallsstichprobe vom Umfang  $n$  nicht abgelehnt. Es konnte statistisch nicht gezeigt werden, dass die wahre Verteilung in der Grundgesamtheit von der hypothetischen Verteilung  $F_0(x)$  abweicht. Das bedeutet jedoch nicht, dass die wahre Verteilung tatsächlich die hypothetische Verteilung  $F_0(x)$  ist. Das Stichprobenergebnis gibt nur keine Veranlassung,  $H_0$  zu verwerfen. Bei dieser Entscheidung besteht die Möglichkeit einen Fehler 2. Art zu begehen, wenn in Wirklichkeit die Alternativhypothese richtig ist.

Unter SPSS kann der  $\chi^2$ -Anpassungstest über

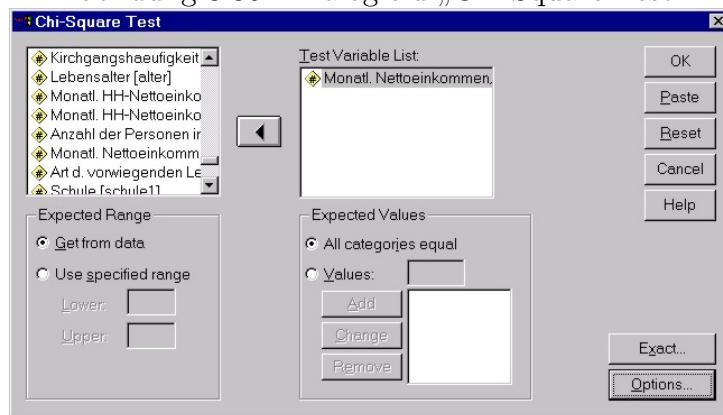
■ Analyze

■ Nonparametric Tests

■ Chi-Square...

aufgerufen werden. In dem erscheinenden Dialogfeld „Chi-Square Test“ erscheinen in der linken Variablenliste nur numerische Variablen. Soll die Häufigkeitsverteilung einer nominalskalierten Variablen geprüft werden, müssen die Variablenausprägungen vorher zahlenmäßig kodiert werden. Die zu testende(n) Variable(n) ist (sind) in das Feld „Test Variable List:“ zu bringen.

Abbildung 3.30.: Dialogfeld „Chi-Square Test“



Voreingestellt ist in dem Feld „Expected Values“ die Prüfung gegen eine Gleichverteilung (All categories equal). Soll nicht auf eine Gleichverteilung getestet werden, muss unter „Values:“ für jeden Beobachtungswert bzw. für jede Klasse die unter  $H_0$  erwartete Häufigkeit eingegeben werden. Dabei ist zu beachten, dass

- die Eingabewerte größer als 0 sein müssen;
- sie in aufsteigender Reihenfolge in Analogie zu den Beobachtungswerten bzw. Klassen einzugeben sind;
- nach der Eingabe eines Wertes auf Add (Hinzufügen) zu klicken ist;
- diese Eingabewerte als Anteile (nicht als absolute Werte) interpretiert werden, d.h., diese Werte werden aufsummiert, jeder Wert wird durch die Summe dividiert und dann in die erwarteten absoluten relativen Häufigkeiten umgewandelt. Es können auch Wahrscheinlichkeiten oder Prozentsätze eingegeben werden. Wenn z.B. die Werte einer Variablen in 4 Klassen eingeteilt wurden und im eben beschriebenen Sinne die Werte 15, 30, 25 und 10 eingegeben wurden, so ist deren Summe 80. Die erwarteten (relativen) Häufigkeiten sind somit 15/80, 30/80, 25/80 und 10/80.

### 3. Prüfung der Verteilungform von Variablen

Voreingestellt ist außerdem in dem Feld „Expected Range“, dass alle Variablenwerte (Klassen, Kategorien) in den Test einbezogen werden (Get from data). Man kann den Test jedoch auch auf eine Teilmenge anwenden, indem auf „Use specified range“ entschieden und eine Unter-(Lower) und Obergrenze (Upper) festgelegt wird.

Über die Schaltfläche „Options...“ gelangt man in ein weiteres Dialogfeld, in dem zusätzlich Maßzahlen der deskriptiven Statistik und die Quartile angefordert werden können.

Über die Schaltfläche „Exact...“ gelangt man in das Dialogfeld „Exact Tests“, wo wie beim Kolmogorov-Smirnov-Test die Möglichkeit besteht, exakte Ergebnisse bei der Berechnung des Signifikanzniveaus (Überschreitungswahrscheinlichkeit) zu erhalten, wenn die Voraussetzungen der asymptotischen Methode durch die Daten nicht erfüllt sind.

Der Test-Output enthält die Klassen bzw. Kategorien, die beobachteten (Observed N) und die unter  $H_0$  erwarteten absoluten Häufigkeiten (Expected N), die Differenz zwischen beiden (Residual), den  $\chi^2$ -Wert (Chi-Square), die Anzahl der Freiheitsgrade (df) und die Überschreitungswahrscheinlichkeit des Testwertes (Asymp. Sig.). Letztere ist mit dem vorgegebenen Signifikanzniveau  $\alpha$  zu vergleichen. Unterhalb der Tabelle „Test Statistics“ wird eine Aussage zu den Approximationsbedingungen getroffen, die auf jeden Fall betrachtet werden sollte.

- Beispiel 3.2 (Fortsetzung):

Das monatliche persönliche Nettoeinkommen der Datei allbus.sav liegt in der Variablen einkomp2 klassiert in 22 Klassen vor. Vor der Testdurchführung werden jedoch die Klassen 15 bis 22 zusammengefaßt, so dass diese letzte Klasse inhaltlich 4000 DM oder mehr bedeutet. Dazu wird, wie im Kapitel 1 beschrieben, eine Umkodierung (siehe Abb. 1.3 und 1.4) in eine andere Variable (hier: einkp2) in der Weise vorgenommen, dass der alte Wertebereich 16-22 auf den neuen Wert 15 gesetzt wird, die Werte 1-15 übernommen und alle anderen Werte als System-Missings vereinbart werden.

Auf einem Signifikanzniveau von 5% soll mit dem  $\chi^2$ -Anpassungstest getestet werden, ob diese Variable in der Grundgesamtheit einer Gleichverteilung entspricht.

Das Ergebnis ist im SPSS-Output 3.2-2 enthalten, wobei die durch SPSS zu sehr gerundeten Angaben in der Häufigkeitstabelle (nur 1 Dezimalstelle) auf 2 Dezimalstellen geändert wurden.

**SPSS-Output 3.2-2:** Ergebnis des Chi-Quadrat-Anpassungstests für die Variable einkp2  
Frequencies

Persönliches Einkommen

	Observed N	Expected N	Residual
unter 400	21	29,67	-8,67
400 - 599	20	29,67	-9,67
600 - 799	20	29,67	-9,67
800 - 999	31	29,67	1,33
1000 - 1249	30	29,67	,33
1250 - 1499	26	29,67	-3,67
1500 - 1749	34	29,67	4,33
1750 - 1999	64	29,67	34,33
2000 - 2249	40	29,67	10,33
2250 - 2499	21	29,67	-8,67
2500 - 2749	31	29,67	1,33
2750 - 2999	27	29,67	-2,67
3000 - 3499	29	29,67	-,67
3500 - 3999	23	29,67	-6,67
4000 und mehr	28	29,67	-1,67
Total	445		

Test Statistics

	Persönliches Nettoeinkommen
Chi-Square <sup>a</sup>	57,753
df	14
Asymp. Sig.	,000

a. 0 cells (,0%) have expected frequencies less than 5.

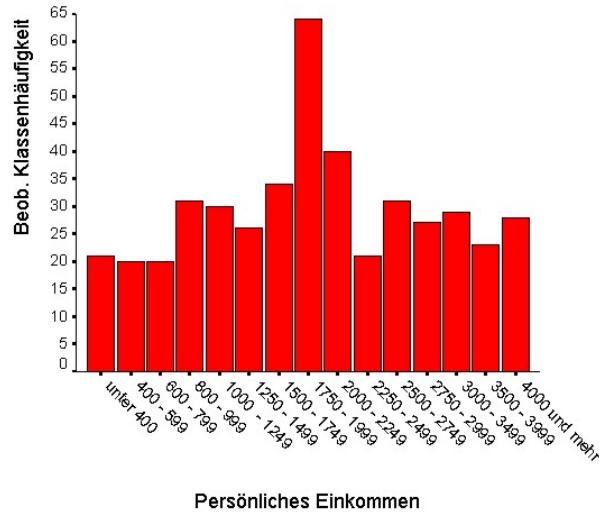
The minimum expected cell frequency is 29,7.

Die Fußnote zur Tabelle „Test Statistics“ signalisiert, dass die Approximationsbedingungen für den Chi-Quadrat-Anpassungstest eingehalten wurden. Die Anzahl der Freiheitsgrade ist hier  $k - 1 = 15 - 1 = 14$ , da für die Gleichverteilung keine Parameter aus der Stichprobe geschätzt werden mußten ( $m = 0$ ).

Da  $Sig < \alpha$ , wird die  $H_0$  abgelehnt; die Variable stammt nicht aus einer gleichverteilten Grundgesamtheit. Zum gleichen Ergebnis gelangt man natürlich, wenn man den berechneten  $\chi^2$ -Wert mit dem kritischen Wert vergleicht, den man für  $1 - \alpha = 0,95$  und die Anzahl der Freiheitsgrade  $f = 14$  aus Anhang C mit  $\chi^2_{14,0,95} = 23,685$  erhält. Da  $v = 57,753$  Element des Ablehnungsbereiches  $\{v | v > 23,685\}$  ist, wird  $H_0$  basierend auf der Stichprobe vom Umfang  $n = 445$  auf dem Signifikanzniveau  $\alpha = 0,05$  verworfen. Betrachtet man die ausgegebene empirische Häufigkeitsverteilung kritisch, so fällt auf, dass vor allem die 8. Einkommensklasse (1750 - 1999 DM) aus dem Rahmen herausfällt und eine sehr große Abweichung (34,33) zur erwarteten Häufigkeit aufweist. Das zeigt auch ein für die Variable einkp2 erzeugtes Balkendiagramm.

### 3. Prüfung der Verteilungform von Variablen

Abbildung 3.31.: Balkendiagramm der Variablen persönliches Nettoeinkommen (klassiert)



Aus diesem Grunde werden nur der darunterliegende bzw. der darüberliegende Bereich auf Gleichverteilung geprüft, indem in Feld „Expected Range“ des Dialogfeldes „Chi-Square Test“ (Abb. 3.30) benutzerdefiniert für Lower 1 und für Upper 7 und bei der nächsten Testdurchführung für Lower 9 und für Upper 15 eingegeben wird. Im erstenen Fall betrifft dies  $n_1 = 182$  Fälle und im 2. Fall  $n_2 = 199$  Fälle. Die Ergebnisse der Tests sind im SPSS-Output 3.2-3 enthalten.

In beiden Fällen wird die  $H_0$  auf Gleichverteilung wegen  $Sig > \alpha$  nicht abgelehnt, so dass offensichtlich die beobachtete Häufigkeit der 8. Einkommensklasse entscheidend für die Verwerfung von  $H_0$  beim Test unter Einbeziehung aller Klassen gewesen ist.

**SPSS-Output 3.2-3:** Ergebnis des Chi-Quadrat-Anpassungstests für die Variable einkp2 für Klasse 1-7 und 9-15

#### Frequencies

#### Persönliches Einkommen

	Observed N	Expected N	Residual
unter 400	21	26,00	-5,0
400 - 599	20	26,00	-6,0
600 - 799	20	26,00	-6,0
800 - 999	31	26,00	5,0
1000 - 1249	30	26,00	4,0
1250 - 1499	26	26,00	,0
1500 - 1749	34	26,00	8,0
Total	182		

#### Test Statistics

	Persönliches Nettoeinkommen
Chi-Square <sup>a</sup>	7,769
df	6
Asymp. Sig.	,256

a. 0 cells (.0%) have expected frequencies less than 5.  
The minimum expected cell frequency is 26,0.

**Frequencies****Persönliches Einkommen**

	Observed N	Expected N	Residual
2000 - 2249	40	28,4	11,6
2250 - 2499	21	28,4	-7,4
2500 - 2749	31	28,4	2,6
2750 - 2999	27	28,4	-1,4
3000 - 3499	29	28,4	,6
3500 - 3999	23	28,4	-5,4
4000 und mehr	28	28,4	-,4
Total	199		

**Test Statistics**

	Persönliches Nettoeinkommen
Chi-Square <sup>a</sup>	8,010
df	6
Asymp. Sig.	,237

a. 0 cells (,0%) have expected frequencies less than 5.  
The minimum expected cell frequency is 28,4.

- Beispiel 3.8:

Aus der gleichen Datei (allbus.sav) soll für die Variable Geschlecht (sex), die kodiert mit 1 - Mann und 2 - Frau vorliegt, so dass sie unter SPSS als numerische Variable behandelt werden kann, auf einem Signifikanzniveau von  $\alpha = 0,1$  geprüft werden, ob unter den gültigen Fällen von  $n = 3052$  ein Geschlechtsverhältnis von 45 : 55 vorliegt. Dazu wird im Feld „Expected Values“ im Dialogfeld „Chi-Square Test“ (siehe Abb. 3.30) Values angeklickt und 45 und 55 eingegeben.

**SPSS-Output 3.8-1:** Ergebnis des Chi-Quadrat-Anpassungstests für die Variable

sex

**Frequencies****Geschlecht**

	Observed N	Expected N	Residual
Mann	1356	1373,4	-17,4
Frau	1696	1676,6	17,4
Total	3052		

**Test Statistics**

	Geschlecht
Chi-Square <sup>a</sup>	,401
df	1
Asymp. Sig.	,527

a. 0 cells (,0%) have expected frequencies less than 5.  
The minimum expected cell frequency is 1373,4.

Die  $H_0$ : „Es liegt eine Verteilung männlich zu weiblich von 45:55 vor“ kann nicht abgelehnt werden.

### 3. Prüfung der Verteilungform von Variablen

#### 3.3.3. Binomial-Test

Der Binomial-Test<sup>22</sup> setzt eine dichotome bzw. dichotomisierte Variablen X voraus, wobei jedes Skalenniveau zulässig ist, und basiert auf der Binomialverteilung<sup>23</sup>. Im weiteren wird vereinbart, dass für die interessierende Eigenschaft die Variable X den Wert 1 annimmt. Die Wahrscheinlichkeit dafür, dass in der Grundgesamtheit ein Element zur Kategorie 1 gehört, ist gleich p und, dass ein Element zur anderen Kategorie gehört, gleich 1 - p. Die Wahrscheinlichkeit p der Grundgesamtheit ist unbekannt. Es existiert jedoch eine Annahme, dass die Wahrscheinlichkeit  $p_0$  ist.

Aus der Grundgesamtheit wird eine einfache Zufallsstichprobe  $X_1, \dots, X_n$  vom Umfang n gezogen. Geprüft wird das Hypothesenpaar

- beim zweiseitigen Test:

$$H_0 : p = p_0 \quad H_1 : p \neq p_0,$$

d.h., die Variable stammt aus einer Grundgesamtheit mit genau dem Anteilswert  $p_0$ ;

- beim einseitigen Test:

$$H_0 : p \leq p_0 \quad H_1 : p > p_0,$$

d.h., die Variable stammt aus einer Grundgesamtheit mit einem Anteilswert von höchstens  $p_0$ ; oder

$$H_0 : p \geq p_0 \quad H_1 : p < p_0,$$

d.h., die Variable stammt aus einer Grundgesamtheit mit einem Anteilswert von mindestens  $p_0$ .

Die Teststatistik ist

$$V = \sum_{i=1}^n X_i, \tag{3.22}$$

d.h. die Anzahl der Fälle der Kategorie 1 in einer Stichprobe vom Umfang n. Die Teststatistik V ist unter  $H_0$   $B(n; p_0)$ -verteilt.

Für „hinreichend“ große Stichproben kann die Verteilung der Teststatistik V für beliebiges  $p_0$  durch die Normalverteilung approximiert werden, wobei die Approximation um so besser ist, je mehr sich  $p_0$  dem Wert 0,5 annähert. Unter  $H_0$  ist  $E(X) = np_0$  und  $Var(x) = np_0(1 - p_0)$ .

Die Teststatistik

$$Z = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}} \tag{3.23}$$

---

<sup>22</sup>Vgl. u.a. Büning, H., Trenkler, G. (1978), S. 103 ff.

<sup>23</sup>Vgl. u.a. Rönnz, B., Strohe, H.G. (Hrsg.) (1994), S. 50 ff.

bzw. unter Berücksichtigung der Stetigkeitskorrektur von Yates für  $20 \leq n \leq 60$

$$Z = \frac{X - np_0 - 0,5}{\sqrt{np_0(1 - p_0)}} \quad (3.24)$$

folgt unter  $H_0$  dann asymptotisch einer Standardnormalverteilung.

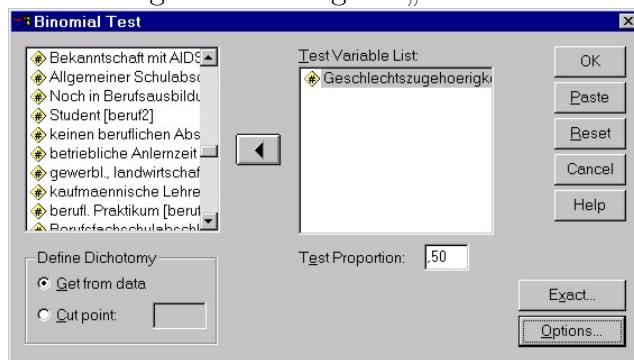
Der Binomial-Test ist unter SPSS ebenfalls als ein nichtparametrischer Test abrufbar:

■ Analyze

■ Nonparametric Tests

■ Binomial...

Abbildung 3.32.: Dialogfeld „Binomial Test“



Im Dialogfeld „Binomial Test“ wird die zu testende Variable in das Feld „Test Variable List:“ gebracht. Im Feld „Define Dichotomy“ bedeutet „Get from data“, dass bereits eine dichotome Variable vorliegt. Ist das nicht der Fall, kann sie durch Eingabe des Trennwertes (Cut point) dichotomisiert werden. Alle Fälle mit Werten, die kleiner oder gleich dem Trennwert sind, gehören dann zu einer Kategorie und Fälle mit Werten, die größer als der Trennwert sind, zu der anderen Kategorie.

Das Feld „Test Proportion:“ nimmt die hypothetische Wahrscheinlichkeit  $p_0$  auf. Sie ist mit  $p_0 = 0,50$  voreingestellt. Ist  $p_0$  unter  $H_0$  ein anderer Wert, so muss er in diesem Feld geändert werden.

Über die Schaltfläche „Options...“ gelangt man in ein weiteres Dialogfeld, in dem zusätzlich Maßzahlen der deskriptiven Statistik und die Quartile angefordert werden können. Über die Schaltfläche „Exact...“ gelangt man in das Dialogfeld „Exact Tests“, wo wie bei den beiden vorherigen Tests die Möglichkeit besteht, exakte Ergebnisse bei der Berechnung des Signifikanzniveaus (Überschreitungswahrscheinlichkeit) zu erhalten, wenn die Voraussetzungen der asymptotischen Methode durch die Daten nicht erfüllt sind.

Der Output umfasst die Anzahl der gültigen Fälle (N) jeder Kategorie und insgesamt, den beobachteten Anteil für jede Kategorie (Observed Prop.), die Wahrscheinlichkeit  $p_0$  (Test Prop.)

### 3. Prüfung der Verteilungform von Variablen

und das Signifikanzniveau (Asymp.Sig. (2-tailed)). Das Signifikanzniveau ergibt sich in unterschiedlicher Weise.

- a) Wurde das voreingestellte  $p_0 = ,50$  belassen, wird ein zweiseitiges, andernfalls ein einseitiges Signifikanzniveau ausgegeben.

Beim einseitigen Signifikanzniveau erfolgt der Test stets in die beobachtete Richtung, d.h., wenn der beobachtete Anteil größer als  $p_0$  ist, bezieht sich das ausgegebene Signifikanzniveau darauf, dass mehr in dieser Kategorie beobachtet werden; wenn der beobachtete Anteil kleiner als  $p_0$  ist, bezieht sich das ausgegebene Signifikanzniveau darauf, dass weniger in dieser Kategorie beobachtet werden.

- b) Für große Stichproben (in den SPSS-Handbüchern wird keine Angabe gemacht, ab welchem n) erfolgt die Approximation durch die Normalverteilung, wie oben beschrieben.

Der Stichprobenumfang n ist unter SPSS durch die Anzahl der gültigen Fälle der Variablen X in der verwendeten Datei gegeben.

#### • Beispiel 3.9:

Mit diesem Beispiel sollen ausschließlich verschiedene Ergebnisse des Binomialtests demonstriert werden.

Für die Variable Bevölkerungsdichte (bev\_di) der Datei beisp1.sav wird eine Dichotomisierung in der Weise durchgeführt, dass als Trennwert (Cut point) 650 eingegeben wird, d.h., Bundesländer mit einer Bevölkerungsdichte kleiner oder gleich 650 bilden eine Kategorie und Bundesländer mit einer Bevölkerungsdichte von mehr als 650 bilden die andere Kategorie. Es soll auf einem Signifikanzniveau von  $\alpha = 0,05$  zunächst getestet werden, ob die dichotomisierte Variable aus einer Grundgesamtheit mit dem Anteil  $p_0 = 0,5$  stammt, d.h., der voreingestellte Anteil (Test Proportion) wird beibehalten.

**SPSS-Output 3.9-1:** Ergebnisse des Binomialtests für die Variable Bevölkerungsdichte mit  $p_0 = 0,5$

Biomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Bevölkerungsdichte	Group 1	$\leq 650$	13	,81	,50	,021
	Group 2	$> 650$	3	,19		
	Total		16	1,00		

Da n klein ist, wird die Teststatistik V verwendet wird, die unter  $H_0 : p = p_0 (= 0,5)$  einer Binomialverteilung  $B(16;0,5)$  folgt, was im Output durch Exact Sig in der letzten Spalte gekennzeichnet wird. Wegen  $p_0 = 0,5$  beinhaltet die Ausgabe das zweiseitige Signifikanzniveau (2-tailed). Unter Verwendung von Tafeln der Verteilungsfunktion der Binomialverteilung

$B(16;0,5)$  findet man

$$P(V \leq 3) = 0,0106 \text{ und } P(V \geq 13) = 1 - P(V \leq 12) = 0,0106,$$

so dass die zweiseitige Überschreitungswahrscheinlichkeit

$$P(V \leq 3) + P(V \geq 13) = 0,0212$$

ist. Wegen  $Sig < \alpha$  wird  $H_0$  aufgrund der Stichprobe auf einem Signifikanzniveau von 5 % abgelehnt.

Prüft man die in o.a. Weise dichotomisierte Variable Bevölkerungsdichte auf dem gleichen Signifikanzniveau, ob sie aus einer Grundgesamtheit mit dem Anteil  $p_0 = 0,6$  stammt, so ist dieser Wert in das Feld Test Proportion einzutragen.

**SPSS-Output 3.9-2:** Ergebnisse des Binomialtests für die Variable Bevölkerungsdichte mit  $p_0 = 0,6$

**Biomial Test**

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
Bevölkerungsdichte	Group 1	<= 650	13	,8125	,60	,065
	Group 2	> 650	3	,2		
	Total		16	1,00		

Hier ist der beobachtete Anteil (Observed Prop.) für Group 1 größer als  $p_0$  (Test Prop), so dass

- die Hypothesen  $H_0 : p \leq p_0 = 0,6$  und  $H_1 : p > p_0 = 0,6$  zugrunde liegen,
- ein einseitiges Signifikanzniveau (1-tailed) ausgegeben wird, das aufgrund der o.g. Regeln  $Sig = P(V \geq v) = P(V \geq 13) = 0,065$  beinhaltet.

Da  $Sig > \alpha$  ist, wird  $H_0$  nicht abgelehnt.

Wird z.B.  $p_0 = 0,9$  angenommen, so erhält man den nachstehenden Output.

**SPSS-Output 3.9-3:** Ergebnisse des Binomialtests für die Variable Bevölkerungsdichte mit  $p_0 = 0,9$

**Biomial Test**

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
Bevölkerungsdichte	Group 1	<= 650	13	,8125	,9	,211 <sup>a</sup>
	Group 2	> 650	3	,2		
	Total		16	1,00		

<sup>a</sup>. Alternative hypothesis states that the proportion of cases in the first group < ,9.

Nunmehr ist der beobachtete Anteil (Observed Prop.) für Group 1 kleiner als  $p_0$  (Test Prop), so dass

- die Hypothesen  $H_0 : p \geq p_0 = 0,9$  und  $H_1 : p < p_0 = 0,9$  zugrunde liegen,
- ein einseitiges Signifikanzniveau (1-tailed) ausgegeben wird, das aufgrund der o.g. Regeln  $Sig = P(V \leq v) = P(V \leq 13) = 0,2108$  beinhaltet.

### 3. Prüfung der Verteilungform von Variablen

Da  $Sig > \alpha$  ist, wird  $H_0$  nicht abgelehnt.

- Beispiel 3.8 (Fortsetzung):

In der Datei allbus.sav ist die Variable Geschlecht (sex) mit 1 = „Mann“ und 2 = „Frau“ kodiert, so dass bereits eine dichotome Variable vorliegt. Es soll auf einem Signifikanzniveau von  $\alpha = 0,05$  geprüft werden, ob die Stichprobe aus einer Grundgesamtheit mit dem Anteil  $p_0 = 0,5$  stammt. Im Dialogfeld „Binomial Test“ (Abb. 3.32) braucht somit nur die Variable sex in das Feld „Test Variable List:“ gebracht werden.

**SPSS-Output 3.8-2:** Ergebnis des Binomialtests für die Variable sex mit  $p_0 = 0,5$

**Binomial Test**

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Geschlechtszugehörigkeit	Group 1	Frau	1696	,56	,50	,000 <sup>a</sup>
	Group 2	Mann	1356	,44		
	Total		3052	1,00		

<sup>a</sup>. Based on Z Approximation.

Der erste Wert in der Datei für diese Variable ist 2, so dass die Kategorie „Frau“ als erstes unter Category erscheint und der Anteil dieser Kategorie geprüft wird. Der beobachtete Anteil beträgt  $1696/3052 = 0,5557$ . Da  $n = 3052$  sehr groß ist, erfolgt die Testdurchführung mit der Teststatistik Z (Z Approximation) und wegen  $p_0 = 0,5$  wird ein zweiseitiges Signifikanzniveau ausgegeben.

Die Testentscheidung soll für dieses Beispiel ausführlich gezeigt werden. Nach (3.23) ergibt sich  $z = (1696 - 3052 \cdot 0,5)(3052 \cdot 0,5 \cdot 0,5)^{-0,5} = 6,15$ . Aus der Tabelle der Standardnormalverteilung findet man für  $1 - \alpha/2 = 0,975$  den Wert  $z_{0,975} = 1,96$ , womit der Nichtablehnungsbereich der  $H_0$  lautet:  $\{z | -1,96 \leq z \leq +1,96\}$  und der Ablehnungsbereich von  $H_0$ :  $\{z | z < -1,96 \text{ oder } z > 1,96\}$ . Da  $z = 6,15$  in den Ablehnungsbereich fällt, wird die Nullhypothese  $H_0$ : „Die Stichprobe stammt aus einer Grundgesamtheit mit einem Frauenanteil von 0,5“ auf einem Signifikanzniveau von 5 % abgelehnt.

Unter SPSS beinhaltet das zweiseitige Signifikanzniveau:

$$Sig = P(\{Z < -6,15\} \cup \{Z > 6,15\}) = 0.$$

Da  $Sig < \alpha$  ist, wird  $H_0$  abgelehnt.

Die Grenzen für den Annahme- und den Ablehnungsbereich der  $H_0$  können auch für die Variable V berechnet werden. Bezeichnet man die Grenzen mit  $v_u$  und  $v_o$ , so ergibt sich:

$$v_u = np_0 + z_{\alpha/2} \cdot \sqrt{np_0(1 - p_0)} \quad (3.25)$$

$$v_o = np_0 + z_{1-\alpha/2} \cdot \sqrt{np_0(1 - p_0)}$$

Da das Ergebnis für  $v_u$  abgerundet und für  $v_o$  aufgerundet wird, erhält man für diesen konkreten Fall  $v_u = 1471$  und  $v_o = 1581$ , so dass sich für den Ablehnungsbereich der  $H_0$ , bezogen auf die Variable V, ergibt:  $\{v | v < 1471 \text{ oder } v > 1581\}$ . Da v = 1696 Element des Ablehnungsbereichs ist, ist die Testentscheidung natürlich wie vorher. Testet man dagegen auf einem Signifikanzniveau von  $\alpha = 0,10$  beispielsweise, ob die Stichprobe aus einer Grundgesamtheit mit einem Frauenanteil von  $p_0 = 0,545$  (oder kleiner) stammt, so liegen die Hypothesen  $H_0 : p \leq p_0 = 0,545$  und  $H_1 : p > p_0 = 0,545$  zugrunde.

**SPSS-Output 3.8-3:** Ergebnis des Binomialtests für die Variable sex mit  $p_0 = 0,545$

#### Biomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Geschlechtszugehörigkeit	Group 1	Frau	1696	,555701	,545	,121 <sup>a</sup>
	Group 2	Mann	1356	,444		
	Total		3052	1,00		

<sup>a</sup>. Based on Z Approximation.

Hier wird nun wieder ein einseitiges Signifikanzniveau ausgegeben, da  $p \neq 0,5$  ist. Da der beobachtete Anteil (Observed Prop.) größer als  $p_0$  ist, beinhaltet das einseitige Signifikanzniveau  $Sig = P(V \geq v) = P(V \geq 1696) = 0,1212$ . Diese Wahrscheinlichkeit ergibt sich wie folgt:

$$\begin{aligned} P(V \geq 1696) &= P\left(Z \geq \frac{1696 - 3052 \cdot 0,545 - 0,5}{\sqrt{3052 \cdot 0,545 \cdot 0,455}}\right) = P(Z \geq 1,169) \\ &= 1 - P(Z \leq 1,169) = 1 - 0,8788 = 0,1212, \end{aligned}$$

mit  $P(Z \leq 1,169) = 0,8788$  aus der Tabelle der Verteilungsfunktion der Standardnormalverteilung. Wegen  $Sig > \alpha$  wird  $H_0$  nicht abgelehnt.

Als Grenze zwischen Nichtablehnungs- und Ablehnungsbereich der  $H_0$  bezogen auf die Teststatistik V folgt:

$$v_{krit} = np_0 + z_{0,9} \cdot \sqrt{np_0(1-p_0)} = 3052 \cdot 0,545 + 1,282 \cdot 27,51 = 1698,61$$

mit  $z_{0,9} = 1,282$  aus der Tabelle der Verteilungsfunktion der Standardnormalverteilung. Da das Ergebnis für  $v_{krit}$  aufgerundet wird, erhält man für den Nichtablehnungsbereich der  $H_0 : \{v | v \leq 1699\}$ . Da v = 1696 Element des Nichtablehnungsbereiches ist, ist die Testentscheidung natürlich wie vorher.

## 3.4. Transformationen

Eine Variablentransformation<sup>24</sup> ist der Übergang von einer metrisch skalierten Variablen X mittels einer Funktion derselben zu einer neuen Variablen Y. Dabei wird jedem Wert der Va-

<sup>24</sup>Vgl. u.a. Rönz, B., Strohe, H.G. (Hrsg.) (1994), S. 368 f.; Hartung, Elpelt, Klösener (1993), S. 349 ff., 832 ff.; Schlittgen, R. (1990), S. 152 ff.; Heiler, S., Michels, P. (1994), S. 157 ff.; Fox, J., Long, J.S. (1990), S. 105 ff.

### 3. Prüfung der Verteilungform von Variablen

riablen X entsprechend der Transformationsvorschrift T ein Wert der Variablen Y zugeordnet:  $y := T(x)$ . Transformationen dienen dem Ziel, eine neue Variable zu erhalten, die die Voraussetzungen für bestimmte statistische Methoden besser erfüllt als die Ausgangsvariable.

Beispiele:

- ◆ Transformation der Werte einer Variablen mit großem Zahlenbereich zur übersichtlichen Darstellung der Häufigkeitsverteilung;
- ◆ Transformation einer Variablen mit schiefer Verteilung in eine mit (approximativ) symmetrischer Verteilung;
- ◆ Transformation zur Erzielung von Gleichheit der Varianzen, wenn mit dem Niveau der Variablenwerte auch das Streuungsverhalten variiert.

Grundsätzlich ist bei der Anwendung von Transformationen zu beachten, dass

- sie nur sinnvoll sind, wenn die Daten eine genügend große Spannweite aufweisen (Faustregel:  $x_{(n)} / x_{(1)} \geq 20$ ),
- die mit den transformierten Variablen erzielten Ergebnisse nicht unbedingt sachlogisch interpretierbar sind.

Die Auswahl einer geeigneten Transformation ist oftmals ein Versuch-Irrtum-Verfahren. Einen breiten Bereich der für a) bis c) notwendigen Transformationen decken Potenztransformationen

$$T_p(x) = \begin{cases} (x + c)^p & p \neq 0 \\ \ln(x + c) & p = 0 \end{cases} \quad (3.26)$$

ab. Die Konstante c wird z.B. so gewählt, dass alle Werte der transformierten Variablen positiv werden, was eine Voraussetzung der Potenztransformationen ist. Da bei  $p < 0$  sich die Ordnung der Daten umkehrt, wird dann auch  $-(x + c)^p$  verwendet. Die Wirkung dieser Transformation zeigt die „Leiter der Transformation“ (Tabelle 3.4). Die Abbildung 3.33 zeigt die Potenzfunktionen für ausgewählte Werte von p und c = 0.

Im Kontext dieses Kapitels ist vor allem der o.g. zweite Grund maßgebend für die Anwendung einer Transformation. Da das wohl kaum ohne Experimentieren abläuft, sollen einige Hinweise gegeben werden, wann welche Transformation geeignet ist. Dazu muß man sich zunächst folgendes vergegenwärtigen: Die Verwendung einer Transformation mit  $p > 1$  bewirkt, dass große Werte eine größere Ausdehnung erfahren als kleinere, dagegen die Verwendung einer Transformation mit  $p < 1$ , dass kleinere Werte eine größere Ausdehnung erfahren als große Werte.

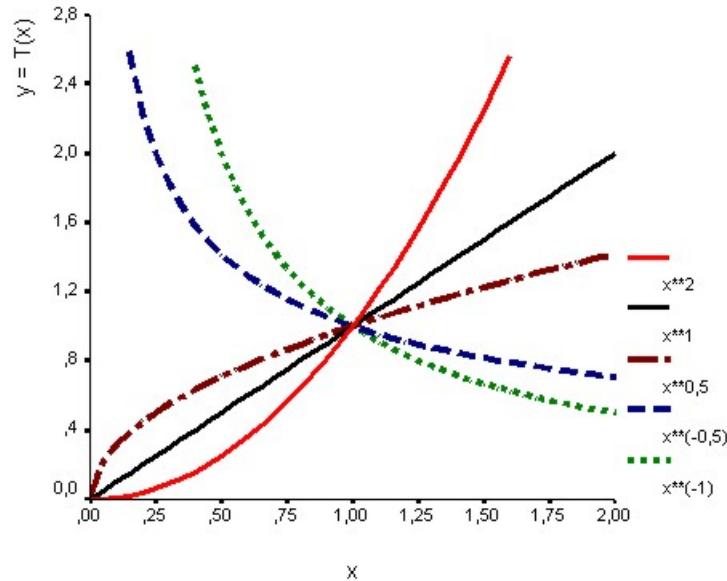
Tabelle 3.4.: Leiter der Transformation

p	transf. Werte $x^p$	Transformation zur Symmetrisierung	Transformation zur Linearisierung
:	:		
:	:	für linksschiefe Verteilungen	wenn überproportional wachsende Änderungen der Y-Werte auftreten
3	$x^3$		
:	:		
2	$x^2$	↑	↑
:	:		
1	$x^1$	ohne Effekt	ohne Effekt
:	:		
0,5	$x^{0,5}$		
:	:		
0	$\ln x; \log x$	↓	↓
:	:		
-0,5	$1/x^{0,5}$	für rechtsschiefe Verteilungen	wenn mit wachsenden X-Werten die Änderungen der Y-Werte schwächer werden
:	:		
-1	$1/x$		
:	:		
-2	$1/x^2$		
:	:		

Quelle: Zusammengestellt nach Schlittgen, R. (1990), S. 155, 427

### 3. Prüfung der Verteilungform von Variablen

Abbildung 3.33.: Potenzfunktionen  $T_p(x)$  für ausgewählte p-Werte



Um eine linksschiefe Verteilung (negative Schiefe) einer symmetrischen Verteilung anzunähern, wird man demzufolge  $p > 1$  wählen, was ein Zusammenziehen des linken Verteilungsschwanzes und eine Ausdehnung des rechten Verteilungsschwanzes bewirkt.

Um eine rechtsschiefe Verteilung (positive Schiefe) einer symmetrischen Verteilung anzunähern, wird man  $p < 1$  wählen, was eine Ausdehnung des linken Verteilungsschwanzes und ein Zusammenziehen des rechten Verteilungsschwanzes zur Folge hat.

Wählt man in (3.26) für  $p > 1$  den Parameter  $c = -m$ , wobei  $m$  ein mittlerer Wert ist, so wird eine Ausdehnung beider Verteilungsschwänze und ein Zusammenziehen des mittleren Bereiches erreicht.

Weitere Transformationen sind:

#### ■ Box-Cox-Transformation

$$T_p^*(x) = \begin{cases} \frac{(x + c)^p - 1}{p} & p \neq 0 \\ \ln(x + c) & p = 0. \end{cases} \quad (3.27)$$

Der Parameter  $p$  dient der Funktionsauswahl und der Parameter  $c$  bewirkt eine Niveauverschiebung, die u.a. wieder notwendig ist, wenn negative Beobachtungswerte auftreten. Die Box-Cox-Transformation enthält einige der Potenztransformationen als Spezialfälle.

■ Arcus-Sinus-Transformation

$$T(x) = \sqrt{n + c_1} \cdot \arcsin \sqrt{\frac{x + c_2}{n + c_3}} \quad (3.28)$$

mit den einstellbaren Konstanten  $c_1$ ,  $c_2$  und  $c_3$ . Sie wird vor allem bei Daten aus einer Binomialverteilung  $B(n;p)$  angewandt, um eine Anpassung an die Normalverteilung zu erreichen.

■ Gefaltete Wurzel- und Log-Transformationen (folded-root, folded-log)

$$\begin{aligned} T_F(x) &= \sqrt{x} - \sqrt{1-x} \\ T_F^*(x) &= \log x - \log(1-x) = \log\left(\frac{x}{1-x}\right) \end{aligned} \quad (3.29)$$

Die gefaltete Log-Transformation wird allgemein auch als Logit-Transformation bezeichnet. Diese Transformationen sind sinnvoll, wenn die Werte der Variable X als Verhältniszahlen, Raten oder Prozentwerte gegeben sind, z.B. wenn X die Wahrscheinlichkeit des Eintretens eines „Erfolges“ beinhaltet. Sie dehnen die Verteilungsschwänze im Verhältnis zum mittleren Verteilungsteil stärker aus. Gegebenenfalls muß eine kleine Konstante vorher zu den x-Werten addiert werden, um die Werte Null und Eins zu vermeiden.

■ Angepaßte Transformation (matched transformations)

Nach der Potenztransformation gemäß (3.26) mit  $c = 0$  fallen die transformierten Werte oftmals aus dem Bereich der Ursprungswerte heraus. Es wird deshalb eine weitere, lineare Transformation der Form

$$z = a \cdot T_p(x) + b$$

durchgeführt. Die Parameter a und b werden dabei so gewählt, dass  $z_0 = x_0$  und  $dz/dx = 1$  für  $x = x_0$ .  $x_0$  ist dabei ein mittlerer Punkt, im allgemeinen der Median.

Wählt man speziell:

$$b = x_0 - a \cdot T_p(x_0) \text{ und } a = 1/T_p^*(x_0),$$

worin  $T_p^*(x_0)$  die erste Ableitung der Transformation  $T_p(x_0)$  ist, so resultiert

$$z = \begin{cases} x_0 + \frac{x^p - x_0^p}{p \cdot x_0^{p-1}} & p \neq 0 \\ x_0 \left[ 1 + \ln \left( \frac{x}{x_0} \right) \right] & p = 0. \end{cases} \quad (3.30)$$

Dies läßt sich leicht wie folgt zeigen. Einsetzen der gewählten Parameter führt zu

$$z = aT_p(x) + b$$

### 3. Prüfung der Verteilungsform von Variablen

$$= \frac{T_p(x)}{T_p^*(x_0)} + x_0 - \frac{T_p(x_0)}{T_p^*(x_0)} = x_0 + \frac{T_p(x) - T_p(x_0)}{T_p^*(x_0)}$$

Für  $c = 0$  und  $p \neq 0$  sind:

$$T_p(x) = (x + c)^p = x^p \quad T_p(x_0) = (x_0 + c)^p = x_0^p \quad T_p^*(x_0) = px_0^{p-1}$$

$$z = aT_p(x) + b = x_0 + \frac{T_p(x) - T_p(x_0)}{T_p^*(x_0)} = x_0 + \frac{x^p - x_0^p}{px_0^{p-1}}.$$

Für  $c = 0$  und  $p = 0$  sind:

$$T_p(x) = \ln(x + c) = \ln x \quad T_p(x_0) = \ln(x_0 + c) = \ln x_0 \quad T_p^*(x_0) = 1/x_0$$

$$\begin{aligned} z &= aT_p(x) + b = x_0 + \frac{T_p(x) - T_p(x_0)}{T_p^*(x_0)} \\ &= x_0 + \frac{\ln x - \ln x_0}{\frac{1}{x_0}} = x_0 + x_0(\ln x - \ln x_0) = x_0 \left[ 1 + \ln \left( \frac{x}{x_0} \right) \right]. \end{aligned}$$

Mit diesem Matching wird erreicht, dass ein großer Teil der Werte (vor allem der mittleren Werte) ein den Ausgangsdaten ähnliches Aussehen erhält und die Änderungen hervorgehoben werden, die durch die Transformation verursacht wurden.

Die durch eine Transformation erreichte Veränderung in Richtung Symmetrie der Verteilung kann durch die Berechnung der Schiefe und/oder der midsummaries der transformierten Werte, durch ein Stem-and-Leaf Plot bzw. mittels der erneuten Durchführung eines Tests eingeschätzt werden. Unter SPSS können Transformationen der Daten über

■ Transform

■ Compute...

vorgenommen werden (siehe Kapitel 1).

- Beispiel 3.4 (Fortsetzung):

Für die Variable Monatsmiete der Datei mieten.sav wurde bereits im Abschnitt 3.3.1 nachgewiesen, dass unter Verwendung der Ausgangsdaten die Prüfung auf Normalverteilung  $N(\mu; \sigma)$  mittels des Kolmogorov-Smirnov-Testes auf dem 5%-Niveau zur Ablehnung der Nullhypothese führt. Diese Variable weist eine deutliche rechtsschiefe Verteilung auf (siehe auch Abb. 3.19). Um eine Annäherung an eine symmetrische Verteilung zu erreichen, muß man somit die Leiter der Transformation hinabsteigen, d.h.,  $p < 1$  wählen. Wählt man speziell  $p = 0$  und  $c =$

0, so wird die Transformation  $T_0(\text{Monatsmiete}) = \ln(\text{Monatsmiete})$  vorgenommen. Für diese ln-Transformation der Monatsmiete wurde bereits anhand des Histogramms (Abb. 3.24) und Wahrscheinlichkeitsplots (Abb. 3.25 und 3.26 im Abschnitt 3.1) explorativ gezeigt, dass eine gute Annäherung an die Normalverteilung erreicht wird. Allerdings führt der Kolmogorov-Smirnov-Test mit Lillefors Korrektur weiterhin zur Ablehnung der Nullhypothese auf Normalverteilung zum 5%-Niveau.

Der folgende SPSS-Output gibt eine Übersicht für die mit verschiedenen Werten von  $p$  transformierte Variable Monatsmiete.

**SPSS-Output 3.4-4:** Schiefe, Kurtosis und Kolmogorov-Smirnov-Test (Lillefors) für die transformierte Monatsmiete

Statistics					
		Monatsmiete	Miete**0,5	Miete**0,1	ln(Miete)
N	Valid	815	815	815	815
	Missing	0	0	0	0
Skewness		1,727	,773	,061	-,116
Std. Error of Skewness		,086	,086	,086	,086
Kurtosis		4,482	1,152	,274	,290
Std. Error of Kurtosis		,171	,171	,171	,171

#### Tests of Normality

	Kolmogorov Smirnov <sup>a</sup>		
	Statistic	df	Sig.
Monatsmiete in DM	,117	815	,000
Miete**0,5	,063	815	,000
Miete**0,1	,028	815	,163
ln(Miete)	,038	815	,007

a. Lillefors Significance Correction

- Beispiel 3.10:

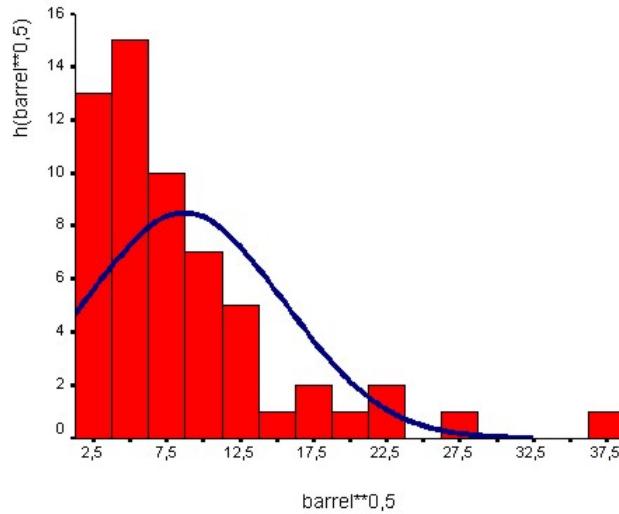
Die Variable barrel der Datei erdöl.sav beinhaltet die geförderte Menge (in Mio. Barrel) und wurde bereits im Abschnitt 2.5 für eine Untersuchung auf Ausreißer verwendet. Es zeigte sich, dass diese Variable eine rechtsschiefe Verteilung aufweist (siehe SPSS-Output 2.5-1 und Abb. 2.25). Darüber hinaus wurde statistisch nachgewiesen, dass die Bohrlöcher 20 und 8 als Ausreißer anzusehen sind. Die Mehrheit der Beobachtungswerte liegt zwischen 0 und 200 Mio. Barrel. Es werden 2 Transformationen durchgeführt, um eine Annäherung an eine symmetrische Verteilung zu erreichen. Dafür muß man einen Wert  $p < 1$  wählen.

$$p = 0,5 \quad T_{0,5}(\text{barrel}) = \text{SQR}(\text{barrel})$$

Der Hauptteil der transformierten Werte liegt zwischen 2 und 11.  $T_{0,5}(\text{barrel})$  weist jedoch noch eine deutliche rechtsschiefe Verteilung auf.

### 3. Prüfung der Verteilungform von Variablen

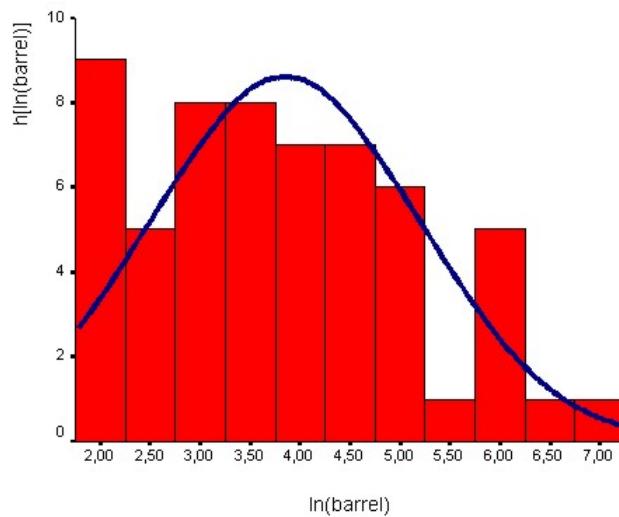
Abbildung 3.34.: Histogramm der Variablen SQR(barrel) des Beispiels 3.10



$$p = 0 \quad T_0(\text{barrel}) = \ln(\text{barrel})$$

Der Hauptteil der transformierten Werte liegt zwischen 2 und 5.  $T_0(\text{barrel})$  weist keine ausgeprägte rechtsschiefe Verteilung mehr auf.

Abbildung 3.35.: Histogramm der Variablen  $\ln(\text{barrel})$  des Beispiels 3.10

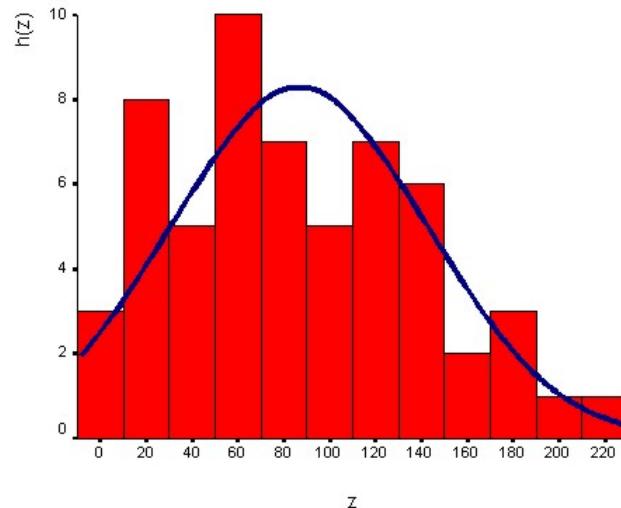


Es wird nunmehr eine matched Transformation angewandt, wobei  $x_0 = x_{0,5} = 41,5$  Mio. Barrel (Median der Ausgangswerte) gewählt wird. Die durchzuführende Transformation ist:

$$z = 41,5[1 + \ln(x/41,5)] + 40,$$

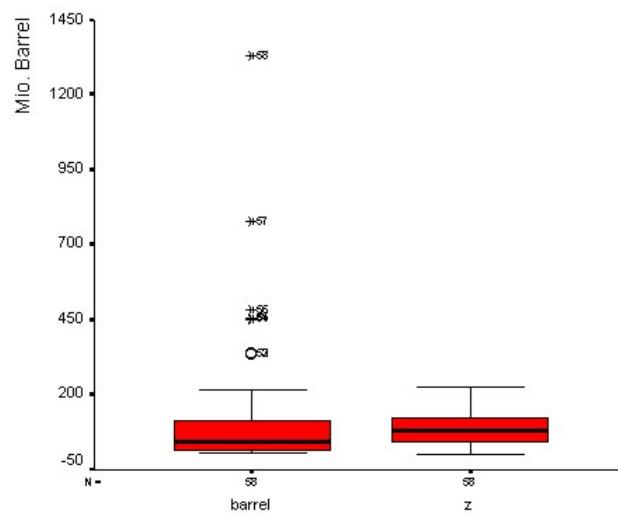
wobei die Addition von 40 vorgenommen wurde, um negative Werte zu vermeiden. Die Werte dieser transformierten Variablen liegen zwischen 0 und 225 Mio. Barrel.

Abbildung 3.36.: Histogramm der Variablen z des Beispiels 3.10



Durch das Matching erfolgt eine Rückverschiebung der Daten in den wesentlichen Bereich der Ausgangswerte. Die Ausreißer sind nicht mehr als solche zu erkennen.

Abbildung 3.37.: Boxplots der Variablen barrel und z des Beispiels 3.10



### *3. Prüfung der Verteilungform von Variablen*

## 4. Parametervergleiche bei unabhängigen Stichproben

Ein nächster möglicher Schritt statistischer Datenanalyse ist die Untersuchung, ob es wesentliche Unterschiede in der Verteilung bzw. den Parametern der Verteilung einer ausgewählten metrischen Variablen gibt, die nach den Ausprägungen einer zweiten Variablen (Faktorvariablen) separiert wird. Die Faktorvariable ist in der Regel eine nominal- oder ordinalskalierte Variable; sie kann jedoch auch eine metrisch diskrete Variable mit wenigen Ausprägungen sein. Da die Merkmalsausprägungen der Faktorvariablen festgelegt sind, dient diese Variable zur Identifizierung der Grundgesamtheiten. Aus den verschiedenen Grundgesamtheiten wird jeweils eine Zufallsstichprobe gezogen und die Stichprobenwerte der interessierenden Variablen erfaßt. Da jedes statistische Element aufgrund der ihm eigenen Ausprägung der Faktorvariablen eindeutig nur zu einer Grundgesamtheit gehört, sind die Beobachtungen der einen Stichprobe unabhängig von den Beobachtungen der anderen Stichprobe. In diesem Sinne spricht man von **unabhängigen Stichproben**. Für die weiteren Abhandlungen in diesem Kapitel werden unabhängige Stichproben zugrunde gelegt.

Zum Beispiel interessiert die Frage, ob Unterschiede in der Einkommensverteilung nach dem Geschlecht existieren. Zu diesem Zweck wird eine Stichprobe aus der Grundgesamtheit der männlichen Personen und eine Stichprobe aus der Grundgesamtheit der weiblichen Personen gezogen. Da jede Person nur zu einer der beiden Grundgesamtheiten gehören kann, sind die beobachteten Einkommenswerte der Stichprobe aus der Grundgesamtheit der männlichen Personen unabhängig von den beobachteten Einkommenswerten der Stichprobe aus der Grundgesamtheit der weiblichen Personen.

**Abhängige Stichproben<sup>25</sup>** (auch als verbundene, korrelierte, gepaarte Stichproben bezeichnet) ergeben sich immer dann, wenn zwischen den Elementen zweier (oder mehrerer) Stichproben eine gegenseitige Beeinflussung ihres Zufallsverhaltens bzw. eine gewisse Informationsbeziehung besteht. Das ist z.B. der Fall, wenn

---

<sup>25</sup>Vgl. u.a. Bortz, J. (1993), S. 135

#### 4. Parametervergleiche bei unabhängigen Stichproben

- die Elemente ein und derselben Stichprobe mehrmals auf ein Merkmal hin beobachtet werden. Die Wiederholung der Beobachtung kann z.B. nach einem zeitlichen Abstand oder unter veränderten Bedingungen vorgenommen werden.

Beispiel: Beobachtung einer biologischen Reaktion an Patienten, während und nach einer ärztlichen Behandlung; Veränderung der Einstellung vergleichbarer Individuen unter verschiedenen Bedingungen (beispielsweise Medieneinwirkung); Einschätzung ein und derselben Leistung durch zwei Prüfer; Veränderung der Wahrnehmung durch mehrmalige Betrachtung.

- nach einem bestimmten Kriterium Parallelstichproben gebildet werden, deren Elemente zufallsbedingt einer der Stichproben zugeordnet werden (parallelisierte Stichproben, matched samples).

Beispiel: Es werden Paare von Probanden gebildet, die hinsichtlich festgelegter Merkmale (z.B. Geschlecht, Alter) aufeinander abgestimmt und somit möglichst homogen sind. Die Zuordnung der Partner zu den beiden Stichproben (Versuchsgruppen) erfolgt zufällig. Jede Stichprobe von Probanden wird dann einer bestimmten Behandlungsmethode unterzogen, beispielsweise die Probanden der einen Gruppe (control subjects) mit einer Standardmethode und die Probanden der anderen Gruppe (experimental subjects) mit einer neuen Methode.

Die in diesem Kapitel durchgeführten Vergleichsprüfungen beziehen sich auf den Mittelwert und die Varianz der Verteilungen der interessierenden Variablen in den Grundgesamtheiten. Sie können mittels explorativer (graphischer) Möglichkeiten zunächst anschaulich und daran anschließend mittels statistischer Signifikanztests erfolgen.

### 4.1. Explorative Analyse

Für den explorativen und beschreibenden Vergleich stehen vor allem, unter Beachtung des jeweils erforderlichen Skalenniveaus, die bereits beschriebenen Möglichkeiten der Ausgabe von

- Häufigkeitstabellen mit entsprechenden statistischen Kennziffern,
- Stem-and-Leaf Plot (vgl. Abschnitt 2.1 und 3.1),
- Boxplot (vgl. Abschnitt 2.2 und 3.1),
- Histogramm (vgl. Abschnitt 3.1),
- Balkendiagramm (vgl. Abschnitt 3.1)

zur Verügung.

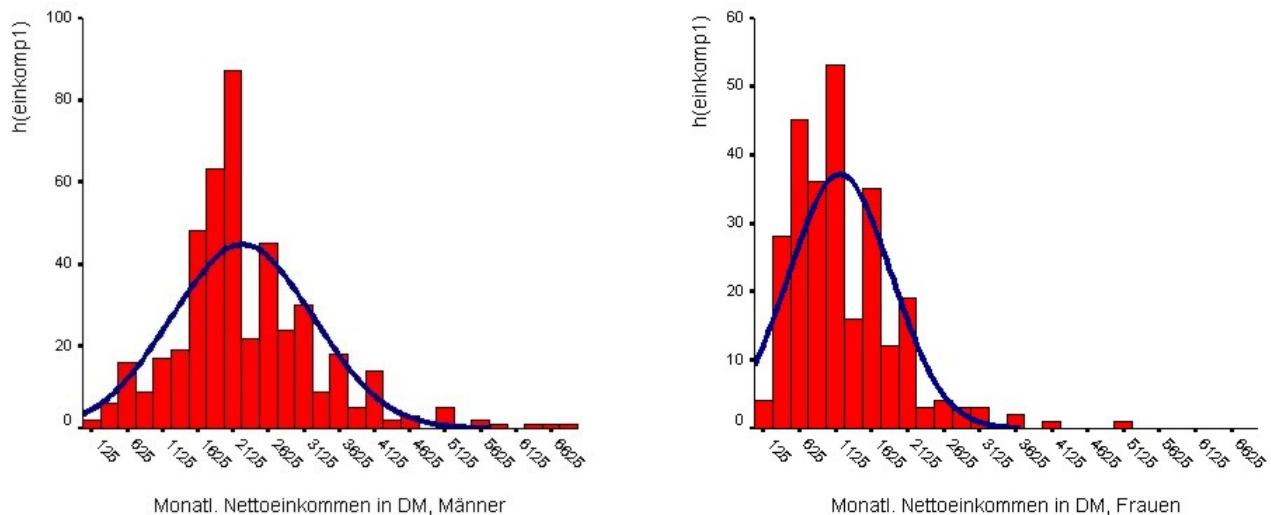
Die unter SPSS im Kontext dieses Kapitels geringfügigen Veränderungen zur Aktivierung dieser Möglichkeiten lassen sich am besten an einem Beispiel darstellen.

- Beispiel 4.1:

Es wird auf die bereits oben erwähnte Frage zurückgegriffen, ob sich Männer und Frauen bezüglich der Verteilung ihres monatlichen persönlichen Nettoeinkommens (Variable einkomp1 der Datei allbus.sav) unterscheiden.

Da das monatliche persönliche Nettoeinkommen bei dieser Variablen nicht klassiert vorliegt, ist die Ausgabe einer Häufigkeitstabelle nicht sinnvoll. Die Stem-and-Leaf Plots und Boxplots des monatlichen persönlichen Nettoeinkommens nach dem Geschlecht wurden bereits im Beispiel 3.2 (Abschnitt 3.1) mit dem SPSS-Output 3.2-1 und der Abbildung 3.9 erstellt. Die Histogramme gibt die Abb. 4.1 wieder, wobei der Ausreißer von 15000 DM bei den Männern herausgelassen wurde und zum besseren Vergleich in beiden Histogrammen Minimum = 0, Maximum 7000 und Klassenbreite = 250 gesetzt wurden.

Abbildung 4.1.: Histogramme des monatlichen persönlichen Nettoeinkommens nach Geschlecht



Die Statistiken der Verteilungen des monatlichen Nettoeinkommens nach dem Geschlecht (ebenfalls ohne den Ausreißer von 15000 DM) erhält man über

■ Analyze

■ Descriptive Statistics

■ Explore...

Im Dialogfeld „Explore“ (siehe Abb. 2.1) wird die Variable einkomp1 in das Feld „Dependent

#### 4. Parametervergleiche bei unabhängigen Stichproben

List:“ und die Variable Geschlechtszugehörigkeit (sex) in das Feld „Factor List.“ gebracht. Im Feld „Display“ wird nur auf „Statistics“ entschieden, da nur Statistiken ausgegeben werden sollen. Im Dialogfeld „Explore: Statistics...“ (siehe Abb. 2.3), das man durch Anwahl der Schaltfläche „Statistics“ erhält, wird die Voreinstellung von „Descriptives“ (univariate Statistiken) belassen. Die Ergebnisse sind im SPSS-Output 4.1-1 enthalten.

**SPSS-Output 4.1-1:** Univariate Statistiken des monatlichen persönlichen Nettoeinkommen nach dem Geschlecht

		Case Processing Summary					
		Cases					
		Valid		Missing		Total	
Geschlecht		N	Percent	N	Percent	N	Percent
Monatl. Nettoeinkommen in DM	Mann	450	33,2%	906	66,8%	1356	100,0%
	Frau	265	15,6%	1431	84,4%	1696	100,0%

Descriptives							
Geschlecht				Statistic	Std. Error		
Monatl. Nettoeinkommen in DM	Mann	Mean		2258,99		47,16	
		95% Confidence Interval for Mean	Lower Bound	2166,30			
			Upper Bound	2351,68			
		5% Trimmed Mean		2208,57			
		Median		2000,00			
		Variance		1000974,837			
		Std. Deviation		1000,49			
		Minimum		200			
		Maximum		7000			
		Range		6800			
		Interquartile Range		1070,00			
		Skewness		1,077		,115	
		Kurtosis		2,616		,230	
	Frau	Mean		1190,69		43,68	
		95% Confidence Interval for Mean	Lower Bound	1104,68			
			Upper Bound	1276,70			
		5% Trimmed Mean		1131,91			
		Median		1000,00			
		Variance		505612,934			
		Std. Deviation		711,06			
		Minimum		120			
		Maximum		5000			
		Range		4880			
		Interquartile Range		870,00			
		Skewness		1,514		,150	
		Kurtosis		4,089		,298	

Auswertung:

Die Stichprobe umfaßt bei den männlichen Befragten  $n_1 = 450$  und bei den weiblichen Befragten  $n_2 = 265$ , wobei auch die Feststellung interessant erscheint, dass weitaus mehr weibliche Befragte die Frage nicht beantwortet haben (Missing gleich 84,4% im Gegensatz zu 66,8% bei

den Männern).

Die Spannweite des Einkommens ist bei den Männern (ohne Ausreißer) um rund 1920 DM größer als bei den Frauen. Die Einkommensverteilung ist bezüglich der Lage bei den Männern deutlich zu den größeren Werten verschoben (Mittelwert und Median). Die Streuung des Nettoeinkommens (IQR und Standardabweichung) ist bei den Frauen geringer. Die beiden Einkommensverteilungen weichen deutlich von einer Normalverteilung ab (Skewness und Kurtosis).

### Fehlerbalken-Diagramm:

Eine zusätzliche grafische Option zur Beurteilung, ob wesentliche Unterschiede zwischen den Mittelwerten in den Stichproben existieren, ist durch das Fehlerbalken-Diagramm gegeben. Dazu ist zu wählen:

#### ■ Graphs

##### ■ Error Bar...

Im Dialogfeld „Error Bar“ (Abb. 4.2) sind die Entscheidungen analog zum Boxplot zu treffen, d.h., die Voreinstellungen für ein einfaches Fehlerbalken-Diagramm (Simple) nach den Kategorien einer Variablen (Summaries for groups of cases) können belassen werden. Nach Klick auf die Schaltfläche „Define“ erscheint das Dialogfeld „Define Simple Error Bar: Summaries for Groups of cases“ (Abb. 4.3).

Abbildung 4.2.: Dialogfeld „Error Bar“

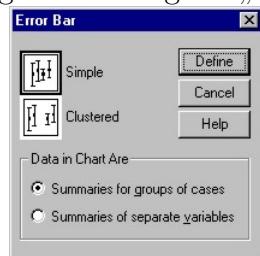
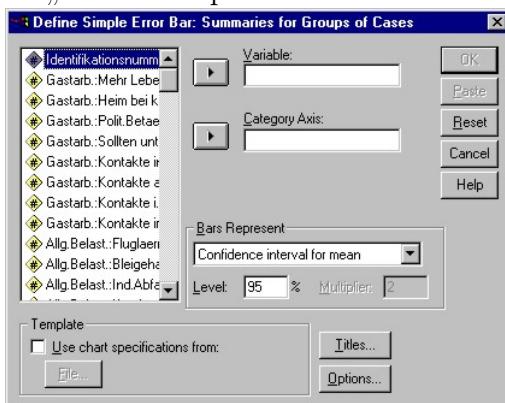


Abbildung 4.3.: Dialogfeld „Define Simple Error Bar: Summaries for groups of cases“



#### 4. Parametervergleiche bei unabhängigen Stichproben

Die zu analysierende Variable wird in das Feld „Variable:“ und die Faktorvariable, nach deren Kategorien unterschieden werden soll, in das Feld „Category Axis:“ gebracht.

Als Inhalt der Fehlerbalken (Feld „Bars Represent“) kann gewählt werden:

1. Confidence interval for mean

Konfidenzintervall für den Mittelwert, wobei das Konfidenzniveau im Feld „Level“ mit 95% vorgegeben ist, jedoch auch verändert werden kann;

2. Standard error of mean

Mittelwert  $\pm$  ein Vielfaches des Standardfehlers des Mittelwertes, wobei der Multiplikator (Multiplier) mit 2 vorgegeben ist, jedoch auch verändert werden kann;

3. Standard deviation

Mittelwert  $\pm$  ein Vielfaches der Standardabweichung, wobei der Multiplikator mit 2 vorgegeben ist, jedoch auch verändert werden kann.

Zu beachten ist, dass nur die Varianten 1 und 2 tatsächliche Konfidenzintervalle für den Erwartungswert  $\mu$  der Grundgesamtheit sind, da in beiden Fällen korrekterweise der Standardfehler des Mittelwertes ( $S_n^{-1/2}$ ) als Präzisionsmaß verwendet wird. Wird die betrachtete Variable mit X und demzufolge der Stichprobenmittelwert mit  $\bar{X}$  bezeichnet, so ergibt sich das Konfidenzintervall für  $\mu$  wie folgt:

$$\left[ \bar{X} - c_{1 - \frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{X} + c_{1 - \frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \right]. \quad (4.1)$$

Bei der ersten Variante kann das Konfidenzniveau  $1 - \alpha$  gewählt werden, für das sich dann aus der relevanten Verteilung (Normalverteilung bzw. t-Verteilung) der Multiplikator c ergibt. Bei der zweiten Variante kann der Multiplikator c gewählt werden, für den man aus der relevanten Verteilung das zugehörige Konfidenzniveau finden kann.

Die dritte Variante ist kein Konfidenzintervall für den Erwartungswert  $\mu$  der Grundgesamtheit, da als Präzisionsmaß die Standardabweichung der Variablen X verwendet wird:

$$\left[ \bar{X} - c_{1 - \frac{\alpha}{2}} \cdot S, \bar{X} + c_{1 - \frac{\alpha}{2}} \cdot S \right]. \quad (4.2)$$

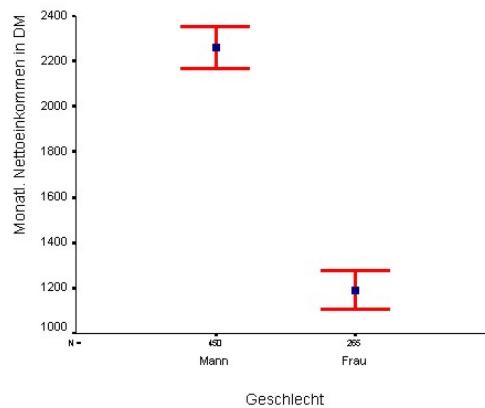
Es sollte deshalb entweder die Variante 1 oder 2 gewählt werden.

Auf jeden Fall sollte man weiterhin im Dialogfeld „Options“, das über die Schaltfläche „Options...“ zu erreichen ist, die Option „Display groups defined by missing values“ ausschalten.

- Beispiel 4.1 (Fortsetzung):

Für die geschlechtsspezifische Einkommensverteilung wird ein Fehlerbalkendiagramm als 95%iges Konfidenzintervall für den Mittelwert erstellt.

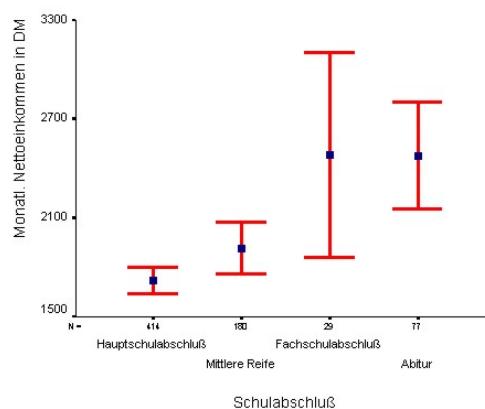
Abbildung 4.4.: Fehlerbalkendiagramm (95%iges Konfidenzintervall für den Mittelwert) für Nettoeinkommen nach Geschlecht



Da sich die beiden Konfidenzintervalle nicht überlappen, besteht offensichtlich ein signifikanter Unterschied im mittleren Einkommen der Männer und Frauen. Fazit ist somit, dass sich die beiden geschlechtsspezifischen Einkommensverteilungen hinsichtlich Niveau und Streuung unterscheiden. Dies gilt es durch Tests zu erhäusern.

In Erweiterung dieses Beispiels wird die Einkommensverteilung separiert nach den Ausprägungen der Faktorvariablen Schulabschluß (schule1) und ein Fehlerbalkendiagramm als 95%iges Konfidenzintervall für den Mittelwert erstellt.

Abbildung 4.5.: Fehlerbalkendiagramm (95%iges Konfidenzintervall für den Mittelwert) für Nettoeinkommen nach dem Schulabschluß



Einige der Konfidenzintervalle überlappen sich, woraus auf einen nichtsignifikanten Unterschied

#### 4. Parametervergleiche bei unabhängigen Stichproben

im mittleren monatlichen Nettoeinkommen geschlossen werden kann. Andererseits besteht z.B. ein signifikanter Mittelwertunterschied beim Nettoeinkommen der Personen mit Hauptschulabschluß und der Personen mit Fachschulabschluß bzw. Abitur.

## 4.2. Statistische Tests

### 4.2.1. Prüfung der Gleichheit der Varianzen

Die statistische Signifikanzprüfung sollte mit der Prüfung der Stichprobenvarianzen beginnen, da die Gleichheit bzw. Ungleichheit der Varianzen für den Test der Mittelwertdifferenzen von Bedeutung ist.

Es wird die Nullhypothese  $H_0$ : „Die Stichproben stammen aus Grundgesamtheiten mit gleichen Varianzen“ geprüft.

#### F-Test

Der F-Test ist ein Test, bei dem die Teststatistik einer F-Verteilung folgt. Er wird u.a. verwendet, um die Hypothese über die Gleichheit zweier Varianzen  $\sigma_1^2$  und  $\sigma_2^2$  zu prüfen. Diesem Test unterliegt das Hypothesenpaar:

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ gegen } H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Die Voraussetzungen des F-Tests sind:

1. Gegeben sind zwei Grundgesamtheiten mit den Zufallsvariablen  $X_1$  und  $X_2$ .
2. Die Zufallsvariablen  $X_1$  und  $X_2$  sind normalverteilt mit den Erwartungswerten  $\mu_1$  bzw.  $\mu_2$  und den Varianzen  $\sigma_1^2$  bzw.  $\sigma_2^2$ :  
$$X_1 \sim N(\mu_1; \sigma_1^2), \quad X_2 \sim N(\mu_2; \sigma_2^2).$$
3. Die zwei Zufallsvariablen  $X_1$  und  $X_2$  sind unabhängig voneinander.
4. Aus jeder Grundgesamtheit wird eine einfache Zufallsstichprobe  $(X_{1,1}, \dots, X_{1,n_1})$  bzw.  $(X_{2,1}, \dots, X_{2,n_2})$  gezogen, womit die Stichprobenvariablen in jeder Stichprobe identisch verteilt sind, die gleiche Verteilungsfunktion wie die Zufallsvariable in der Grundgesamtheit besitzen und unabhängige Zufallsvariablen sind. Die Stichprobenumfänge  $n_1$  und  $n_2$  müssen nicht gleich sein.
5. Die Beobachtungen der einen Stichprobe sind unabhängig von den Beobachtungen der anderen Stichproben, d.h., es handelt sich um unabhängige Stichproben.

Die Teststatistik des F-Test lautet

$$V = \frac{S_1^2}{S_2^2} \tag{4.3}$$

und folgt bei Gültigkeit der Nullhypothese einer F-Verteilung mit den Freiheitsgraden  $f_1 = n_1 - 1$  bzw.  $f_2 = n_2 - 1$ .

### Herleitung

Für jede Stichprobe wird der Stichprobenmittelwert und die Stichprobenvarianz bestimmt:

$$\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1,i}, \quad \bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2,i}. \quad (4.4)$$

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1,i} - \bar{X}_1)^2, \quad (4.5)$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2,i} - \bar{X}_2)^2$$

Die beiden Zufallsvariablen

$$\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} = \sum_{i=1}^{n_1} \left( \frac{X_{1,i} - \bar{X}_1}{\sigma_1} \right)^2 \quad (4.6)$$

$$\chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} = \sum_{i=1}^{n_2} \left( \frac{X_{2,i} - \bar{X}_2}{\sigma_2} \right)^2$$

folgen jeweils einer Chi-Quadrat-Verteilung mit  $f_1 = n_1 - 1$  bzw.  $f_2 = n_2 - 1$  Freiheitsgraden. Dies lässt sich wie folgt begründen:

- Laut Voraussetzung sind die beiden Zufallsvariablen  $X_1$  und  $X_2$  normalverteilt.
- Dann sind die Stichprobenvariablen  $X_{1i}$  und  $X_{2i}$  normalverteilt mit dem Erwartungswert  $\mu_1$  bzw.  $\mu_2$  und Varianz  $\sigma_1^2$  bzw.  $\sigma_2^2$ :  

$$X_{1i} \sim N(\mu_1; \sigma_1^2), i = 1, \dots, n_1 \quad \text{und} \quad X_{2i} \sim N(\mu_2; \sigma_2^2), i = 1, \dots, n_2$$
- Aufgrund dieser Gegebenheiten sind die Stichprobenfunktionen  $\bar{X}_1$  und  $\bar{X}_2$  normalverteilt mit dem Erwartungswert  $\mu_1$  bzw.  $\mu_2$  und Varianz  $\sigma_1^2/n_1$  bzw.  $\sigma_2^2/n_2$ :  

$$\bar{X}_1 \sim N(\mu_1; \sigma_1^2/n_1), \quad \text{und} \quad \bar{X}_2 \sim N(\mu_2; \sigma_2^2/n_2).$$
- Die standardisierten Zufallsvariablen

$$Z_{1,i} = \frac{X_{1,i} - \bar{X}_1}{\sigma_1}, \quad Z_{2,i} = \frac{X_{2,i} - \bar{X}_2}{\sigma_2}$$

folgen dementsprechend einer Standardnormalverteilung  $N(0;1)$ :

$$Z_{1i} \sim N(0; 1), i = 1, \dots, n_1 \quad \text{und} \quad Z_{2i} \sim N(0; 1), i = 1, \dots, n_2.$$

#### 4. Parametervergleiche bei unabhängigen Stichproben

- Die Zufallsvariablen  $Z_{1i}$  und  $Z_{2i}$  sind laut Voraussetzung innerhalb der Stichproben und zwischen den Stichproben unabhängig.
- $\chi_1^2$  und  $\chi_2^2$  beinhalten somit jeweils die Summe von quadrierten, unabhängigen und identisch standardnormalverteilten Zufallsvariablen. Derartig definierte Zufallsvariablen sind chi-quadrat-verteilt. Die Anzahl der Freiheitsgrade der chi-quadrat-verteilten Zufallsvariablen ergibt sich hier entsprechend der Anzahl der Freiheitsgrade der Stichprobenfunktionen  $S_1^2$  bzw.  $S_2^2$  zu  $f_1 = n_1 - 1$  bzw.  $f_2 = n_2 - 1$ .

Es wird nunmehr der Quotient dieser beiden auf ihre Freiheitsgrade bezogenen Zufallsvariablen  $\chi_1^2$  und  $\chi_2^2$  gebildet:

$$V = \frac{\frac{\chi_1^2}{f_1}}{\frac{\chi_2^2}{f_2}} = \frac{\frac{(n_1 - 1)S_1^2}{\sigma_1^2(n_1 - 1)}}{\frac{(n_2 - 1)S_2^2}{\sigma_2^2(n_2 - 1)}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \quad (4.7)$$

Da  $\sigma_1^2$  und  $\sigma_2^2$  feste Größen sind, ist V der Quotient zweier unabhängiger, chi-quadrat-verteilter Zufallsvariablen mit  $f_1$  und  $f_2$  Freiheitsgraden. Eine solcherart definierte Zufallsvariable folgt einer F-Verteilung mit  $f_1 = n_1 - 1$  und  $f_2 = n_2 - 1$  Freiheitsgraden. Bei Gültigkeit der Hypothese  $H_0$  ( $\sigma_1^2 = \sigma_2^2$ ) vereinfacht sich (4.7) zu (4.3).

Die Nullhypothese auf Gleichheit der Varianzen wird verworfen, wenn  $V < F_{n_1-1, n_2-1; \alpha/2}$  oder  $V > F_{n_1-1, n_2-1; 1-\alpha/2}$  ausfällt. Dabei sind  $F_{n_1-1, n_2-1; \alpha/2}$  und  $F_{n_1-1, n_2-1; 1-\alpha/2}$  die  $\alpha/2$ - bzw.  $(1-\alpha/2)$ -Quantile der F-Verteilung mit  $(n_1 - 1, n_2 - 1)$  Freiheitsgraden und  $\alpha$  das vorgegebene Signifikanzniveau.

Bei der praktischen Durchführung des F-Tests geht man im allgemeinen wie folgt vor: Von den Realisierungen  $s_1^2$  und  $s_2^2$  aus den beiden Stichproben verwendet man die größere Varianz als Zähler in (4.3) und die kleinere Varianz als Nenner. Es sei hier zur Vereinfachung angenommen, dass  $s_1^2$  die größere der beiden Stichprobenvarianzen sei. Damit wird  $H_0 : \sigma_1^2 = \sigma_2^2$  gegen  $H_1 : \sigma_1^2 > \sigma_2^2$  geprüft und als Ablehnungsbereich der Teststatistik  $V > F_{n_1-1, n_2-1; 1-\alpha}$  verwendet. Bei Gültigkeit der Nullhypothese wird die Teststatistik V einen Wert nahe Eins annehmen. Man lehnt deshalb die Nullhypothese  $H_0$  genau dann ab, wenn der so gebildete Quotient größer als  $F_{n_1-1, n_2-1; 1-\alpha}$  ausfällt, wobei  $n_1$  der Umfang der Zählerstichprobe und  $n_2$  der Umfang der Nennerstichprobe sind. Der F-Test ist durch die Verwendung von Varianzen ausreißerempfindlich, d.h., er ist kein sehr robuster Test.

#### Levene-Test

Da der F-Test empfindlich auf Abweichungen von der Normalverteilung reagiert, hat Levene einen approximativen Test zur Prüfung der Gleichheit von Varianzen vorgeschlagen.

Gegenüber dem F-Test ändern sich die erste und zweite Voraussetzung:

1. Gegeben sind  $m$  Grundgesamtheiten ( $m \geq 2$ ) mit den Zufallsvariablen  $X_1, \dots, X_m$ . Mit dem Levene-Test kann somit nicht nur die Gleichheit zweier, sondern mehrerer Varianzen geprüft werden.
2. Die Zufallsvariablen  $X_1, \dots, X_m$  weisen in den Grundgesamtheiten eine stetige Verteilung auf, d.h., die Normalverteilung ist nicht unbedingte Voraussetzung.

Die Voraussetzungen 3-5 bleiben unverändert, jedoch bezogen auf  $m \geq 2$ .

Geprüft wird  $H_0 : \sigma_1^2 = \dots = \sigma_m^2$  gegen  $H_1 : \sigma_j^2 \neq \sigma_k^2$  ( $j \neq k$ ).

Der Teststatistik von Levene liegen die absoluten Abweichungen der Stichprobenvariablen  $X_{j,i}$  ( $j = 1, \dots, m; i = 1, \dots, n_j$ ) vom Mittelwert  $\bar{X}_j$  der jeweiligen Stichprobe zugrunde:

$$Y_{j,i} = |X_{j,i} - \bar{X}_j| \quad j = 1, \dots, m, \quad i = 1, \dots, n_j. \quad (4.8)$$

Die Teststatistik ist definiert als

$$L = \frac{n-m}{m-1} \frac{\sum_{j=1}^m n_j (\bar{Y}_j - \bar{Y})^2}{\sum_{j=1}^m \sum_{i=1}^{n_j} (Y_{j,i} - \bar{Y}_j)^2} \quad (4.9)$$

mit

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{j,i}, \quad \bar{Y} = \frac{1}{n} \sum_{j=1}^m n_j \bar{Y}_j, \quad n = \sum_{j=1}^m n_j. \quad (4.10)$$

Im Zähler von (4.9) steht die Summe der Abweichungsquadrate zwischen den Stichproben, die durch die Faktorvariable erklärt wird. Im Nenner steht die Summe der Abweichungsquadrate innerhalb der Stichproben, die nicht durch die Wirkung der Faktorvariable erklärt werden kann. Die Teststatistik von Levene entspricht der Anwendung einer einfachen Varianzanalyse auf die absoluten Differenzen. Es wird somit geprüft, ob die Stichproben aus den Grundgesamtheiten (Verteilungen) mit gleichen mittleren Abweichungen stammen. Wenn dem nicht so ist, dann kann bei der Varianzanalyse die Nullhypothese der Gleichheit der Varianzen der Stichproben nicht aufrecht erhalten werden.

Die Teststatistik  $L$  folgt unter  $H_0$  einer F-Verteilung mit  $f_1 = m - 1$  und  $f_2 = n - m$  Freiheitsgraden.

Die Nullhypothese auf Gleichheit der Varianzen wird verworfen, wenn  $L > F_{m-1, n-m; 1-\alpha}$  ist. Dabei ist  $F_{m-1, n-m; 1-\alpha}$  das  $(1 - \alpha)$ -Quantil der F-Verteilung mit  $(m - 1, n - m)$  Freiheitsgraden und  $\alpha$  das vorgegebene Signifikanzniveau.

Unter SPSS ist nur der Levene-Test erhältlich, da er den Vergleich zweier Varianzen (wie beim F-Test) einschließt und nur eine stetige Verteilung in den Grundgesamtheiten voraussetzt.

#### 4. Parametervergleiche bei unabhängigen Stichproben

Er ist über

- Analyze
  - Descriptive Statistics
  - Explore

erhältlich. Im Dialogfeld „Explore“ (siehe Abb. 2.1) wird die zu analysierende Variable in das Feld „Dependent List:“ und die Faktorvariable in das Feld „Factor List:“ gebracht, im Feld „Display“ auf Both entschieden und die Schaltfläche „Plots...“ betätigt. Im folgenden Dialogfeld „Explore: Plots“ (siehe Abb. 2.2) wird nur im Feld „Spread vs. Level with Levene Test“ die Option untransformed angeklickt.

- Beispiel 4.1 (Fortsetzung):

Da das monatliche persönliche Nettoeinkommen in einem vorgegebenen Intervall sehr viele mögliche Werte annehmen kann und somit als eine quasi-stetige Variable behandelt wird, kann die Voraussetzung einer stetigen Verteilung in den geschlechtsspezifischen Grundgesamtheiten als erfüllt angesehen werden.

Zur Prüfung der Gleichheit der Varianzen wird der Levene-Test angewandt ( $\alpha = 0,05$ ), wobei wiederum der Ausreißer herausgelassen wird. Der für den Levene-Test relevante Teil des Outputs ist im SPSS-Output 4.1-2 enthalten. Die Zeile mit „Based on Mean“ ist im Kontext dieses Abschnittes die zu interpretierende Zeile.

Wegen  $m = 2$  und  $n_1 + n_2 = 450 + 265 = 715$  beträgt die Anzahl der Freiheitsgrade  $f_1 = 2 - 1 = 1$  (df1) und  $f_2 = 715 - 2 = 713$  (df2). Da  $Sig = 0,000 < \alpha = 0,05$  wird die Nullhypothese auf einem Signifikanzniveau von 5% verworfen, d.h., die beiden Stichproben stammen aus Grundgesamtheiten (Verteilungen) mit verschiedenen Varianzen.

**SPSS-Output 4.1-2:** Levene-Test des monatlichen persönlichen Nettoeinkommen nach dem Geschlecht

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Monatl. Nettoeinkommen in DM	Based on Mean	18,938	1	713	,000
	Based on Median	14,353	1	713	,000
	Based on Median and with adjusted df	14,353	1	651,250	,000
	Based on trimmed mean	17,701	1	713	,000

#### Spread-and-Level Plot

Neben dem Ergebnis des Levene-Tests erhält man im SPSS Viewer ein Spread vs. Level Plot (Streuung vs. Niveau Plot), wobei für Spread der Interquartilsabstand (IQR) und für Level der Median verwendet wird. Für jede Stichprobe wird ein Punkt (Median, IQR) in dem Koordinatensystem abgetragen.

Mit dieser Grafik kann eine Einschätzung getroffen werden, inwieweit eine Beziehung zwischen Niveau und Streuung besteht. Wenn keine solche Beziehung existiert, liegen die Punkte annähernd auf einer horizontalen Linie.

- Beispiel 4.1 (Fortsetzung):

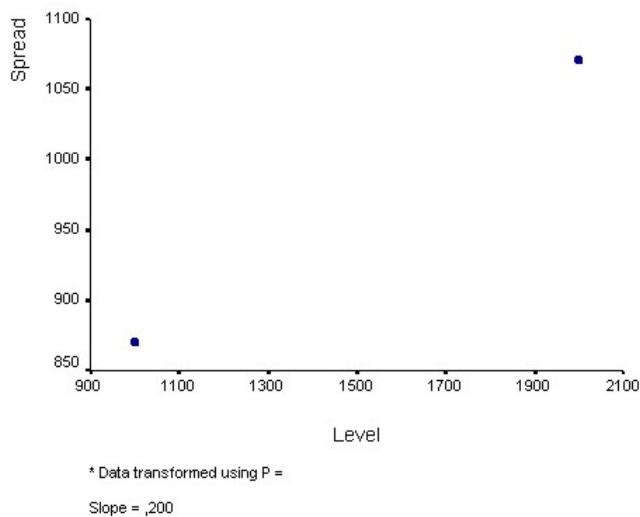
Aus dem SPSS-Output 4.1-1 können entnommen werden:

Mann:  $x_{0,5} = 2000$  IQR = 1070

Frau:  $x_{0,5} = 1000$  IQR = 870

Im Spread vs. Level Plot (Abb. 4.6) sind der Punkt (1000; 870) für die Einkommensstichprobe der Frauen und der Punkt (2000; 1070) für die Einkommensstichprobe der Männer enthalten.

Abbildung 4.6.: Spread vs. Level Plot von monatlichem Nettoeinkommen nach Geschlecht



Unterhalb des Koordinatensystems ist eine Information enthalten, in welcher Weise die Daten transformiert werden müßten, um annähernde Varianzgleichheit zu erzielen (siehe Abschnitt 3.4):

\* Data transformed using P =

Slope = ,200

P gibt den Wert der power of transformation an (siehe Tabelle 3.4) an. Da im Beispiel keine Transformation vorgenommen wurde, wird P = ausgegeben. Slope = 0,20 ist der Anstieg einer den Punkten angepaßten Geraden, d.h. der Regressionskoeffizient von Spread auf Level im Fall von nur zwei Stichproben:

$$slope = \frac{\sum_{j=1}^m (M_j - \bar{M})(IQR_j - \bar{IQR})}{\sum_{j=1}^m (M_j - \bar{M})^2}, \quad (4.11)$$

#### 4. Parametervergleiche bei unabhängigen Stichproben

wobei M für Level (Median) und IQR für Spread (Interquartilsabstand) steht.

Um eine Abschätzung des P-Wertes für die durchzuführende Transformation zu erhalten, wird nochmals „Spread vs. Level with Levene Test“, jedoch „Power Estimation“ angefordert. In diesem Fall ist die Ausschrift unterhalb der Grafik:

\* Plot of LN of Spread vs. LN of Level

Slope = ,299 Power for transformation = ,701

Die Power-Abschätzung erfolgt als  $P = 1 - \text{slope}$  und sollte dann zum nächstliegenden ,5-Wert gerundet werden. Im Beispiel ist eine Abrundung auf  $P = 0,5$  vorzunehmen, d.h., es sollte eine Transformation  $x^{0,5}$  durchgeführt werden. Den erreichten Effekt kann man durch Wiederholung des Levene-Tests und des Spread vs. Level Plots (bei Wahl von „Transformed“ mit Power: Square root) erkennen. Für die  $p = 0,5$  transformierten Werte des monatlichen persönlichen Nettoeinkommens wird die Nullhypothese auf Gleichheit der Varianzen auf dem 5%-Niveau nicht mehr abgelehnt.

Verwendet man (wie bereits weiter oben) die Faktorvariable Schulabschluß (schule1), so sind die 4 Punkte im Spread vs. Level Plot:

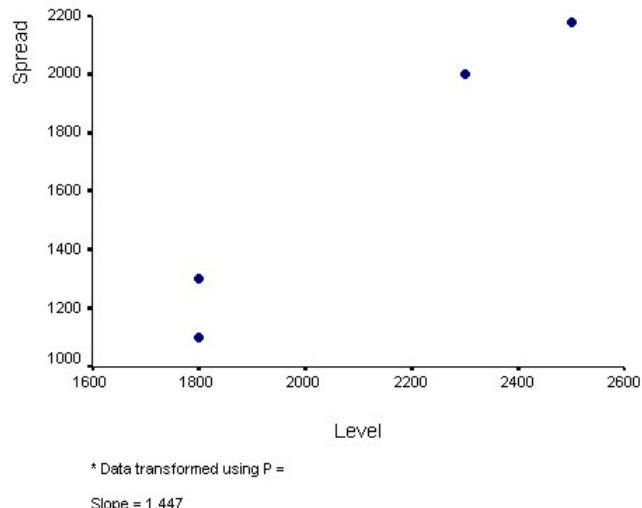
Hauptschulabschluß:  $(x_{0,5}; IQR) = (1800; 1100)$

Mittlere Reife:  $(x_{0,5}; IQR) = (1800; 1300)$

Fachschulabschluß:  $(x_{0,5}; IQR) = (2300; 2000)$

Abitur:  $(x_{0,5}; IQR) = (2500; 2175)$

Abbildung 4.7.: Spread vs. Level Plot von monatlichem Nettoeinkommen nach Schulabschluß



Auch bei der Faktorvariablen Schulabschluß wird die Nullhypothese auf Gleichheit der Varianzen des monatlichen Nettoeinkommens auf dem 5%-Niveau abgelehnt.

## 4.2.2. Prüfung der Gleichheit der Mittelwerte mittels parametrischer Tests

Bei dieser Prüfung muss man verschiedene Fälle unterscheiden, die sich aufgrund unterschiedlicher Voraussetzungen ergeben, die vor der Testdurchführung zu überprüfen sind<sup>26</sup>.

### 4.2.2.1. Test der Mittelwerte zweier Grundgesamtheiten mit gleichen, unbekannten Varianzen

Bei diesem Test handelt es sich um einen Parametertest, da eine Hypothese über einen unbekannten Parameter, die Differenz zweier Erwartungswerte  $\mu_1 - \mu_2$ , geprüft wird. Er beruht auf den Ergebnissen zweier Zufallsstichproben und wird deshalb als Zweistichprobentest bezeichnet.

Die Voraussetzungen dieses Zweistichproben-t-Tests sind:

1. Gegeben sind zwei Grundgesamtheiten. In der ersten Grundgesamtheit weist die Zufallsstichprobe  $X_1$  den Erwartungswert  $E(X_1) = \mu_1$  und die Varianz  $Var(X_1) = \sigma_1^2$  und in der zweiten Grundgesamtheit die Zufallsvariable  $X_2$  den Erwartungswert  $E(X_2) = \mu_2$  und die Varianz  $Var(X_2) = \sigma_2^2$  auf.  $\mu_1$  und  $\mu_2$  sind unbekannt.
2. Aus jeder Grundgesamtheit wird eine einfache Zufallsstichprobe gezogen bzw. es wird unterstellt, dass die Umfänge der beiden Grundgesamtheiten  $N_1$  und  $N_2$  genügend groß sind, dass von der Realisierung einfacher Zufallsstichproben ausgegangen werden kann. Die Stichprobenumfänge sind  $n_1$  und  $n_2$ .
3. Die beiden Zufallsstichproben  $X_{1,1}, \dots, X_{1,n_1}$  und  $X_{2,1}, \dots, X_{2,n_2}$  sind unabhängig voneinander.
4. Entweder sind die Zufallsvariablen  $X_1$  und  $X_2$  in den Grundgesamtheiten normalverteilt, d.h.  $X_1 \sim N(\mu_1, \sigma_1)$  und  $X_2 \sim N(\mu_2, \sigma_2)$ , oder die Stichprobenumfänge  $n_1$  und  $n_2$  sind genügend groß, dass der Zentrale Grenzwertsatz wirksam wird (approximativer Test auf  $\mu_1 - \mu_2$ ).
5. Die beiden Grundgesamtheiten haben eine gleiche, jedoch unbekannte Varianz:  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  (Varianzhomogenität).

Für die weiteren Herleitungen wird außerdem unterstellt, dass die Umfänge der beiden Grundgesamtheiten  $N_1$  und  $N_2$  genügend groß sind, so dass der Korrekturfaktor für endliche Grundgesamtheiten vernachlässigt werden kann.

---

<sup>26</sup>Die nachfolgend aufgeführten Tests sind in Standard-Lehrbüchern der Statistik enthalten.

#### 4. Parametervergleiche bei unabhängigen Stichproben

Über die Differenz der beiden Erwartungswerte existiert eine Annahme mit hypothetischen Wert  $\mu_1 - \mu_2 = \omega_0$ . Von besonderem Interesse bei der praktischen Anwendung dieses Tests ist oftmals die Gleichheit der beiden Erwartungswerte  $\mu_1 = \mu_2$ , womit  $\omega_0 = 0$  ist. Diese Hypothese steht auch hier im Vordergrund des Interesses.

Der Test wird auf dem Signifikanzniveau  $\alpha$  durchgeführt.

Geprüft werden kann eine

- zweiseitige Fragestellung

$$H_0 : \mu_1 = \mu_2 \text{ bzw. äquivalent } H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 \neq \mu_2 \text{ bzw. äquivalent } H_1 : \mu_1 - \mu_2 \neq 0$$

- einseitige Fragestellung

- linksseitiger Test

$$H_0 : \mu_1 \geq \mu_2 \text{ bzw. äquivalent } H_0 : \mu_1 - \mu_2 \geq 0$$

$$H_1 : \mu_1 < \mu_2 \text{ bzw. äquivalent } H_1 : \mu_1 - \mu_2 < 0$$

- rechtsseitiger Test

$$H_0 : \mu_1 \leq \mu_2 \text{ bzw. äquivalent } H_0 : \mu_1 - \mu_2 \leq 0$$

$$H_1 : \mu_1 > \mu_2 \text{ bzw. äquivalent } H_1 : \mu_1 - \mu_2 > 0.$$

#### Herleitung der Teststatistik

Wegen der 4. Voraussetzung sind die Stichprobenmittelwerte  $\bar{X}_1$  und  $\bar{X}_2$  zumindest approximativ normalverteilt:

$$\bar{X}_1 \sim N(\mu_1; \sigma_1^2/n_1) \quad \text{und} \quad \bar{X}_2 \sim N(\mu_2; \sigma_2^2/n_2).$$

Aufgrund der Reproduktivitätseigenschaft der Normalverteilung folgt, dass die Differenz der beiden Stichprobenmittelwerte (als eine spezielle Linearkombination)

$$D = \bar{X}_1 - \bar{X}_2 \tag{4.12}$$

ebenfalls (zumindest approximativ) normalverteilt ist mit dem Erwartungswert

$$E(D) = E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2. \tag{4.13}$$

Wegen der Unabhängigkeit der Stichprobenvariablen (2. und 3. Voraussetzung) ist die Varianz von D gegeben mit<sup>27</sup>:

$$Var(D) = \sigma_D^2 = Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \tag{4.14}$$

---

<sup>27</sup>Die Varianz der Differenz zweier Zufallsvariablen X - Y ergibt sich zu:

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y).$$

Wenn X und Y unabhängig voneinander sind, ist  $\text{Cov}(X, Y) = 0$ .

Die standardisierte Zufallsvariable

$$Z = \frac{D - E(D)}{\sigma_D} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (4.15)$$

ist standardnormalverteilt  $N(0;1)$ . Wegen der 5. Voraussetzung (Varianzhomogenität in den Grundgesamtheiten) folgt

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}} \quad (4.16)$$

Hierin ist jedoch  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  unbekannt und muss aus den Stichproben geschätzt werden. Geeignete Schätzfunktionen sind mit (4.5) angegeben. Die Zufallsvariablen  $\chi_1^2 = (n_1 - 1)S_1^2/\sigma^2$  und  $\chi_2^2 = (n_2 - 1)S_2^2/\sigma^2$  (4.6) folgen jeweils einer Chi-Quadrat-Verteilung mit  $f_1 = n_1 - 1$  bzw.  $f_2 = n_2 - 1$  Freiheitsgraden, wie bereits gezeigt wurde. Wegen der Unabhängigkeit dieser beiden Zufallsvariablen und aufgrund der Reproduktivitätseigenschaft der Chi-Quadrat-Verteilung ist auch

$$\chi^2 = \chi_1^2 + \chi_2^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \quad (4.17)$$

chi-quadrat-verteilt mit  $f_1 + f_2 = n_1 + n_2 - 2$  Freiheitsgraden.

Die Zufallsvariable

$$T = \frac{Z}{\sqrt{\frac{\chi^2}{f}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2} \cdot \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \quad (4.18)$$

folgt einer t-Verteilung mit  $f = n_1 + n_2 - 2$  Freiheitsgraden. Darin ist der Schätzer für die gemeinsame Varianz  $\sigma^2$  beider Grundgesamtheiten

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad (4.19)$$

das gewogene arithmetische Mittel aus den beiden Stichprobenvarianzen und wird als pooled variance bezeichnet.

Wird speziell die Nullhypothese  $H_0 : \mu_1 - \mu_2 = 0$  getestet, so ist die für den Zweistichproben-t-Test zu verwendende Teststatistik mit

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2} \cdot \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \quad (4.20)$$

gegeben, die bei Gültigkeit der Nullhypothese einer t-Verteilung mit  $f = n_1 + n_2 - 2$  Freiheitsgraden folgt.

#### 4. Parametervergleiche bei unabhängigen Stichproben

Für das vorgegebene Signifikanzniveau  $\alpha$  und die Freiheitsgrade  $f = n_1 + n_2 - 2$  findet man die kritischen Werte  $t_{f;1-\alpha/2}$ ,  $t_{f;\alpha}$  bzw.  $t_{f;1-\alpha}$  als Quantile der Ordnung  $1 - \alpha/2$ ,  $\alpha$  bzw.  $1 - \alpha$  aus der Tabelle der Verteilungsfunktion der t-Verteilung. Für die einzelnen Testvarianten erhält man die nachstehenden Entscheidungsbereiche bei Gültigkeit der Nullhypothese  $H_0$  und vorgegebenem Signifikanzniveau  $\alpha$ :

Testvariante	Ablehnungsbereich der $H_0$	Nichtablehnungsbereich der $H_0$
zweiseitig	$\{t   t < -t_{f;1-\alpha/2} \text{ oder } t > t_{f;1-\alpha/2}\}$	$\{t   -t_{f;1-\alpha/2} \leq t \leq t_{f;1-\alpha/2}\}$
rechtsseitig	$\{t   t > t_{f;1-\alpha}\}$	$\{t   t \leq t_{f;1-\alpha}\}$
linksseitig	$\{t   t < -t_{f;1-\alpha}\}$	$\{t   t \geq -t_{f;1-\alpha}\}$

Bei genügend großen Stichprobenumfängen ( $n_1 > 30$  und  $n_2 > 30$ ) ist aufgrund der Wirksamkeit des Zentralen Grenzwertsatzes die Teststatistik  $T$  unter  $H_0$  approximativ  $N(0;1)$ -verteilt. Es können dann die kritischen Werte aus der Tabelle der Verteilungsfunktion der Standardnormalverteilung entnommen werden.

##### 4.2.2.2. Test der Mittelwerte zweier Grundgesamtheiten mit ungleichen, unbekannten Varianzen

Die Voraussetzungen 1 - 4 sind wie vorher. Die 5. Voraussetzung verändert sich zu: Die beiden Grundgesamtheiten haben verschiedene, jedoch unbekannte Varianzen:  $\sigma_1^2 \neq \sigma_2^2$  (Varianzheterogenität).

Die zu prüfende Nullhypothese ist wie vorher, und die Ausführungen zur Herleitung der Teststatistik und ihrer Verteilung sind bis Formel (4.15) analog zum Abschnitt 4.2.2.1.

$$Z = \frac{D - E(D)}{\sigma_D} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (4.15)$$

In (4.15) sind nunmehr die beiden Varianzen  $\sigma_1^2$  und  $\sigma_2^2$  ungleich, und man steht vor dem sogenannten, bisher nicht gelösten Behrens-Fisher-Problem. Welch<sup>28</sup> (1947) hat einen Lösungsansatz vorgeschlagen, indem die Verteilung der Teststatistik

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (4.21)$$

---

<sup>28</sup>Welch, B. L. (1947), The generalization of Students problem when several different population variances are involved, Biometrika 34, 28-35.

mit  $S_1^2$  und  $S_2^2$  als die Stichprobenvarianzen gemäß (4.5) durch eine t-Verteilung mit

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \quad (4.22)$$

Freiheitsgraden (gerundet zur nächsten ganzen Zahl) approximiert wird. Der Nenner von (4.21) ist in diesem Fall der Standardfehler der Zufallsvariablen  $D = \bar{X}_1 - \bar{X}_2$  (unpooled variance, separate variance). Dieser Welch-Test ist ein Näherungstest. Die Testentscheidung ist wie im Abschnitt 4.2.2.1 angegeben zu treffen.

Diese beiden Tests aus Abschnitt 4.2.2.1 und 4.2.2.2 zum Zweistichproben-Problem sind unter SPSS über

- Analyze
  - Compare Means
  - Independent-Samples T Test...

verfügbar. Im sich öffnenden Dialogfeld „Independent-Samples T Test“ (Abb. 4.8) ist die zu testende Variable in das Feld „Test Variable(s):“ und die Faktorvariable, nach deren Ausprägungen sich die beiden Grundgesamtheiten unterscheiden, in das Feld „Grouping Variable“ zu bringen.

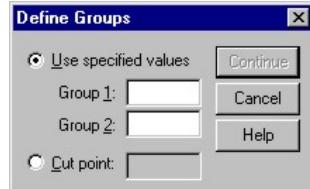
Abbildung 4.8.: Dialogfeld „Independent-Samples T Test“



Da es sich hier um einen Zweistichprobentest handelt, müssen noch die beiden Gruppen, d.h. die beiden Ausprägungen der Faktorvariablen, nach denen sich die Stichproben unterscheiden, festgelegt werden. Um die beiden Gruppen, die zunächst durch Fragezeichen hinter der Gruppierungsvariable gekennzeichnet sind, zu definieren, ist die Schaltfläche „Define Groups...“ anzuklicken. Es erscheint das Dialogfeld „Define Groups“.

#### 4. Parametervergleiche bei unabhängigen Stichproben

Abbildung 4.9.: Dialogfeld „Define Groups“



Dieses Dialogfeld erhält man, wenn die Gruppierungavariable (im Sinne von SPSS) eine numerische Variable ist. Es werden für Group 1 und Group 2 die entsprechenden Werte der Gruppierungsvariable eingetragen, sofern die beiden Gruppen dadurch bereits eindeutig definiert sind. Fälle mit anderen Werten werden bei der Testdurchführung nicht berücksichtigt. Für eine metrische Gruppierungsvariable muss dagegen ein Wert (Cut point) festgelegt werden, der die Elemente der beiden Gruppen trennt. Dabei gilt, dass alle Fälle mit einem Beobachtungswert größer oder gleich dem Trennwert zu einer Gruppe und alle anderen Fälle zur anderen Gruppe gehören.

Für eine (im Sinne von SPSS) String-Variable als Gruppierungsvariable besteht das Dialogfeld „Define Groups“ nur aus den beiden Eingabefeldern für Group 1 und Group 2, wobei die Kategorien (Begriffe) einzugeben sind, die die beiden Gruppen kennzeichnen. Alle anderen Fälle werden von der Analyse ausgeschlossen.

- Beispiel 4.1 (Fortsetzung):

Bevor der Test auf Mittelwertunterschiede der geschlechtsspezifischen Einkommensverteilung (ohne Ausreißer von 15000 DM) durchgeführt werden kann, müssen die Voraussetzungen überprüft werden:

- Die beiden Stichproben sind unabhängig, da die Beobachtungen der Stichprobe der Männer unabhängig von den Beobachtungen der Stichprobe der Frauen sind.
- Sie stammen jedoch nicht aus normalverteilten Grundgesamtheiten, wie oben geprüft wurde. Die Stichprobenumfänge sind mit  $n_1 = 450$  (Mann) und  $n_2 = 265$  (Frau) aber genügend groß, so dass der zentrale Grenzwertsatz Anwendung findet.
- Die Varianzen der beiden Grundgesamtheiten sind ungleich, wie ebenfalls bereits getestet wurde, d.h., es ist der Welch-Test zu verwenden.

Es soll auf einem Signifikanzniveau von  $\alpha = 0,05$  geprüft werden, ob die Einkommensverteilung der Männer einen größeren Mittelwert aufweist als die der Frauen:

$$H_0 : \mu_1 \leq \mu_2 \text{ gegen } H_1 : \mu_1 > \mu_2.$$

Im Dialogfeld „Independent-Samples T Test“ wird die Variable einkomp1 in das Feld „Test Variable(s):“ und die Variable sex in das Feld „Grouping Variable:“ gebracht. Da die Gruppierungsvariable sex eine nominalskalierte Variable ist, deren Ausprägungen zahlenmäßig kodiert wurden (1 - Mann, 2 - Frau), sind die Gruppen (Stichproben) dadurch eindeutig definiert. Im Dialogfeld „Define Groups“ wird 1 bei Group 1 und 2 bei Group 2 eingetragen. Man erhält nachstehenden Output.

**SPSS-Output 4.1-3:** Zweistichproben-t-Test für monatliches persönliches Nettoeinkommen nach Geschlecht

Group Statistics					
		Geschlecht	N	Mean	Std. Deviation
Monatl. Nettoeinkommen in DM	Mann	450	2258,99	1000,49	47,16
	Frau	265	1190,69	711,06	43,68

Independent Samples Test			
			Levene's Test for Equality of Variances
			F Sig.
Monatl. Nettoeinkommen in DM	Equal variances assumed	18,939	,000
	Equal variances not assumed		

t-test for Equality of Means						
t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
					Lower	Upper
15,258	713	,000	1068,30	70,01	930,84	1205,76
16,619	688,313	,000	1068,30	64,28	942,09	1194,52

Zunächst werden in der Teiltabelle „Group Statistics“ für den Test wichtige Stichprobenergebnisse, wie Stichprobenumfänge  $n_j$ , die Realisationen der Stichprobenfunktionen  $\bar{X}_j$  (Mean),  $S_j$  (Std. Deviation),  $S_j(\bar{X}_j)$  (Std. Error Mean) für  $j = 1,2$  ausgegeben.

In der Teiltabelle „Independent Samples Test“ weist der Levene-Test auch hier nochmals auf Varianzheterogenität hin. In den darauffolgenden Spalten sind die Ergebnisse des t-Tests bzw. des Welch-Tests enthalten. Unter SPSS wird der Zweistichproben-t-Test sowohl für den Fall der Varianzhomogenität (Zeile Equal variances assumed) als auch der Varianzheterogenität (Zeile Equal variances not assumed) durchgeführt. Der Nutzer hat die für seine Gegebenheiten gültige Zeile zu wählen. Für dieses Beispiel ist es die Zeile Equal Variances not assumed, da

#### 4. Parametervergleiche bei unabhängigen Stichproben

beim Levene-Test die Nullhypothese auf Varianzgleichheit abgelehnt wurde.

Der aufgrund der Stichprobenergebnisse berechnete Werte der Teststatistik T (4.21) und die zugehörige Anzahl der Freiheitsgrade nach (4.22) sind in der Spalte t bzw. df angegeben. Es wird stets ein zweiseitiger Test durchgeführt (Sig. 2-tailed). Bei einem einseitigen Test ist dieses ausgegebene Signifikanzniveau durch 2 zu teilen (man achte auch auf das Vorzeichen des t-value). Der Stichprobenwert der Zufallsvariablen D nach (4.12) ist in der Spalte Mean Difference, der Standardfehler für D (Nenner von (4.21)) ist in der Spalte Std. Error Difference und ein 95%-iges Konfidenzintervall für D in den letzten beiden Spalten enthalten.

Hinweis: Durch die Betätigung der Schaltfläche „Options...“ im Dialogfeld „Independent-Samples T Test“ gelangt man in ein Dialogfeld „Independent-Samples T Test: Options“, in dem die Wahrscheinlichkeit  $1 - \alpha$  für das Konfidenzintervall verändert werden kann.

Da  $Sig/2 < \alpha$  ist, wird die Nullhypothese verworfen, d.h. auf einem Signifikanzniveau von  $\alpha = 0,05$  und basierend auf Stichproben vom Umfang 450 und 265 konnte statistisch bewiesen werden, dass das mittlere persönliche Nettoeinkommen der Männer größer als das der Frauen ist.

- Beispiel 4.2:

Für das monatliche persönliche Nettoeinkommen (ohne Ausreißer von 15000 DM) der Datei allbus.sav soll auf dem 5%-Niveau geprüft werden, ob es Mittelwertunterschiede hinsichtlich des Alters gibt (zweiseitiger Test). Da das Alter eine metrische Variable ist, muss sie für die Durchführung des Zweistichproben-t-Test dichotomisiert werden. Dazu wird im Dialogfeld „Define Groups“ (Abb. 4.9) ein Cut point eingegeben. Der Test wird hier mit zwei verschiedenen Cut points durchgeführt, um zu zeigen, dass die Wahl des Cut point entscheidend für die Testbeantwortung sein kann.

- a) Cut point = 45

Befragte Personen in einem Alter größer oder gleich 45 Jahre bilden die eine Gruppe und alle anderen befragten Personen mit einem Alter kleiner als 45 Jahre gehören zur anderen Gruppe.  
Prüfung der Voraussetzungen:

- Die beiden Stichproben sind unabhängig, da die Beobachtungen der Stichprobe der Personen mit einem Alter  $\geq 45$  Jahre unabhängig von den Beobachtungen der Stichprobe der Personen mit einem Alter  $< 45$  Jahre sind.
- Sie stammen nicht aus einer normalverteilten Grundgesamtheiten. Die Stichprobenumfänge sind mit  $n_1 = 339$  und  $n_2 = 376$  aber genügend groß, so dass der zentrale Grenzwertsatz Anwendung findet.
- Die Varianzen der beiden Grundgesamtheiten können aufgrund des Levene-Tests als gleich angesehen werden, so dass zur Auswertung des Zweistichproben-t-Tests die Zeile “Equal

variances assumed“ verwendet wird.

**SPSS-Output 4.2-1:** Zweistichproben-t-Test für monatliches Nettoeinkommen nach Altersgruppen (Cut point: 45)

Group Statistics					
	Lebens-alter	N	Mean	Std. Deviation	Std. Error Mean
Monatl. Nettoeinkommen in DM	>= 45	339	2014,63	1066,66	57,93
	< 45	376	1726,38	998,52	51,49
Independent Samples Test					
					Levene's Test for Equality of Variances
					F Sig.
Monatl. Nettoeinkommen in DM	Equal variances assumed			,891	,346
	Equal variances not assumed				
t-test for Equality of Means					
t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
					Lower Upper
3,732	713	,000	288,25	77,25	136,59 439,90
3,719	693,112	,000	288,25	77,51	136,06 440,43

Da  $Sig < \alpha$  ist, wird die Nullhypothese verworfen, d.h., auf einem Signifikanzniveau von  $\alpha = 0,05$  und basierend auf den Stichproben vom Umfang 339 und 376 konnte statistisch bewiesen werden, dass das mittlere persönliche Nettoeinkommen der Personen mit einem Alter  $\geq 45$  verschieden vom mittleren persönlichen Nettoeinkommen der Personen mit einem Alter  $< 45$  Jahre ist.

b) Cut Point = 50

Befragte Personen mit einem Alter größer oder gleich 50 Jahre bilden die eine Gruppe und alle befragten mit einem Alter kleiner 50 Jahre gehören zur anderen Gruppe.

Die Prüfung der Voraussetzungen führt zu den gleichen Ergebnissen wie unter a).

#### 4. Parametervergleiche bei unabhängigen Stichproben

**SPSS-Output 4.2-2:** Zweistichproben-t-Test für monatliches Nettoeinkommen nach Altersgruppen (Cut point: 50)

Group Statistics					
	Lebens-alter	N	Mean	Std. Deviation	Std. Error Mean
Monatl. Nettoeinkommen in DM	$\geq 50$	282	1893,72	951,76	56,68
	< 50	433	1843,07	1095,32	52,64

Independent Samples Test			
			Levene's Test for Equality of Variances
			F Sig.
Monatl. Nettoeinkommen in DM	Equal variances assumed Equal variances not assumed	1,415	,235

t-test for Equality of Means						
t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
					Lower	Upper
,636	713	,525	50,65	79,67	-105,76	207,06
,655	656,913	,513	50,65	77,35	-101,23	202,53

Hier zeigt sich jedoch, dass wegen  $Sig > \alpha$  keine Veranlassung besteht, die Nullhypothese zu verwerfen. Auf einem Signifikanzniveau von  $\alpha = 0,05$  und basierend auf Stichproben vom Umfang 282 und 433 konnte statistisch nicht bewiesen werden, dass das mittlere persönliche Nettoeinkommen der Personen mit einem Alter  $\geq 50$  Jahre verschieden vom mittleren persönlichen Nettoeinkommen der Personen mit einem Alter  $< 50$  Jahre ist.

Hinweis:

Die Tests auf Normalverteilung, Gleichheit der Varianzen und Gleichheit der Mittelwerte sollten nicht auf der Grundlage der gleichen Stichproben durchgeführt werden, da sonst als Ganzes das vorgegebene Signifikanzniveau nicht eingehalten wird.

Da mit der Datei allbus.sav nur eine Untersuchung zur Verfügung steht und entsprechend Vorversuche hier nicht möglich sind, wurden zu Demonstrationszwecken die Tests aufgrund derselben Stichproben durchgeführt.

#### 4.2.2.3. Test der Mittelwerte mehrerer Grundgesamtheiten

Soll die Gleichheit der Mittelwerte von m Grundgesamtheiten geprüft werden, so kann die Varianzanalyse<sup>29</sup> (one-way analysis of variance; ANOVA) verwendet werden. Ihr Name ist darauf zurückzuführen, dass der Mittelwertvergleich über die Analyse der Varianzen erfolgt, d.h., die Stichprobenvarianzen in die Teststatistik eingehen.

Einfache oder einfaktorielle Varianzanalyse bedeutet, dass eine abhängige und eine unabhängige Variable in die Analyse eingehen, wie das im Kontext dieses Kapitels der Fall ist. Die Varianzanalyse dient jedoch auch der Untersuchung mehrerer abhängiger und unabhängiger Variablen (multiple oder mehrfaktorielle Varianzanalyse, multivariate analysis of variance, MANOVA).

Die Voraussetzungen der ANOVA sind:

1. Die abhängige Variable X (auch als Zielgröße bezeichnet) muss metrisches Skalenniveau aufweisen. Bezuglich der unabhängigen Faktorvariablen (Faktor) sind an das Skalenniveau keine Voraussetzungen gebunden; eine stetige Faktorvariable macht jedoch i.a. wegen der vielen Faktorausprägungen keinen Sinn.
2. Der Faktor weise m verschiedene Faktorstufen (Ebenen) auf, nach denen sich die Grundgesamtheiten unterscheiden.
3. Die Zufallsvariablen  $X_j$  ( $j = 1, \dots, m$ ) sind in den Grundgesamtheiten normalverteilt mit Erwartungswert  $\mu_j = E(X_j)$  und Varianz  $\sigma_j^2 : X_j \sim N(\mu_j; \sigma_j^2)$ .
4. Die Varianzen in den m Grundgesamtheiten sind gleich groß, wenn auch unbekannt (Varianzhomogenität, Homoskedastizität):  $\sigma_1^2 = \dots = \sigma_m^2 = \sigma^2$ . Diese Voraussetzung kann mit dem Levene-Test überprüft werden.
5. Gegeben sind m ( $m \geq 2$ ) unabhängige einfache Zufallsstichproben  $X_{j,1}, \dots, X_{j,n_j}$  ( $j = 1, \dots, m$ ) mit Stichprobenumfängen  $n_j$  ( $i = 1, \dots, n_j$ ) aus diesen Grundgesamtheiten. Sind die Stichprobenumfänge  $n_j$  für alle j gleich, spricht man vom balancierten Fall, andernfalls vom unbalancierten Fall.

Die Problemstellung lautet: Existieren signifikante Mittelwertunterschiede in den Grundgesamtheiten, die auf die Wirkung der Faktorstufen zurückzuführen sind?

Das allgemeine Modell der ANOVA ist wie folgt formuliert:

$$X_{j,i} = \mu_j + E_{j,i} = \mu + \tau_j + E_{j,i}, \quad j = 1, \dots, m; \quad i = 1, \dots, n_j \quad (4.23)$$

---

<sup>29</sup>Vgl. u.a. Hochstädtter, D., Kaiser, U. (1988), S. 4 ff.; Bosch, K. (1992), S. 497 ff.; Bortz, J. (1993), S. 226 ff.

#### 4. Parametervergleiche bei unabhängigen Stichproben

mit

$$\sum_{j=1}^m n_j \tau_j = 0. \quad (4.24)$$

Darin sind

-  $\mu_j = E(X_{j,i})$  der Erwartungswert der StichprobenvARIABLEN  $X_{j,i}$ ,

-  $\mu$  der (unbekannte) GesamtMittelwert gemäß

$$\mu = \frac{1}{n} \sum_{j=1}^m n_j \mu_j \quad (4.25)$$

mit

$$n = \sum_{j=1}^m n_j, \quad (4.26)$$

-  $\tau_j = \mu_j - \mu$  die Abweichung des Mittelwertes der j-ten Faktorstufe vom GesamtMittelwert, die durch die Wirkung des Faktors verursacht wird, d.h. der (unbekannte) feste Effekt<sup>30</sup> der j-ten Faktorstufe, und

-  $E_{j,i} = X_{j,i} - \mu_j = X_{j,i} - (\mu + \tau_j)$  der Beobachtungsfehler oder Versuchsfehler auf der j-ten Faktorstufe, der nicht aus der Wirkung des Faktors erklärt werden kann. Die  $E_{j,i}$  sind unabhängige,  $N(0; \sigma^2)$ -verteilte Zufallsvariablen, deren Werte nicht beobachtbar sind, sondern sich als Residuen nach der Modellschätzung ergeben.

(4.24) ist die sogenannte Reparametrisierungsbedingung, die eine eindeutige Schätzung der Parameter des Modells (auch bei ungleichen Stichprobenumfängen) ermöglicht, denn in (4.23) sind  $m + 1$  Parameter ( $\mu$  plus  $m$  Effekte  $\tau_j$ ) unbekannt und aus den Stichproben zu schätzen.

Zusammengefaßt beinhaltet dies, dass alle StichprobenvARIABLEN  $X_{j,i}$  wegen der 3. und 4. Voraussetzung  $N(\mu_j, \sigma^2)$ -verteilt sind ( $j = 1, \dots, m; i = 1, \dots, n_j$ ) und in den Grundgesamtheiten allenfalls Mittelwertunterschiede  $\mu_j = \mu + \tau_j$  existieren.

Daraus ergibt sich die Hypothesenformulierung:

$$H_0: \mu_1 = \dots = \mu_j = \dots = \mu_m = \mu$$

$H_1$ : mindestens zwei Mittelwerte sind voneinander verschieden,

oder äquivalent

---

<sup>30</sup>Hiervon ausgehend wird dieses Modell der Varianzanalyse auch als Modell mit festen Effekten oder Modell I bezeichnet. Des weiteren gibt es bei der Varianzanalyse noch das Modell mit zufälligen Effekten oder Modell II und das Modell mit gemischten Effekten oder Modell III. Für die Fragestellung im Kontext dieses Abschnittes trifft jedoch das Modell mit festen Effekten zu.

$$H_0 : \tau_1 = \dots = \tau_j = \dots = \tau_m = 0$$

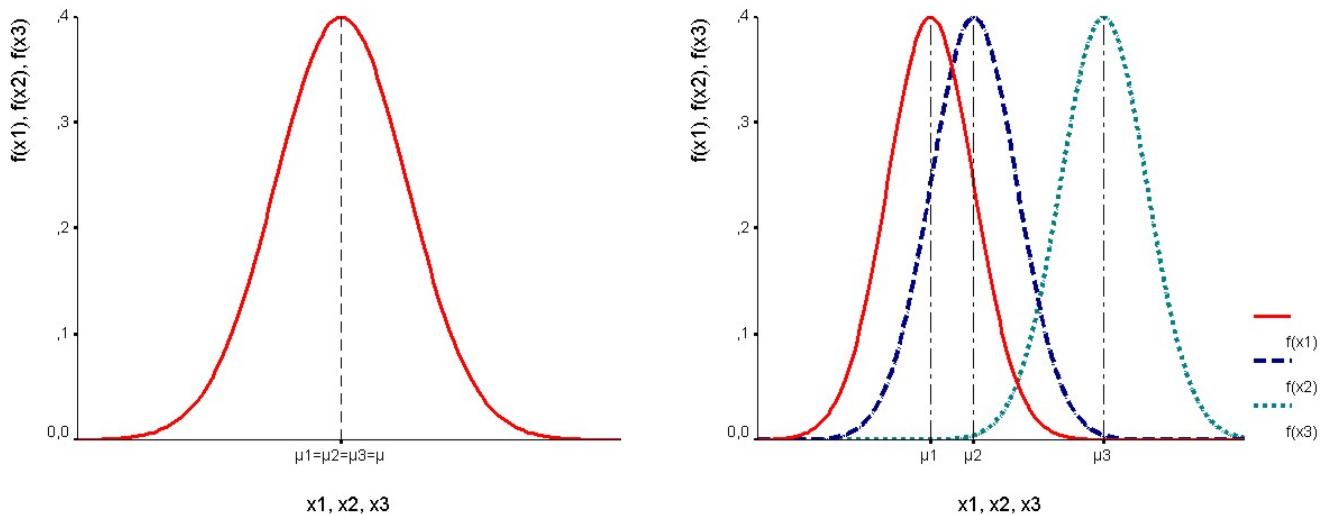
$H_1$ : nicht alle  $\tau_j$  sind gleich Null.

Abb. 4.10 zeigt für  $m = 3$  die Verteilung der Zufallsvariablen in den drei Grundgesamtheiten bei Gültigkeit der  $H_0$  und bei Gültigkeit der  $H_1$ .

Die Idee der Testdurchführung ist nun wie folgt:

Werden Faktorstufen (Grundgesamtheiten) zusammengefaßt, so dürfte sich aufgrund der vorausgesetzten Varianzhomogenität und bei Gültigkeit der Nullhypothese die Varianz der sich durch die Zusammenfassung ergebenden Grundgesamtheit nur unwesentlich, d.h. zufällig ändern. Bei deutlicher Vergrößerung dieser Varianz wird wohl eher die Alternativhypothese wahr sein.

Abbildung 4.10.: Verteilung der Zufallsvariablen  $X_1$ ,  $X_2$  und  $X_3$  in den drei Grundgesamtheiten  
 bei Gültigkeit der  $H_0$  bei Gültigkeit der  $H_1$



## Herleitung der Teststatistik:

Die m Stichproben werden in einer Tabelle zusammengefaßt.

Tabelle 4.1.: Ausgangstabelle der Varianzanalyse

Stichproben- element Nr.	Faktorstufen				
	1	...	j	...	m
1	$X_{1,1}$	...	$X_{j,1}$	...	$X_{m,1}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	$X_{1,i}$	...	$X_{j,i}$	...	$X_{m,i}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$n_j$	$X_{1,n_1}$	...	$X_{j,n_j}$	...	$X_{m,nm}$
Stichprobenmittelwert	$\bar{X}_1$	...	$\bar{X}_j$	...	$\bar{X}_m$

Die Stichprobenmittelwerte  $\bar{X}_j$  sind analog zu (4.4) und die Stichprobenvarianzen  $S_j^2$  analog zu

#### 4. Parametervergleiche bei unabhängigen Stichproben

(4.5) zu berechnen. Nach der Ziehung der  $m$  Stichproben stehen in dieser Tabelle anstelle der Zufallsvariablen die konkreten Realisationen  $x_{j,i}$  und  $\bar{x}_j$ .

Werden die  $m$  Stichproben zu einer einzigen zusammengefaßt, so gilt für deren Stichprobenmittelwert  $\bar{X}$ :

$$\bar{X} = \frac{1}{n} \sum_{j=1}^m n_j \bar{X}_j = \frac{1}{n} \sum_{j=1}^m \sum_{i=1}^{n_j} X_{j,i} \quad (4.27)$$

und für die Stichprobenvarianz  $S^2$

$$S^2 = \frac{1}{n-1} \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,i} - \bar{X})^2 \quad (4.28)$$

Die Summe der Abweichungsquadrate in der Gesamtstichprobe wird mit SQG bezeichnet:

$$SQG = (n-1)S^2 = \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,i} - \bar{X})^2 = \sum_{j=1}^m \sum_{i=1}^{n_j} X_{j,i}^2 - n\bar{X}^2 \quad (4.29)$$

SQG kann in zwei Komponenten zerlegt werden:

$$\begin{aligned} SQG &= \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,i} - \bar{X})^2 \\ &= \sum_{j=1}^m \sum_{i=1}^{n_j} [(X_{j,i} - \bar{X}_j) + (\bar{X}_j - \bar{X})]^2 \\ &= \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,i} - \bar{X}_j)^2 + 2 \sum_{j=1}^m (\bar{X}_j - \bar{X}) \sum_{i=1}^{n_j} (X_{j,i} - \bar{X}_j) + \sum_{j=1}^m \sum_{i=1}^{n_j} (\bar{X}_j - \bar{X})^2 \end{aligned}$$

Aufgrund der Nulleigenschaft des arithmetischen Mittels ist der zweite Ausdruck auf der rechten Seite gleich Null und es folgt:

$$\begin{aligned} SQG &= \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,i} - \bar{X}_j)^2 + \sum_{j=1}^m n_j (\bar{X}_j - \bar{X})^2 \\ &= \left( \sum_{j=1}^m \sum_{i=1}^{n_j} X_{j,i}^2 - \sum_{j=1}^m n_j \bar{X}_j^2 \right) + \left( \sum_{j=1}^m n_j \bar{X}_j^2 - n \bar{X}^2 \right) \\ &= SQI + SQZ \end{aligned} \quad (4.30)$$

SQZ ist die Summe der Abweichungsquadrate zwischen den Stichproben, d.h. die Summe der Abweichungsquadrate der jeweiligen Stichprobenmittelwerte  $\bar{X}_j$  vom Gesamtmittelwert  $\bar{X}$ . Sie ist auf den Faktor zurückzuführen und kann somit durch das Modell (4.23) erklärt werden.

SQI ist die Summe der Abweichungsquadrate innerhalb der Stichproben. Sie ist die sogenannte Restvariation, die durch das Modell, d.h. durch den Faktor, nicht erklärt werden kann.

Bezeichnet man mit  $S_I^2$  die Stichprobenvarianz innerhalb der Stichproben und mit  $S_Z^2$  die Stichprobenvarianz zwischen den Stichproben, so gilt (siehe Anhang D):

$$E(S^2) = E\left(\frac{SQG}{n-1}\right) = \sigma^2 + \frac{1}{n-1} \sum_{j=1}^m n_j(\mu_j - \mu)^2 \quad (4.31)$$

$$\begin{aligned} E(SQG) &= (n-1)\sigma^2 + \sum_{j=1}^m n_j(\mu_j - \mu)^2 \\ E(S_I^2) &= E\left(\frac{SQI}{n-m}\right) = \sigma^2 \end{aligned} \quad (4.32)$$

$$E(SQI) = (n-m)\sigma^2$$

$$E(S_Z^2) = E\left(\frac{SQZ}{m-1}\right) = \sigma^2 + \frac{1}{m-1} \sum_{j=1}^m n_j(\mu_j - \mu)^2 \quad (4.33)$$

$$E(SQZ) = (m-1)\sigma^2 + \sum_{j=1}^m n_j(\mu_j - \mu)^2$$

Unter  $H_0$  gilt somit:

$$E(SQG) = E(SQI) + E(SQZ)$$

$$(n-1)\sigma^2 = (n-m)\sigma^2 + (m-1)\sigma^2$$

$SQG/\sigma^2 = (n-1)S^2/\sigma^2$  ist unter  $H_0$   $\chi^2$ -verteilt mit  $f = n - 1$  Freiheitsgraden.

$SQI/\sigma^2 = (n-m)S_I^2/\sigma^2$  ist unter  $H_0$   $\chi^2$ -verteilt mit  $f = n - m$  Freiheitsgraden.

$SQZ/\sigma^2 = (m-1)S_Z^2/\sigma^2$  ist unter  $H_0$   $\chi^2$ -verteilt mit  $f = m - 1$  Freiheitsgraden.

Während  $S_I^2$  in jedem Fall, also auch wenn  $H_0$  nicht zutreffend ist, eine erwartungstreue Schätzfunktion für  $\sigma^2$  ist, sind  $S_Z^2$  und  $S^2$  nur im Falle der Gültigkeit der  $H_0$  erwartungstreue Schätzfunktionen für  $\sigma^2$ . Im Falle der Gültigkeit der  $H_1$  fallen die Schätzungen von  $S_Z^2$  und  $S^2$  zu groß aus. Sie weisen jeweils eine Verzerrung [2. Term auf der rechten Seite von (4.31) bzw. (4.33)] auf, die durch die Abweichungen  $\mu_j - \mu$  hervorgerufen werden. An dieser Stelle wird deutlich, dass mit der Varianzanalyse Mittelwertunterschiede geprüft werden.

$S_Z^2$  und  $S_I^2$  haben unter  $H_0$  den gleichen Erwartungswert und ihre Schätzwerte sind bis auf Zufallsschwankungen identisch. Unter  $H_1$  ist dagegen der Erwartungswert von  $S_Z^2$  und damit auch der Schätzwert (tendenziell) größer, was auf die Mittelwertunterschiede zurückzuführen ist. Es liegt also nahe, das Verhältnis dieser beiden Varianzen zu bilden, was zur Teststatistik

#### 4. Parametervergleiche bei unabhängigen Stichproben

der einfachen Varianzanalyse führt:

$$F = \frac{S_Z^2}{S_I^2} = \frac{\frac{SQZ}{m-1}}{\frac{SQI}{n-m}} = \frac{(n-m)SQZ}{(m-1)SQI} = \frac{MQZ}{MQI}, \quad (4.34)$$

wobei MQ die Abkürzung für mittlere Quadratsumme ist, die häufig bei Computerausgaben verwendet wird.

F ist das Verhältnis zweier chi-quadrat-verteilter Zufallsvariablen und daher bei Gültigkeit der Nullhypothese  $F(f_1; f_2)$ -verteilt mit  $f_1 = m - 1$  und  $f_2 = n - m$  Freiheitsgraden. Bei Richtigkeit der Nullhypothese wird F Werte nahe Eins annehmen, so dass große Werte von F zur Ablehnung der Nullhypothese führen.

$H_0$  wird auf dem vorgegebenen Signifikanzniveau abgelehnt, wenn  $F > F_{f_1; f_2; 1-\alpha}$  ist, wobei  $F_{f_1; f_2; 1-\alpha}$  das  $1 - \alpha$  Quantil der F-Verteilung mit  $f_1 = m - 1$  und  $f_2 = n - m$  Freiheitsgraden ist.

Die Ergebnisse der Testdurchführung werden i.a. in einer Varianztabelle der einfachen Varianzanalyse zusammengefaßt.

Tabelle 4.2.: Varianztabelle der einfachen Varianzanalyse

Source	Sum of Squares	Degrees of freedom (df)	Mean Square	F
Between Groups	SQZ	$f_1 = m - 1$	$S_Z^2 = \frac{SQZ}{m-1}$	$F = \frac{S_Z^2}{S_I^2}$
Within Groups	SQI	$f_2 = n - m$	$S_I^2 = \frac{SQI}{n-m}$	
Total	SQG	$f = n - 1$	-	-

Die Varianzanalyse ist im balancierten Fall relativ unempfindlich gegenüber der Verletzung der Voraussetzung der Normalverteilung, , d.h., sie führt auch dann noch zu brauchbaren Ergebnissen, wenn diese Voraussetzung nicht exakt erfüllt ist. Man sollte deshalb Stichproben gleichen Umfangs ziehen.

Beispiele zur ANOVA werden nach der Behandlung der multiplen Mittelwertvergleiche angegeben.

##### 4.2.2.4. Multiple Mittelwertvergleiche

Wird  $H_0$  bei der ANOVA verworfen, so sind signifikante Unterschiede zwischen den Mittelwerten der betrachteten Grundgesamtheiten gegeben. Im Fall  $m > 2$  bleibt dabei aber offen, zwischen welchen Grundgesamtheiten, d.h. auf welchen Faktorstufen, diese Unterschiede auftreten. Um dies herauszufinden, müssen weitere Tests angewandt werden. Bevor einige dieser

Tests behandelt werden, sind einige Vorbemerkungen notwendig.

### Zum Signifikanzniveau:

Man könnte in der Weise vorgehen, dass man mittels des Zweistichproben-t-Tests jeweils die Differenz zwischen den Mittelwerten zweier Grundgesamtheiten, d.h. alle möglichen Mittelwertpaare  $(j; j^*)$  mit  $j \neq j^*$  und  $j < j^*$ , prüft. Liegen  $m$  unabhängige Stichproben vor, so sind insgesamt

$$g = \binom{m}{2} = \frac{m(m-1)}{2}, \quad (4.35)$$

berechnet als die Anzahl der Kombinationen ohne Wiederholung, solcher Mittelwertvergleiche  $H_{0,k} : \mu_j = \mu_{j^*}$  gegen  $H_{1,k} : \mu_j \neq \mu_{j^*}$  ( $k = 1, \dots, g$ ) durchzuführen. Da jedoch nur ein Datensatz vorliegt, ergibt sich ein Problem bezüglich des Signifikanzniveaus.

Es bezeichne:

- $H_{0,G} : \mu_1 = \dots = \mu_m = \mu$  die globale Nullhypothese, die mittels der Varianzanalyse geprüft wird;
- $P("H_{0,G}" | H_{0,G}) = 1 - \alpha_G$  die Wahrscheinlichkeit, die globale Nullhypothese nicht abzulehnen<sup>31</sup>;
- $P("H_{1,G}" | H_{0,G}) = \alpha_G$  die Wahrscheinlichkeit, die globale Nullhypothese irrtümlich abzulehnen (Fehler 1. Art), d.h. die maximale Wahrscheinlichkeit, die Nullhypothese bei allen  $g$  paarweisen Mittelwertvergleichen mindestens einmal irrtümlich abzulehnen;  $H_{0,G}$  wird also abgelehnt, falls eine der Hypothesen  $H_{0,k}$ ,  $k = 1, \dots, g$ , verworfen wird;
- $P("H_{0,k}" | H_{0,k}) = 1 - \alpha$  die Wahrscheinlichkeit, die Nullhypothese  $H_{0,k} : \mu_j = \mu_{j^*}$  nicht abzulehnen, für alle  $k = 1, \dots, g$ ;
- $P("H_{1,k}" | H_{0,k}) = \alpha$  die Wahrscheinlichkeit, die Nullhypothese  $H_{0,k} : \mu_j = \mu_{j^*}$  irrtümlich abzulehnen (Fehler 1. Art), für alle  $k = 1, \dots, g$ .

---

<sup>31</sup>“ $H_{0,G}$ “ symbolisiert die Testentscheidung “Beibehaltung der globalen Nullhypothese“ und “ $H_{1,G}$ “ die Testentscheidung “Ablehnung der globalen Nullhypothese“ aufgrund der Stichproben. Analog gilt diese Schreibweise beim Test zweier Mittelwerte.

#### 4. Parametervergleiche bei unabhängigen Stichproben

Aufgrund der Unabhängigkeit der Stichproben gilt nun<sup>32</sup>

$$\begin{aligned}
 P("H_{0,G}"|H_{0,G}) = 1 - \alpha_G &= P\left(\bigcap_{k=1}^g ("H_{0,k}"|H_{0,k})\right) \\
 &= \prod_{k=1}^g P("H_{0,k}"|H_{0,k}) \\
 &= (1 - \alpha)^g = (1 - \alpha)^{\frac{m(m-1)}{2}}.
 \end{aligned} \tag{4.36}$$

Die Wahrscheinlichkeit eines Fehlers 1. Art für die globale Hypothese beträgt dann

$$\alpha_G = 1 - (1 - \alpha)^g = 1 - (1 - \alpha)^{\frac{m(m-1)}{2}}, \tag{4.37}$$

d.h., sie nimmt mit steigender Anzahl von Mittelwerten zu.

Beispiel: Bei  $m = 4$  und einem Signifikanzniveau von  $\alpha = 0,05$  für alle paarweisen Mittelwertvergleiche beträgt  $g = 6$ ,  $1 - \alpha_G = (1 - 0,05)^6 = 0,95^6 = 0,7351$  und  $\alpha_G = 0,2649$ ; bei  $m = 5$  ist  $g = 10$ ;  $1 - \alpha_G = (1 - 0,05)^{10} = 0,95^{10} = 0,5987$  und  $\alpha_G = 0,4013$ .

Will man umgekehrt ein bestimmtes  $\alpha_G$  sichern, dann müssen die paarweisen Mittelwertvergleiche mit einem Signifikanzniveau von

$$\alpha^* = 1 - \sqrt[g]{1 - \alpha_G} \tag{4.38}$$

getestet werden. Für z.B.  $m = 4$ ,  $g = 6$  und  $\alpha = 0,10$  wäre dies ein  $\alpha^* = 0,0174$ .

<sup>32</sup>Es sei

- $A = \{\text{berechtigte Beibehaltung von } H_0 \text{ bei einem paarweisen Mittelwertvergleich}\} = "H_{0,k}"|H_{0,k} \text{ mit } P(A) = 1 - \alpha;$
- $\bar{A} = \{\text{irrtümliche Ablehnung von } H_0 \text{ bei einem paarweisen Mittelwertvergleich}\} = "H_{1,k}"|H_{0,k} \text{ mit } P(\bar{A}) = \alpha;$
- $Y = \{\text{Anzahl der Mittelwertvergleiche, bei denen A eintritt}\}$  mit  $y = 0,1,\dots,g$ .

Dann folgt die Zufallsvariable  $Y$  einer Binomialverteilung  $B(g; 1 - \alpha)$ :

$$P(Y = y) = \binom{g}{y} (1 - \alpha)^y (1 - (1 - \alpha))^{g-y}.$$

Für  $y = g$  (bei allen Vergleichen wird  $H_0$  nicht irrtümlich abgelehnt) folgt:

$$P(Y = g) = \binom{g}{g} (1 - \alpha)^g (1 - (1 - \alpha))^0 = (1 - \alpha)^g.$$

Lineare Kontraste<sup>33</sup>:

Ein linearer Kontrast ist eine Linearkombination von m Mittelwerten  $\mu_j$  ( $j = 1, \dots, m$ )

$$\Lambda = c_1\mu_1 + \dots + c_j\mu_j + \dots + c_m\mu_m = \sum_{j=1}^m c_j\mu_j, \quad (4.39)$$

worin die Größen  $c_j$  ( $j = 1, \dots, m$ ) vorgegebene Konstanten sind, für die die zusätzliche Bedingung

$$\sum_{j=1}^m c_j = 0 \quad (4.40)$$

gilt Bei  $m = 2$  spricht man von einem einfachen linearen Kontrast. Sind die Mittelwerte  $\mu_j$  alle gleich, ist  $\Lambda = 0$  und ist  $\Lambda = 0$ , sind alle Mittelwerte gleich.

Sind  $\Lambda_1$  und  $\Lambda_2$  zwei lineare Kontraste und gilt

$$\sum_{j=1}^m c_{1j}c_{2j} = 0, \quad (4.41)$$

so heißen  $\Lambda_1$  und  $\Lambda_2$  orthogonale Kontraste.

Lineare Kontraste dienen der Signifikanzprüfung von zwei Gruppen von Mittelwerten in einem Schritt.

Beispiele:

Ist  $m = 2$  und die Hypothese lautet:  $H_0 : \mu_1 = \mu_2$ , so ist  $\Lambda = \mu_1 - \mu_2$  ein linearer Kontrast mit  $c_1 = 1$  und  $c_2 = -1$ .

Für  $m = 4$  lassen sich u.a. folgende lineare Kontraste formulieren:

- $\Lambda_1 = \mu_2 - \mu_4 = 0$  mit  $c_1 = c_3 = 0$ ,  $c_2 = 1$  und  $c_4 = -1$  für die Hypothese  $H_0 : \mu_2 = \mu_4$ ;
- $\Lambda_2 = \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$  mit  $c_1 = c_2 = 1$  und  $c_3 = c_4 = -1$  für die Hypothese  $H_0 : \mu_1 + \mu_2 = \mu_3 + \mu_4$  bzw.  $(\mu_1 + \mu_2) - (\mu_3 + \mu_4) = 0$ .

$$L = c_1\bar{x}_1 + \dots + c_j\bar{x}_j + \dots + c_m\bar{x}_m = \sum_{j=1}^m c_j\bar{x}_j \quad (4.42)$$

ist eine erwartungstreue Schätzung für  $\Lambda$  aufgrund der m unabhängigen Stichproben. Mittels L wird nun ein Konfidenzintervall für  $\Lambda$  angegeben, indem ein Vielfaches d der Standardabweichung von L zu L addiert bzw. subtrahiert wird, so dass die Wahrscheinlichkeit dafür, dass  $\Lambda$  von diesem Intervall überdeckt wird, gleich  $1 - \alpha$  ist:

$$[L - dS_L; L + dS_L] \quad (4.43)$$

---

<sup>33</sup>Vgl. u.a. Bortz, J. (1993), S. 240 ff.; Weber, E. (1972), S. 249 ff.

#### 4. Parametervergleiche bei unabhängigen Stichproben

zum Konfidenzniveau

$$P(L - dS_L \leq \Lambda \leq L + dS_L) = 1 - \alpha \quad (4.44)$$

Darin sind:

- die Standardabweichung von L

$$S_L = S_I \sqrt{\sum_{j=1}^m \left( \frac{c_j^2}{n_j} \right)} \quad (4.45)$$

- $S_I^2 = SQI/(n - m)$  die Varianz innerhalb der Stichproben mit SQI aus (4.30).

Schließt dieses Konfidenzintervall den Wert Null ein, so wird die Nullhypothese auf Gleichheit der Mittelwertgruppen nicht abgelehnt, andernfalls wird sie abgelehnt. Solche zu prüfenden linearen Kontraste sind vor der Ziehung der Stichproben zu formulieren, da sonst das Signifikanzniveau (Wahrscheinlichkeit für den Fehler 1. Art) statistisch nicht angebar ist.

Studentisierte Variationsbreite<sup>34</sup>:

$\bar{X}_1, \dots, \bar{X}_m$  seien unabhängige und  $N(\mu; \sigma^2/r)$ -verteilte Zufallsvariablen und  $S^2$  eine von den m Mittelwerten unabhängige und  $\sigma^2 \cdot \chi^2/f$ -verteilte Schätzfunktion für die gemeinsame Varianz  $\sigma^2$  der Mittelwerte, wobei r den gleichen Stichprobenumfang der m Stichproben (balanciertes Modell) und  $f = m \cdot r - 1$  die Anzahl der Freiheitsgrade von  $S^2$  bezeichnet. Dann heißt die Zufallsvariable

$$Q(m; f) = \frac{|\bar{X}_j - \bar{X}_{j^*}|}{S} \sqrt{r}, \quad j, j^* = 1, \dots, m \quad (4.46)$$

studentisierte Variationsbreite (durch  $S(\bar{X}) = Sr^{-1/2}$  normierte Spannweite). In der Variationsbreite  $\bar{X}_j - \bar{X}_{j^*}$  liegen eventuell mehrere Mittelwerte, die in die Nullhypothese auf Gleichheit eingeschlossen werden. Die Verteilung der studentisierten Variationsbreite hängt nur von der Anzahl der Mittelwerte m und der Anzahl der Freiheitsgrade f ab. Aus dieser Verteilung findet man für ein vorgegebenes  $\alpha$  Signifikanzpunkte  $q(\alpha; m; f)$ , mit denen Konfidenzintervalle bestimmt werden können, so dass

$$P(|\bar{X}_j - \bar{X}_{j^*}| - q(\alpha; m; f) \frac{S}{\sqrt{r}} \leq \mu_j - \mu_{j^*} \leq |\bar{X}_j - \bar{X}_{j^*}| + q(\alpha; m; f) \frac{S}{\sqrt{r}}) = 1 - \alpha \quad (4.47)$$

für alle  $j$  und  $j^*$  gilt.  $q(\alpha; m; f)$  gibt die maximale studentisierte Variationsbreite an, die eine Gruppe von m Mittelwerten aufweisen darf, um zum Signifikanzniveau  $\alpha$  als nicht signifikant verschieden bezeichnet zu werden.

---

<sup>34</sup>Vgl. u.a. Läuter, H., Pincus, R. (1989), S. 113 f.; Rasch, Enderlein, Herrendörfer (1973), S. 91; Bosch, K. (1992), S. 509

Schließt das Konfidenzintervall die Null ein, kann die Hypothese auf Gleichheit der betrachteten Mittelwerte nicht abgelehnt werden. Verwendet man z.B. die größte Differenz der Stichprobenmittelwerte und stellt fest, dass das Konfidenzintervall die Null einschließt, dann kann die Hypothese auf Gleichheit dieser Mittelwerte nicht abgelehnt werden. Gleichzeitig sind aber auch alle dazwischen liegenden Mittelwerte gleich, d.h., für sie kann die Hypothese auf Gleichheit ebenfalls nicht abgelehnt werden. In diesem Sinne werden also gleichzeitig mehrere Mittelwerte auf Gleichheit geprüft.

Werden nur zwei Mittelwerte miteinander verglichen, so ist  $q(\alpha; m; f) = t_{f;1-\alpha/2} \cdot 2^{1/2}$ , da nicht wie beim t-Test  $S_D$ , sondern  $S(\bar{X})$  im Nenner von (4.46) steht.<sup>35</sup>

Im unbalancierten Modell (d.h. bei ungleichen Stichprobenumfängen) kann  $Sr^{-1/2}$  durch eine Approximation von Kramer

$$S_{\bar{X}} = S \sqrt{\frac{1}{2} \left( \frac{1}{n_j} + \frac{1}{n_{j^*}} \right)} \quad (4.48)$$

ersetzt werden.

Diese Aussagen lassen sich auf lineare Kontraste verallgemeinern, für die dann gilt:

$$P \left( L - \frac{1}{2} q(\alpha; m; f) \frac{S}{\sqrt{r}} \sum_{j=1}^m |c_j| \leq \Lambda \leq L + \frac{1}{2} q(\alpha; m; f) \frac{S}{\sqrt{r}} \sum_{j=1}^m |c_j| \right) = 1 - \alpha, \quad (4.49)$$

wobei sich  $L$  gemäß (4.42) und  $\Lambda$  gemäß (4.39) ergibt und die Standardabweichung mit  $1/2$  multipliziert werden muss, da ein zweiseitiges Konfidenzintervall bestimmt wird. Die Verwendung linearer Kontraste ermöglicht, die Anzahl durchzuführender Tests von  $g = m(m - 1)/2$  auf  $m - 1$  zu verringern, da sich bei  $m$  Mittelwerten nur  $m - 1$  unabhängige lineare Kontraste ergeben.

Die Tests von Duncan, Newman-Keuls und Tukey verwenden die studentisierte Variationsbreite, jedoch mit unterschiedlichen kritischen Werten  $q(\alpha; m; f)$ .

Nachfolgend werden einige der unter SPSS verfügbaren Tests für multiple Mittelwertvergleiche ausführlicher behandelt.

## Least Significant Difference (LSD - Test)

Der LSD-Test ist äquivalent zu mehrfachen t-Tests zwischen allen Paaren von Mittelwerten und beruht auf der Teststatistik (4.20) des Zweistichproben-t-Tests. Diese kann bei Gültigkeit von  $H_0 : \mu_j = \mu_{j^*}$  auch in folgender Weise geschrieben werden:

$$T_{j,j^*} = \frac{\bar{X}_j - \bar{X}_{j^*}}{S \sqrt{\frac{1}{n_j} + \frac{1}{n_{j^*}}}} \quad (4.50)$$

<sup>35</sup>Bei gleichem Umfang  $r$  für beide Stichproben ergibt sich:

$$S \sqrt{\frac{1}{n_j} + \frac{1}{n_{j^*}}} = S \sqrt{\frac{1}{r} + \frac{1}{r}} = S \sqrt{\frac{2}{r}} = \frac{S}{\sqrt{r}} \sqrt{2} = S_{\bar{X}} \sqrt{2}.$$

#### 4. Parametervergleiche bei unabhängigen Stichproben

mit  $j \neq j^*$ ,  $j < j^*$ ,  $j = 1, \dots, m-1$  und  $j^* = 2, \dots, m$ .

Beim multiplen Mittelwertvergleich wird der Schätzer für die gemeinsame Varianz  $\sigma^2$  jedoch nicht nur aufgrund der beiden in den Test einbezogenen Stichproben, sondern auf Basis aller  $m$  Stichproben bestimmt:

$$S_I^2 = \frac{1}{n-m} \sum_{j=1}^m (n_j - 1) S_j^2, \quad (4.51)$$

was seine Berechtigung darin hat, dass alle Zufallsvariablen  $E_{j,i}$  im Modell (4.23) laut Voraussetzung die gleiche Varianz aufweisen und somit  $\sigma^2$  genauer geschätzt werden kann.  $S_I^2$  findet man als Varianz innerhalb der Stichproben in der Tabelle der einfachen Varianzanalyse.

Damit folgen alle  $T_{j,j^*}$  unter  $H_0$  jedoch einer t-Verteilung mit  $f = n - m$  Freiheitsgraden.  $H_0$  wird abgelehnt, wenn  $|T_{j,j^*}| > t_{n-m; 1-\alpha/2}$  ist, wobei  $t_{n-m; 1-\alpha/2}$  das  $(1 - \alpha/2)$ -Quantil der t-Verteilung ist. Es ist hier  $1 - \alpha/2$  zu verwenden, da es sich um einen zweiseitigen Test handelt.

Bei diesem multiplen t-Test wird keine Korrektur des Signifikanzniveaus für paarweise Vergleiche durchgeführt. Man erhält somit ein hohes Signifikanzniveau  $\alpha_G$  für die globale Nullhypothese  $H_{0,G}$  und u.U. einen Widerspruch zum Ergebnis des F-Tests bei der ANOVA, d.h., der globale Test führt bei  $\alpha$  nicht zur Ablehnung der Nullhypothese und der multiple Vergleich findet zum selben  $\alpha$  ein oder mehrere signifikante Differenzen bzw. umgekehrt. Man sollte deshalb nur diejenigen Mittelwertdifferenzen prüfen, von denen man von vorherein sachlogisch Signifikanz vermutet.<sup>36</sup>

Die Entscheidungsregel  $|T_{j,j^*}| > t_{n-m; 1-\alpha/2}$  kann auch wie folgt geschrieben werden:

$H_0$  wird abgelehnt, falls gilt:

$$\frac{|\bar{X}_j - \bar{X}_{j^*}|}{S_j \sqrt{\frac{1}{n_j} + \frac{1}{n_{j^*}}}} > t_{n-m; 1-\alpha/2} \quad (4.52)$$

$$|\bar{X}_j - \bar{X}_{j^*}| > t_{n-m; 1-\alpha/2} S_j \sqrt{\frac{1}{n_j} + \frac{1}{n_{j^*}}} \doteq LSD_t$$

↑ ↑ ↑

Der Ausdruck auf der rechten Seite der Ungleichung (4.52) wird als least significant difference bezeichnet und ist derjenige Grenzwert, ab dem eine Mittelwertdifferenz als signifikant erkannt wird.

Die drei Pfeile unterhalb von Formel (4.52) weisen nochmals auf die bereits dargelegten Charakteristika des LSD-Tests hin:

---

<sup>36</sup>Zum Problem des Testens von a priori und a posteriori Hypothesen ist das Studium von Bortz (1993), Abschnitt 7.3.4., S. 249 f. sehr zu empfehlen.

1. Verwendung des kritischen Wertes aus der t-Verteilung,
2. Bestimmung der Freiheitsgrade unter Einbeziehung aller m Stichproben,
3. kein korrigiertes Signifikanzniveau  $\alpha$ .

### Least Significant Difference - Bonferroni (modifizierter LSD-Test)<sup>37</sup>

Die least significant difference nach Bonferroni wird aufgrund der gleichen Überlegungen bestimmt. Es erfolgt aber eine Korrektur des Signifikanzniveaus, die auf die Ungleichung von Bonferroni zurückgeht:

$$P\left(\bigcup_{k=1}^g H_k\right) \leq \sum_{k=1}^g P(H_k). \quad (4.53)$$

Beinhaltet z.B.  $H_k$  das Ereignis, die Nullhypothese eines paarweisen Mittelwertvergleiches irrtümlich abzulehnen, so bedeutet  $\bigcup_k H_k$ , mindestens einen paarweisen Vergleich irrtümlich abzulehnen, d.h. die irrtümliche Ablehnung der globalen Nullhypothese. Ist  $P(\bigcup_k H_k) = \alpha_G$  und  $P(H_k) = \alpha_k$ , so geht (4.53) über in

$$\alpha_G \leq \sum_{k=1}^g \alpha_k. \quad (4.54)$$

Werden alle paarweisen Mittelwertvergleiche zum gleichen Signifikanzniveau durchgeführt und wählt man speziell  $\alpha^* = \alpha_G/g$ , so wird das Signifikanzniveau  $\alpha_G$  des globalen Tests eingehalten. Dies gilt auch für eine beliebige Untergruppe von paarweisen Mittelwertvergleichen.

Beispiel: Es sei  $m = 4$  und damit  $g = 6$ . Die Varianzanalyse (Prüfung der globalen Hypothese) wurde mit einem (globalen) Signifikanzniveau  $\alpha_G = 0,10$  durchgeführt und die Nullhypothese sei abgelehnt worden. Um die Mittelwertdifferenzen aufzufinden, wird der LSD-Bonferroni Test durchgeführt. Verwendet man für die Durchführung aller paarweisen Mittelwertvergleiche ein Signifikanzniveau von  $\alpha = 0,05$ , so wird insgesamt das globale Signifikanzniveau nicht eingehalten, da

$$\alpha_G = 0,10 \leq \sum_{k=1}^g \alpha_k = 6 \cdot 0,05 = 0,3$$

ist. Verwendet man für die Durchführung aller paarweisen Mittelwertvergleiche dagegen ein Signifikanzniveau von  $\alpha^* = \alpha_G/g = 0,1/6 = 0,01667$ , so wird insgesamt das globale Signifikanzniveau eingehalten.

---

<sup>37</sup>Vgl. u.a. Läuter, H., Pincus, R. (1989), S. 117; Griliches, Z. Intriligator, M.D. (1992), S. 834, 846 ff.

#### 4. Parametervergleiche bei unabhängigen Stichproben

Das Quantil der t-Verteilung ist somit für  $\alpha^*$  aufzusuchen. Die least significant difference ergibt sich als

$$|\bar{X}_j - \bar{X}_{j^*}| > t_{n-m; 1-\alpha^*/2} S_j \sqrt{\frac{1}{n_j} + \frac{1}{n_{j^*}}} \doteq LSD_{Bonferroni} \quad (4.55)$$

↑    ↑    ↑

Die drei Pfeile unterhalb der Formel (4.55) weisen nochmals auf die bereits dargelegten Charakteristika des LSD-Bonferroni Tests hin:

1. Verwendung des kritischen Wertes der t-Verteilung,
2. Bestimmung der Freiheitsgrade unter Einbeziehung aller m Stichproben,
3. korrigiertes Signifikanzniveau  $\alpha^*$ .

Da sich  $t_{n-m; 1-\alpha^*/2}$  oftmals in den Tabellen der t-Verteilung nicht auffinden lässt, sind spezielle Tabellen<sup>38</sup> oder z.B. die Näherungsformel

$$t_{f; 1-\alpha^*/2} \approx \lambda_\alpha \left( 1 - \frac{\lambda_\alpha^2 + 1}{4f} \right)^{-1} \quad (4.56)$$

mit  $\lambda_\alpha$  als das  $\alpha$ -Quantil der Standardnormalverteilung  $N(0;1)$  und f als Anzahl der Freiheitsgrade zu verwenden.

Da  $t_{n-m; 1-\alpha^*/2} > t_{n-m; 1-\alpha/2}$  und somit unter sonst gleichen Bedingungen  $LSD_{Bonferroni} > LSD$  ist, werden bei  $LSD_{Bonferroni}$  nicht so viele Differenzen als signifikant ausgewiesen als bei LSD.

$LSD_{Bonferroni}$  sollte nur bei nicht zu großem g angewandt werden, da sich sonst die Nichtablehnungsbereiche der t-Tests überschneiden und der Test sinnlos wird, d.h., die Trennschärfe des Tests wird immer geringer. Z.B. für  $g = 10$  und  $\alpha_G = 0,05$  ist  $\alpha^*/10 = 0,005$ .

#### Tukey-Test<sup>39</sup>

Der Tukey-Test kann nur im balancierten Fall (gleicher Umfang r aller Stichproben) angewandt werden. Er dient dem paarweisen Vergleich aller Mittelwerte und dem Auffinden von homogenen Untergruppen von Mittelwerten. Die Teststatistik des Tukey-Tests basiert auf der studentisierten Variationsbreite  $|\bar{X}_j - \bar{X}_{j^*}|$  für  $j, j^* = 1, \dots, m$  und  $j < j^*$  (die auch als linearer Kontrast mit  $c_j = 1, c_{j^*} = -1$  und alle anderen  $m-2$  c-Koeffizienten gleich 0 aufgefaßt werden kann).

<sup>38</sup>Siehe u.a. Rasch, Enderlein, Herrendörfer (1973), S. 374 f.; Dunn, O.J. (1961); Miller, R.G.jr. (1966); Bailey, J.R. (1977)

<sup>39</sup>Vgl. u.a. Hartung, Elpelt, Klösener (1993), S. 616; Bosch, K. (1992), S. 509 f.; Lohse, Ludwig, Röhr (1982), S. 272 ff.; Hochstädter, Kaiser (1988), S. 38 ff.; Läuter, Pincus (1989), S. 113 f.; Weber, E. (1972), S. 262 ff.

Die studentisierte Variationsbreite  $|\bar{X}_j - \bar{X}_{j^*}|$  ist mit der kritischen Variationsbreite  $HSD_\alpha$  zu vergleichen (HSD - honestly significant difference):

$$HSD_\alpha = q(\alpha; m; f) \cdot \frac{S_I}{\sqrt{r}} \quad (4.57)$$

Darin sind

- $q(\alpha; m; f)$  der Signifikanzpunkt der studentisierten Variationsbreite für ein vorgegebenes Signifikanzniveau  $\alpha$ ; sie liegen für ausgewählte  $m$  und  $f$  und für  $\alpha = 0,05$  bzw.  $\alpha = 0,01$  tabelliert vor<sup>40</sup>,
- $m$  die Anzahl der insgesamt gegebenen Stichprobenmittelwerte,
- $f = n - m$  die Anzahl der Freiheitsgrade von  $S_I$  mit  $n = r \cdot m$ ,
- $S_I$  die Quadratwurzel aus  $S_I^2$  der einfachen Varianzanalyse (Varianz innerhalb der Stichproben).

Der Tukey-Test verwendet also nur einen kritischen Wert  $q(\alpha; m; f)$  für alle paarweisen Mittelwertvergleiche und erreicht dadurch eine Beschränkung von  $\alpha_G$  auf das Niveau des Varianzanalyse-Tests, denn aus (4.47) für alle  $j$  und  $j^*$  folgt wegen (4.46):

$$P \left( \max_{j,j^*} \frac{|\bar{X}_j - \bar{X}_{j^*}|}{S_I} \sqrt{r} \leq q(\alpha; m; f) \right) = P \left( \max_{j,j^*} Q(m; f) \leq q(\alpha; m; f) \right) = 1 - \alpha.$$

Testentscheidung:

Wenn

$$|\bar{X}_j - \bar{X}_{j^*}| > q(\alpha; m; f) \cdot \frac{S_I}{\sqrt{r}} = HSD_\alpha \quad (4.58)$$

↑      ↑      ↑

ist, wird die Nullhypothese, dass die durch die Variationsbreite  $|\bar{X}_j - \bar{X}_{j^*}|$  überspannten  $p$  aus  $m$  Mittelwerten gleich sind, abgelehnt. Die drei Pfeile unterhalb von Formel (4.58) weisen nochmals auf die bereits dargelegten Charakteristika des Tukey-Tests hin:

1. Verwendung eines kritischen Wertes der Verteilung der studentisierten Variationsbreite,
2. Verwendung von  $m$ , der Anzahl der insgesamt gegebenen Stichprobenmittelwerte,
3. gleicher Stichprobenumfang  $r$  für alle Stichproben (balancierter Fall).

---

<sup>40</sup>Siehe u.a. Läuter, H. Pincus, R. (1989), S. 366 f.; Bosch, K. (1992), S. 785 f.; Rasch, Enderlein, Herrendörfer (1973), S. 349 ff.; Hartung, Elpelt, Klösener (1993), S. 902 f.

#### 4. Parametervergleiche bei unabhängigen Stichproben

Der Tukey-Test ist trennschärfster als der LSD-Bonferroni-Test, wenn eine große Anzahl von paarweisen Vergleichen durchzuführen ist.

Der Tukey-Test lässt sich auch auf allgemeinere lineare Kontraste anwenden. Beziehen sich diese jedoch auf weniger als m Mittelwerte, wird das Konfidenzintervall wegen der Verwendung von  $q(\alpha; m; f)$  unnötig ausgedehnt und damit die Güte des Tests verringert. Der unter SPSS auch angegebene Tukey b-Test arbeitet mit einem kritischen Signifikanzpunkt, der sich als Durchschnitt von  $q(\alpha; m; f)$  aus HSD und  $q(\alpha; p; f)$  des Newman-Keuls-Tests ergibt.

#### Student-Newman-Keuls-Test<sup>41</sup>

Der Student-Newman-Keuls-Test kann nur im balancierten Fall (gleicher Umfang r aller Stichproben) angewandt werden. Er dient dem Auffinden von homogenen Untergruppen von Mittelwerten. Ausgegangen wird von den der Größe nach aufsteigend geordneten Mittelwerten  $\bar{X}_{(1)}, \dots, \bar{X}_{(m)}$  der m Stichproben.

Die Teststatistik dieses Tests basiert auf der studentisierten Variationsbreite  $|\bar{X}_j - \bar{X}_{j^*}|$  einer Gruppe von p aus den m Mittelwerten, die mit der kritischen Differenz  $R_p$  zu vergleichen ist:

$$R_p = q(\alpha; p; f) \cdot \frac{S_I}{\sqrt{r}}. \quad (4.59)$$

Darin sind

- $q(\alpha; p; f)$  der Signifikanzpunkt der studentisierten Variationsbreite für ein vorgegebenes Signifikanzniveau  $\alpha$ ;
- p die Anzahl der Mittelwerte in der zu prüfenden Gruppe: p = oberer Index - unterer Index + 1 (Wird z.B. die Differenz  $\bar{X}_{(8)} - \bar{X}_{(3)}$  geprüft, so enthält diese Gruppe 6 Mittelwerte:  $p = 8 - 3 + 1 = 6$ ),
- f = n - m die Anzahl der Freiheitsgrade von  $S_I$  mit  $n = r \cdot m$ ,
- $S_I$  die Quadratwurzel aus  $S_I^2$  der Tabelle der einfachen Varianzanalyse (Varianz innerhalb der Stichproben).

Testentscheidung:

Wenn

$$|\bar{X}_j - \bar{X}_{j^*}| > q(\alpha; p; f) \cdot \frac{S_I}{\sqrt{r}} = R_p \quad (4.60)$$

↑      ↑      ↑

gilt, ist die Differenz signifikant zum Niveau  $\alpha$ .

Die drei Pfeile unterhalb von Formel (4.60) weisen nochmals auf die bereits dargelegten Charakteristika des Student-Newman-Keuls-Tests hin:

---

<sup>41</sup>Vgl. u.a. Hochstädter, Kaiser (1988), S. 41 ff; Läuter, Pincus (1989), S. 115 f.; Weber, E. (1972), S. 257 ff.

1. Verwendung kritischer Werte der Verteilung der studentisierten Variationsbreite,
2. Verwendung von p, der Anzahl der von der Variationsbreite überspannten Mittelwerte,
3. gleicher Stichprobenumfang r für alle Stichproben (balancierter Fall).

Testprozedur:

Die Testprozedur erfolgt schrittweise. Im 1. Schritt wird  $|\bar{x}_{(m)} - \bar{x}_{(1)}|$ , d.h. die Variationsbreite über alle Mittelwerte, getestet. In diesem Schritt ist  $p = m$ . Ist diese Differenz nicht signifikant, kann  $H_0 : \mu_1 = \dots = \mu_m$  nicht abgelehnt werden. Das Testverfahren wird beendet, denn es sind dann auch alle Differenzen von Mittelwerten, die zwischen  $\bar{x}_{(m)}$  und  $\bar{x}_{(1)}$  liegen, nicht signifikant.

Wird im 1. Schritt  $H_0$  abgelehnt, werden im 2. Schritt die Variationsbreiten  $|\bar{x}_{(m-1)} - \bar{x}_{(1)}|$  und  $|\bar{x}_{(m)} - \bar{x}_{(2)}|$  geprüft. Diese überspannen jedoch jetzt  $p = m - 1$  Mittelwerte, so dass in (4.59) ein anderer Signifikanzpunkt  $q(\alpha; p; f)$  gewählt wird. Sind beide Differenzen nicht signifikant, kann  $H_0 : \mu_1 = \dots = \mu_{m-1}$  bzw.  $H_0 : \mu_2 = \dots = \mu_m$  nicht abgelehnt werden und das Verfahren bricht ab.

Andernfalls werden für den Fall, bei dem sich eine signifikante Differenz ergeben hat, im 3. Schritt die Variationsbreiten über  $p = m - 2$  Mittelwerte geprüft; usw.

Im  $(m-1)$ -ten Schritt (falls dieser erreicht wird) werden schließlich  $p = 2$  Mittelwerte verglichen.

Die in den einzelnen Schritten zu verwendenden Signifikanzpunkte  $q(\alpha; p; f)$  sind:

Schritt	Anzahl der überspannten Mittelwerte p	$q(\alpha; p; f)$
1	m	$q(\alpha; m; f)$
2	$m - 1$	$q(\alpha; m - 1; f)$
3	$m - 2$	$q(\alpha; m - 2; f)$
:	:	:
$m - 1$	2	$q(\alpha; 2; f)$

Durch diese Anpassung von  $q(\alpha; p; f)$  wird ein gleiches Signifikanzniveau bei der globalen Hypothese und bei den multiplen Tests gesichert. Der Student-Newman-Keuls-Test ist deshalb trennschärfer als der Tukey-Test, d.h., bei Ungleichheit der Mittelwerte in der Grundgesamtheit werden mehr Teilhypthesen abgelehnt.

## Duncan Test<sup>42</sup>

Der Duncan-Test unterscheidet sich nur durch die Wahl des Signifikanzniveaus vom Student-Newman-Keuls-Test. Er dient ebenfalls dem Auffinden von homogenen Untergruppen von Mittelwerten.

Die Teststatistik des Duncan-Tests, die studentisierte Variationsbreite  $|\bar{X}_j - \bar{X}_{j^*}|$ , ist mit der

<sup>42</sup>Vgl. u.a. Hochstädter, Kaiser (1988), S. 44 f.; Weber, E. (1972), S. 256.

#### 4. Parametervergleiche bei unabhängigen Stichproben

kritischen Variationsbreite  $R_D$  (D für Duncan)

$$R_D = q(\alpha_p; p; f) \cdot \frac{S_I}{\sqrt{r}} \quad (4.61)$$

zu vergleichen. Darin ist  $q_D(\alpha_p; p; f)$  der Signifikanzpunkt der studentisierten Variationsbreite für ein vorgegebenes Signifikanzniveau  $\alpha$  nach Duncan mit

$$\alpha_p = 1 - (1 - \alpha)^{p-1}. \quad (4.62)$$

Testentscheidung: Wenn

$$|\bar{X}_j - \bar{X}_{j^*}| > q(\alpha_p; p; f) \cdot \frac{S_I}{\sqrt{r}} = R_p \quad (4.63)$$

↑ ↑ ↑      ↑

gilt, ist die Differenz signifikant zum Niveau  $\alpha_p$ .

Die vier Pfeile unterhalb von Formel (4.63) weisen auf die Charakteristika des Duncan-Tests hin:

1. Verwendung kritischer Werte der Verteilung der studentisierten Variationsbreite,
2. Verwendung von  $p$ , der Anzahl der von der Variationsbreite überspannten Mittelwerte,
3. an  $p$  angepaßtes Signifikanzniveau  $\alpha_p$ ,
4. gleicher Stichprobenumfang  $r$  für alle Stichproben (balancierter Fall).

Die Signifikanzwerte  $\alpha_p$  liegen für ausgewählte  $p$  und  $f$  für  $\alpha = 0,05$  und  $\alpha = 0,01$  tabelliert vor<sup>43</sup>. Erst bei  $p = 2$  ist  $\alpha_p = \alpha$ . Sind z.B.  $m = 5$  und  $\alpha = 0,05$ , so ergeben sich in den einzelnen Schritten des Testverfahrens nachstehende  $\alpha_p$ :

Schritt	Anzahl der überspannten Mittelwerte $p$	$\alpha_p$
1	5	0,1855
2	4	0,1426
3	3	0,0975
4	2	0,05

#### Scheffé-Test<sup>44</sup>

Der Scheffé-Test kann auch auf den unbalancierten Fall sowie zur Prüfung von a posteriori-Hypothesen über lineare Kontraste verwendet werden. Er dient dem paarweisen Vergleich aller Mittelwerte und dem Auffinden von homogenen Untergruppen von Mittelwerten.

<sup>43</sup>Siehe u.a. Harter, H.L. (1960)

<sup>44</sup>Vgl. u.a. Lohse, Ludwig, Röhr (1982), S. 264 f.; Hochstädtter (1988), S. 47 ff; Läuter, Pincus (1989), S. 112 f.; Bortz, J. (1993), S. 250 ff.

Der Scheffé-Test prüft die Nullhypothese  $H_{0,u} : \Lambda_u = 0$  für alle voneinander linear unabhängigen linearen Kontraste (gekennzeichnet mit u). Dabei ist  $\Lambda_u$  gemäß (4.39) definiert. Aus den Stichproben heraus erhält man die erwartungstreue Schätzung L gemäß (4.42). Die Varianz  $S_L^2$  von L ergibt sich als Quadrat von (4.45):

$$S_L^2 = S_I^2 \sum_{j=1}^m \left( \frac{c_j^2}{n_j} \right). \quad (4.64)$$

Die Zufallsvariable  $S_L^2$  ist chi-quadrat-verteilt mit  $n - m$  Freiheitsgraden und liefert eine erwartungstreue Schätzung für  $\sigma^2$ , da  $S_I^2$  stets ein erwartungstreuer Schätzer für  $\sigma^2$  ist, wie bereits bei der Varianzanalyse gezeigt wurde.

Gleichfalls gilt auch  $Var(L) = E[(L - E(L))^2] = E(L^2) - [E(L)]^2$ . Nur unter Gültigkeit von  $H_0$  ist  $E(L) = 0$  und somit

$$Var(L) = E(L^2) = \left[ \left( \sum_{j=1}^m c_j \bar{X}_j \right)^2 \right] = S_{L|H_0}^2 \quad (4.65)$$

Diese Größe ist chi-quadrat-verteilt mit  $m - 1$  Freiheitsgraden. Wenn die Koeffizienten  $c_j$  festgelegt sind, gibt es nur  $m - 1$  unabhängige  $\bar{X}_j$ , um  $L = 0$  unter  $H_0$  zu garantieren. Der Quotient aus (4.65), dividiert durch  $m - 1$ , und (4.64)

$$F = \frac{\frac{1}{m-1} \left( \sum_{j=1}^m c_j \bar{X}_j \right)^2}{S_I^2 \sum_{j=1}^m \frac{c_j^2}{n_j}} \quad (4.66)$$

folgt unter  $H_0$  einer F-Verteilung mit  $f_1 = m - 1$  und  $f_2 = n - m$  Freiheitsgraden.

Ist  $F > F_{m-1;n-m;1-\alpha}$ , so kann  $H_0$  auf Gleichheit der im linearen Kontrast enthaltenen Mittelwerte nicht beibehalten werden, wobei  $F_{m-1;n-m;1-\alpha}$  das Quantil der Ordnung  $1 - \alpha$  der F-Verteilung mit  $f_1 = m - 1$  und  $f_2 = n - m$  Freiheitsgraden und  $\alpha$  das vorgegebene Signifikanzniveau sind.

Diese Quantil  $F_{m-1;n-m;1-\alpha}$  stimmt jedoch mit dem F-Quantil des Tests zur Prüfung der globalen Hypothese bei der Varianzanalyse überein, so dass der Scheffé-Test das Niveau  $\alpha$  einhält. Nach Einsetzen von (4.66) in  $F > F_{m-1;n-m;1-\alpha}$  und einigen Umformungen kann die Testentscheidung auch in folgender Weise getroffen werden:

Wenn

$$\sum_{j=1}^m c_j \bar{X}_j > S_I \sqrt{(m-1) F_{m-1;n-m;1-\alpha} \sum_{j=1}^m \frac{c_j^2}{n_j}} \quad (4.67)$$

gilt, ist  $H_0$  zum Niveau  $\alpha$  abzulehnen.

#### 4. Parametervergleiche bei unabhängigen Stichproben

Für den Vergleich zweier Mittelwerte wird  $H_0 : \mu_j = \mu_{j^*}$  geprüft. Es sind  $c_j = 1$ ,  $c_{j^*} = -1$  und alle anderen  $m-2$  c-Koeffizienten gleich Null. (4.67) geht in diesem Fall über in

$$|\bar{X}_j - \bar{X}_{j^*}| > S_I \sqrt{(m-1)F_{m-1;n-m;1-\alpha}} \sqrt{\frac{1}{n_j} + \frac{1}{n_{j^*}}}. \quad (4.68)$$

↑↑                      ↑                      ↑

Die Pfeile unterhalb von Formel (4.68) weisen auf die Charakteristika des Scheffé-Tests hin:

1. Verwendung kritischer Werte aus der F-Verteilung,
2. Bestimmung der Freiheitsgrade unter Einbeziehung aller m Stichproben,
3. Verwendung des Signifikanzniveaus  $\alpha$  der Varianzanalyse,
4. ungleicher Stichprobenumfang möglich.

Für diese einfachen linearen Kontraste werden im allgemeinen die Konfidenzintervalle breiter als z.B. beim Tukey-Test, so dass nicht notwendig alle signifikanten Mittelwertdifferenzen gefunden werden.

Der Scheffé-Test ist relativ robust gegenüber Verletzungen der Voraussetzungen (Normalverteilung, gleiche Varianz). Er ist jedoch ein konservativer Test, d.h., entscheidet tendenziell eher zugunsten der  $H_0$ .

### Einfache Varianzanalyse und multiple Mittelwertvergleiche unter SPSS

Unter SPSS sind die einfache Varianzanalyse und die multiplen Mittelwertvergleiche über

■ Analyze

■ Compare Means

■ One-Way ANOVA...

abrufbar.

Abbildung 4.11.: Dialogfeld „One-Way ANOVA“

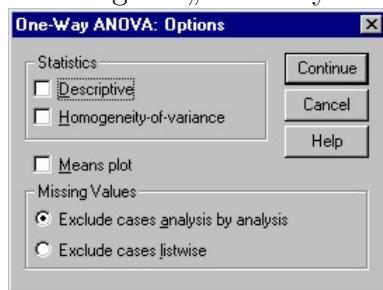


In diesem Dialogfeld ist zunächst die zu analysierende Variable in das Feld „Dependent List:“ und die Faktorvariable in das Feld „Factor:“ zu bringen.

Über die Schaltfläche „Options...“ lassen sich im Dialogfeld „One-Way ANOVA: Options“ (Abb. 4.12) zusätzliche Informationen zur Varianzanalyse anfordern:

- deskriptive Statistiken (Descriptive) für die einzelnen Stichproben (Faktorstufen) und für alle Stichproben zusammen:
  - Anzahl der gültigen Fälle ( $n_j$  bzw.  $n$ ),
  - Mittelwert ( $\bar{x}_j$  bzw.  $\bar{x}$ ),
  - Standardabweichung ( $s_j$  bzw.  $s$ ),
  - Standardfehler des Mittelwertes ( $s_j(\bar{x}_j)$  bzw.  $s(\bar{x})$ ),
  - 95%-Konfidenzintervalle für den Mittelwert,
  - kleinsten und größten beobachteten Wert;
- Test auf Varianzhomogenität nach Levene (Homogeneity-of-variance), der eigentlich vor der Varianzanalyse (wie im Abschnitt 4.2.1 behandelt), jedoch spätestens hier durchgeführt werden sollte;
- ein Plot der Mittelwerte über den Faktorausprägungen (Means plot).

Abbildung 4.12.: Dialogfeld „One-Way ANOVA: Options“



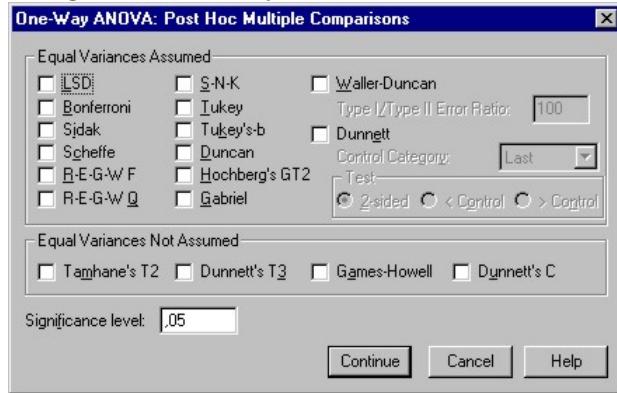
Im Verfahrensablauf sollte nunmehr erst einmal die Varianzanalyse durchgeführt werden, bevor weitere Entscheidungen für zusätzliche Tests getroffen werden. Dies hat seine Berechtigung im oben ausgeführten Sinne: Wenn die globale Hypothese  $H_{0,G} : \mu_1 = \dots = \mu_m$  durch den Test der Varianzanalyse nicht abgelehnt werden kann, sind weitergehende Einzelvergleiche nicht notwendig.

Wird  $H_{0,G}$  abgelehnt, ist sicherlich von Interesse, zwischen welchen Faktorstufen Mittelwertunterschiede existieren. Um die multiplen Mittelwertvergleiche durchführen zu lassen, wird wiederum die One-Way ANOVA aufgerufen und im zugehörigen Dialogfeld (Abb. 4.11) die

#### 4. Parametervergleiche bei unabhängigen Stichproben

Schaltfläche „Post hoc...“ betätigt.

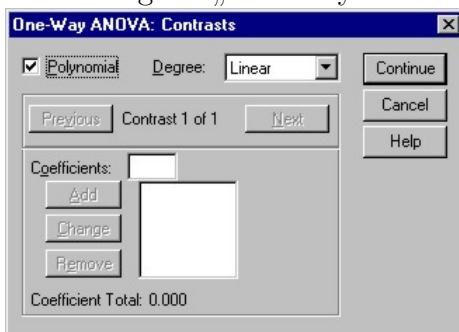
Abbildung 4.13.: Dialogfeld „One-Way ANOVA: Post Hoc Multiple Comparisons“



Im Dialogfeld „One-Way ANOVA: Post Hoc Multiple Comparisons“ stehen multiple Mittelwerttests zur Auswahl, wobei R-E-G-W für Ryan-Einot-Gabriel-Welsch F-Test, R-E-G-W Q für Ryan-Einot-Gabriel-Welsch RangeTest, S-N-K für Student-Newman-Keuls steht. Von diesen ist entsprechend den gegebenen Voraussetzungen der geeignete Test (bzw. Tests) auszuwählen. Des weiteren kann das Signifikanzniveau (Significance level) gegenüber der Voreinstellung von 0,05 verändert werden.

Wurden vor der Stichprobenerhebung (d.h. *a priori*) Hypothesen über lineare Kontraste formuliert, so können diese ebenfalls mittels der One-Way ANOVA getestet werden. Dazu ist im Dialogfeld „One-Way ANOVA“ (Abb. 4.11) die Schaltfläche „Contrasts...“ zu betätigen. Es erscheint das Dialogfeld „One-Way ANOVA: Contrasts“, in dem die unter der Hypothese festgelegten Koeffizienten  $c_j$  eingegeben werden.

Abbildung 4.14.: Dialogfeld „One-Way ANOVA: Contrasts“



Nach der Eingabe eines Koeffizienten in das Feld „Coefficients:“ darf die Betätigung der Schaltfläche „Add“ nicht vergessen werden.

Zu beachten ist, dass die Reihenfolge der einzugebenden Koeffizienten der aufsteigenden Reihenfolge der Faktorvariablenwerte entsprechen muss. Außerdem muss aufgrund der Definition linearer Kontraste (Formeln (4.39) und (4.40)) die Summe der  $c_j$  gleich Null sein, was unter dem Eingabefeld für die Koeffizienten kontrolliert werden kann.

Weist z.B. die Faktorvariable 5 Faktorstufen ( $m = 5$ ) auf und wurde a priori die Nullhypothese formuliert, dass der Mittelwert der 1. und 2. Faktorstufe gleich dem Mittelwert der 4. und 5. Faktorstufe ist, so entspricht dies dem zu prüfenden linearen Kontrast

$$H_0 : \Lambda = 0,5\mu_1 + 0,5\mu_2 + 0\mu_3 + (-0,5)\mu_4 + (-0,5)\mu_5 = 0,$$

und es sind nacheinander die Koeffizienten 0,5, 0,5, 0, -0,5, -0,5 einzugeben.

Wurden a priori mehrere lineare orthogonale (d.h. unabhängige) Kontraste festgelegt, so können sie in diesem Dialogfeld gleichzeitig eingegeben werden. Dazu ist nach Eingabe der Koeffizienten des ersten linearen Kontrastes die innere Schaltfläche „Next“ (hinter dem Text „Contrast 1 of 1“) zu betätigen und anschließend die Koeffizienten für den zweiten linearen Kontrast einzugeben, usw. Insgesamt können bis zu 10 lineare Kontraste mit bis zu 50 Koeffizienten spezifiziert werden.

Der Output enthält dann für jeden spezifizierten Kontrast zusätzlich eine Liste der Koeffizienten sowie den Wert des Kontrastes, den Standardfehler des Kontrastes, die t-Statistik, die Anzahl der Freiheitsgrade für die t-Statistik und das zweiseitige Signifikanzniveau von t sowohl für eine gepoolte Varianzschätzung (Assume equal variances) als auch eine separate Varianzschätzung (Does not assume equal).

### • Beispiel 4.3:

Für die Durchführung der Varianzanalyse und der multiplen Mittelwertvergleiche soll zunächst ein sehr einfaches, nachrechenbares Beispiel<sup>45</sup> gezeigt werden. Es soll geprüft werden, ob spezielle Zusätze (Faktor) zum Benzin einen Effekt auf die Klopffestigkeit hat. Es werden 5 unterschiedliche Zusätze (Faktorstufen,  $m = 5$ ,  $j = 1, \dots, 5$ ) verwendet, wobei immer nur ein Zusatz beigefügt wird.

Die Hypothesenformulierung ist wie folgt:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1 : \mu_j \neq \mu_{j^*} \text{ für mindestens ein Mittelwertpaar } j, j^* \ (j, j^* = 1, \dots, 5, j \neq j^*).$$

Die Nullhypothese soll auf einem Signifikanzniveau von  $\alpha = 0,05$  getestet werden. Bei einem Versuch werden aus jeder der 5 Grundgesamtheiten 4 Proben entnommen, d.h., es werden 5 Stichproben mit  $n_1 = \dots = n_5 = r = 4$  ( $i = 1, \dots, 4$ ) gezogen, deren Unabhängigkeit durch die Versuchsanlage gegeben ist. Es ergaben sich nachstehende Beobachtungswerte.

---

<sup>45</sup>Dieses Beispiel wurden entnommen aus: Kockelkorn, U. (1995)

#### 4. Parametervergleiche bei unabhängigen Stichproben

Tabelle 4.3.: Beobachtungswerte zum Beispiel 4.3

i	Zusatz 1	Zusatz 2	Zusatz 3	Zusatz 4	Zusatz 5
1	91,7	91,7	92,4	91,8	93,1
2	91,2	91,9	91,2	92,2	92,9
3	90,9	90,9	91,6	92,0	92,4
4	90,6	90,9	91,0	91,4	92,4

Diese Beobachtungswerte sind in der Datei benzin.sav in der Variablen kf (Klopffestigkeit) enthalten. Die Faktorstufe (Zusatz zum Benzin) zu jedem Beobachtungswert enthält die Variable gr (Gruppe).

Es wird die One-Way ANOVA aufgerufen, die Variable kf nach „Dependent List:“ und die Variable gr nach „Factor:“ gebracht. Unter Options werden alle drei Möglichkeiten gewählt, denn auf jeden Fall muss der Levene-Test auf Varianzhomogenität durchgeführt werden.

**SPSS-Output 4.3-1:** One-Way ANOVA, Descriptives, Homogeneity-of-variance für Beispiel 4.3

#### Descriptives

KF

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	4	91,100	,469	,235	90,354	91,846	90,6	91,7
2	4	91,350	,526	,263	90,513	92,187	90,9	91,9
3	4	91,550	,619	,310	90,565	92,535	91,0	92,4
4	4	91,850	,342	,171	91,306	92,394	91,4	92,2
5	4	92,700	,356	,178	92,134	93,266	92,4	93,1
Total	20	91,710	,706	,158	91,379	92,041	90,6	93,1

#### Test of Homogeneity of Variance

KF

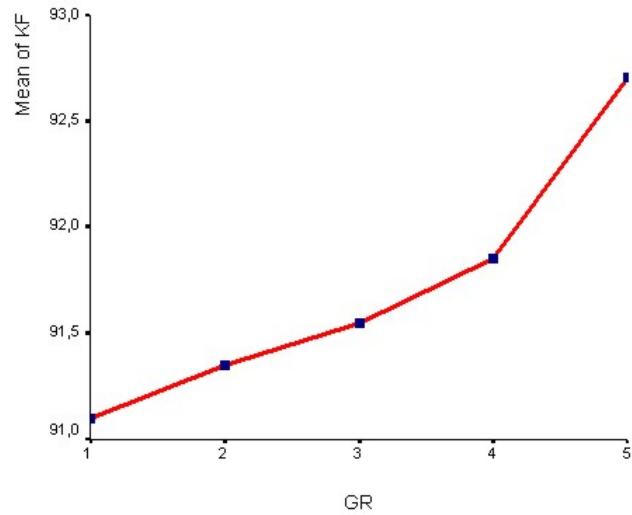
Levene Statistic	df1	df2	Sig.
,738	4	15	,580

#### ANOVA

KF

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6,108	4	1,527	6,797	,002
Within Groups	3,370	15	,225		
Total	9,478	19			

Abbildung 4.15.: Means Plot für Beispiel 4.3



Der Levene-Test zeigt keine signifikanten Unterschiede der Varianzen zum 5%-Niveau an, d.h. diese Voraussetzung der Varianzanalyse ist erfüllt.

Die ANOVA-Übersicht entspricht Tabelle 4.2. Für  $\alpha = 0,05$ ,  $f_1 = 5 - 1 = 4$  und  $f_2 = 20 - 5 = 15$  findet man in der Tabelle der F-Verteilung  $F_{4;15;g} = 3,056$ . Wegen  $F = 6,7967 > F_{4;15;g} = 3,056$  bzw. aufgrund von  $\text{Sig} = 0,002 < \alpha = 0,05$  wird die Nullhypothese basierend auf den 5 Stichproben mit gleichem Stichprobenumfang  $n_j = 4$  auf dem Signifikanzniveau von 5% abgelehnt. Die Zusätze führen zu signifikanten Unterschieden bei der Klopffestigkeit des Benzins.

Nun geht das Interesse dahin, festzustellen, zwischen welchen Zusätzen solche Unterschiede in der mittleren Klopffestigkeit bestehen. Im folgenden werden die oben behandelten multiplen Mittelwerttests nacheinander aufgerufen und der dafür relevante Teil des Outputs im SPSS Viewer wiedergegeben und kommentiert.

### LSD-Test (Geringste signifikante Differenz)

Die hier anzuwendende Testentscheidung ist (4.52).

Mit

- $t_{15;0,975} = 2,131$  aus der Tabelle der t-Verteilung,
- $S_I^2 = 0,2247$  als Mean Squares Within Groups in der Varianztabelle,
- $n_j = n_{j*} = r = 4$

folgt für die rechte Seite von (4.52):  $LSD_t = 0,7143$ .

#### 4. Parametervergleiche bei unabhängigen Stichproben

##### SPSS-Output 4.3-2: LSD-Test für Beispiel 4.3

###### Multiple Comparisons

Dependent Variable: KF

LSD

I (GR)	J (GR)	Mean Difference (I - J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-,250	,335	,467	-,964	,464
	3	-,450	,335	,199	-1,164	,264
	4	-,750*	,335	,041	-1,464	-3,562E-02
	5	-1,600*	,335	,000	-2,314	-,886
2	1	,250	,335	,467	-,464	,964
	3	-,200	,335	,560	-,914	,514
	4	-,500	,335	,156	-1,214	,214
	5	-1,350*	,335	,001	-2,064	-,636
3	1	,450	,335	,199	-,264	1,164
	2	,200	,335	,560	-,514	,914
	4	-,300	,335	,385	-,414	1,014
	5	-1,150*	,335	,004	-1,864	-,436
4	1	,750*	,335	,041	3,562E-02	1,464
	2	,500	,335	,156	-,214	1,214
	3	,300	,335	,385	-,414	1,014
	5	-,850*	,335	,023	-1,564	-,136
5	1	1,600*	,335	,000	,886	2,314
	2	1,350*	,335	,001	,636	2,064
	3	1,150*	,335	,004	,436	1,864
	4	,850*	,335	,023	,136	1,564

\*. The mean difference is significant at the .05 level.

Jede Mittelwertdifferenz ist mit diesem Wert zu vergleichen. So ist z.B.

$$|\bar{x}_1 - \bar{x}_2| = |91,1 - 91,35| = 0,25 < 0,7143 \text{ keine signifikante Differenz};$$

$$|\bar{x}_1 - \bar{x}_5| = |91,1 - 92,7| = 1,6 > 0,7143 \text{ dagegen eine signifikante Differenz.}$$

Alle signifikanten Differenzen sind im Output mit einem \* gekennzeichnet. Die folgende Tabelle soll eine bessere Übersicht geben.

Tabelle 4.4.: Signifikante Mittelwertdifferenzen beim LSD-Test

Gruppe	1	2	3	4	5
1					
2					
3					
4		★			
5	★	★	★	★	★

So ist  $\mu_5$  signifikant verschieden von den Mittelwerten aller anderen Grundgesamtheiten und  $\mu_4$  signifikant verschieden von  $\mu_1$ .

Da jeder paarweise Mittelwertvergleich mit einem Signifikanzniveau von  $\alpha = 0,05$  durchgeführt wurde, wird wegen (4.37) das globale Signifikanzniveau nicht eingehalten. Nun bietet SPSS jedoch die Möglichkeit, das Signifikanzniveau im Dialogfeld „One-Way ANOVA: Post Hoc Multiple Comparisons“ zu verändern. Wählt man nach (4.38)  $\alpha^* = 1 - 0,95^{1/10} = 0,005$ , so wird das globale Signifikanzniveau eingehalten, führt dann aber nur noch zu signifikanten Differenzen der Mittelwertpaare (1;5), (2;5) und (3;5).

### LSD-Bonferroni

Die Testentscheidung erfolgt hier nach (4.55) mit  $g = 10$ ,  $\alpha^* = \alpha/g = 0,05/10 = 0,005$ ,  $t_{15;0,9975} = 3,29$  (aus Rausch, Enderlein, Herrendörfer (1973), S. 375),  $S_I^2 = 0,2247$  und  $n_j = n_{j^*} = r = 4$ . Für die rechte Seite der Ungleichung ergibt sich somit rund 1,1028, mit dem alle Mittelwertdifferenzen zu vergleichen sind.

#### SPSS-Output 4.3-3: LSD-Bonferroni-Test für Beispiel 4.3

##### Multiple Comparisons

Dependent Variable: KF

Bonferroni

I (GR)	J (GR)	Mean Difference (I - J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-,250	,335	1,000	-1,351	,851
	3	-,450	,335	1,000	-1,551	,651
	4	-,750	,335	,408	-1,851	,351
	5	-1,600*	,335	,002	-2,701	-,499
2	1	,250	,335	1,000	-,851	1,351
	3	-,200	,335	1,000	-1,301	,901
	4	-,500	,335	1,000	-1,601	,601
	5	-1,350*	,335	,011	-2,451	-,249
3	1	,450	,335	1,000	-,651	1,551
	2	,200	,335	1,000	-,901	1,301
	4	-,300	,335	1,000	-1,401	,801
	5	-1,150*	,335	,037	-2,251	-4,865E-02
4	1	,750	,335	,408	-,351	1,851
	2	,500	,335	1,000	-,601	1,601
	3	,300	,335	1,000	-,801	1,401
	5	-,850	,335	,228	-1,951	,251
5	1	1,600*	,335	,002	,499	2,701
	2	1,350*	,335	,011	,249	2,451
	3	1,150*	,335	,037	4,865E-02	2,251
	4	,850	,335	,228	-,251	1,951

\*. The mean difference is significant at the .05 level.

#### 4. Parametervergleiche bei unabhängigen Stichproben

Die folgende Tabelle soll wieder eine bessere Übersicht geben.

Tabelle 4.5.: Signifikante Mittelwertdifferenzen beim LSD-Bonferroni-Test

Gruppe	1	2	3	4	5
1					
2					
3					
4					
5	★	★	★		

Im Vergleich zum LSD-Test ist  $t_{n-m;1-\alpha^*/2} > t_{n-m;1-\alpha/2}$  ( $1,102 > 0,7143$ ), d.h., die beim Bonferroni-Test konstruierten Konfidenzintervalle sind breiter. Dadurch werden die Differenzen  $\mu_4 - \mu_5$  und  $\mu_1 - \mu_4$  beim Bonferroni-Test nicht als signifikant ausgewiesen.

Im Vergleich zum LSD-Test mit angepaßtem Signifikanzniveau  $\alpha^* = 0,005$  ergeben sich jedoch genau die gleichen signifikanten Mittelwertdifferenzen, da dann  $t_{n-m;1-\alpha^*/2}$  für beide Tests gleich sind.

### Tukey-Test

Die Mittelwertdifferenzen sind alle mit dem kritischen Wert  $HSD_\alpha$  gemäß (4.57) zu vergleichen und die Testentscheidung erfolgt nach (4.58).  $S_I$  und  $r$  sind wie vorher;  $q(\alpha; m; f) = q(0,05; 5; 15)$  ergibt sich aus der Tabelle der Signifikanzpunkte der studentisierten Variationsbreite zu 4,37 und damit  $HSD_{0,05} = 1,0357$ .

Da der Tukey-Test dem paarweisen Vergleich aller Mittelwerte und dem Auffinden von homogenen Untergruppen von Mittelwerten dient, sind im SPSS-Output die Tabelle Multiple Comparison und die Tabelle Homogenous Subsets enthalten.

Auch beim Tukey-Test werden die Differenzen der Mittelwertpaare (1;5), (2;5) und (3;5) als signifikant angegeben. Damit gibt es zwei Unterteilungen der Gruppen, in denen die Mittelwerte nicht signifikant verschieden sind (homogenous Subsets): die Gruppen 1, 2, 3, 4 einerseits und die Gruppen 4, 5 andererseits.

**SPSS-Output 4.3-4:** Tukey-Test für Beispiel 4.3**Multiple Comparisons**

Dependent Variable: KF

Tukey HSD

I (GR)	J (GR)	Mean Difference (I - J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-,250	,335	,942	-1,285	,785
	3	-,450	,335	,671	-1,485	,585
	4	-,750	,335	,219	-1,785	,285
	5	-1,600*	,335	,002	-2,635	-,565
2	1	,250	,335	,942	-,785	1,285
	3	-,200	,335	,973	-1,235	,835
	4	-,500	,335	,583	-1,535	,535
	5	-1,350*	,335	,008	-2,385	-,315
3	1	,450	,335	,671	-,585	1,485
	2	,200	,335	,973	-,835	1,235
	4	-,300	,335	,894	-1,335	,735
	5	-1,150*	,335	,026	-2,185	-,115
4	1	,750	,335	,219	-,285	1,785
	2	,500	,335	,583	-,535	1,535
	3	,300	,335	,894	-,735	1,335
	5	-,850	,335	,134	-1,885	,185
5	1	1,600*	,335	,002	,565	2,635
	2	1,350*	,335	,008	,315	2,385
	3	1,150*	,335	,026	,115	2,185
	4	,850	,335	,134	-,185	1,885

\*. The mean difference is significant at the .05 level.

**Homogenous Subsets****KF**Tukey HSD<sup>a</sup>

GR	N	Subset for alpha = .05	
		1	2
1	4	91,100	
2	4	91,350	
3	4	91,550	
4	4	91,850	91,850
5	4		92,700
Sig.		,219	,134

Means for groups in homogenous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.000.

Auf die Wiedergabe des Outputs des Tukey-B-Tests soll hier verzichtet werden. Er resultiert in den gleichen homogenous Subsets.

#### 4. Parametervergleiche bei unabhängigen Stichproben

### Student-Newman-Keuls-Test

Er dient dem Auffinden von homogenen Untergruppen von Mittelwerten, weshalb im Output nur die Tabelle der homogenous Subsets enthalten ist.

**SPSS-Output 4.3-5:** Student-Newman-Keuls-Test für Beispiel 4.3

#### Homogenous Subsets KF

Student-Newman-Keuls<sup>a</sup>

GR	N	Subset for alpha = .05	
		1	2
1	4	91,100	
2	4	91,350	
3	4	91,550	
4	4	91,850	
5	4		92,700
Sig.		,158	1,000

Means for groups in homogenous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.000.

Der kritische Wert einer Mittelwertdifferenz ergibt sich gemäß (4.59).  $q(\alpha; p; f)$  mit  $f = n - m = 15$  entnimmt man für das entsprechende  $p$  und  $\alpha = 0,05$  Tabellen der Signifikanzpunkte der studentisierten Variationsbreite. Die Testentscheidung erfolgt nach (4.60). Man erhält:

1. Schritt:

$q(0,05; 5; 15) = 4,37$  und somit für  $R_5 = 1,0357$ ;

$\bar{x}_5 - \bar{x}_1 = 92,7 - 91,1 = 1,6 > 1,0357$ , die Differenz ist signifikant.

2. Schritt:

$q(0,05; 4; 15) = 4,07$  und somit für  $R_4 = 0,9646$ ;

$\bar{x}_4 - \bar{x}_1 = 91,85 - 91,1 = 0,75 < 0,9646$ , diese Differenz und damit alle Differenzen von dazwischenliegenden Mittelwerten sind nicht signifikant. Sie ergeben homogenous Subset 1.

$\bar{x}_5 - \bar{x}_2 = 92,7 - 91,35 = 1,35 > 0,9646$ , diese Differenz ist signifikant.

3. Schritt:

$q(0,05; 3; 15) = 3,67$  und somit für  $R_3 = 0,8698$ ;

$\bar{x}_5 - \bar{x}_3 = 92,7 - 91,55 = 1,15 > 0,8698$ , diese Differenz ist signifikant.

4. Schritt:

$q(0,05; 2; 15) = 3,02$  und somit für  $R_2 = 0,7158$ ;

$\bar{x}_5 - \bar{x}_4 = 92,7 - 91,85 = 0,85 > 0,7158$ , diese Differenz ist signifikant.  $\bar{x}_5$  bildet homogenous Subset 2.  $\mu_5$  ist signifikant verschieden von den Mittelwerten aller anderen Grundgesamtheiten. Durch die Anpassung von  $q(\alpha; p; f)$  ist der Student-Newman-Keuls-Test trennschärfer als der Tukey-Test. Er weist auch die Differenz  $\mu_5 - \mu_4$  als signifikant aus.

## Duncan-Test

Er dient ebenfalls dem Auffinden von homogenen Untergruppen von Mittelwerten, so dass im Output nur eine Tabelle der homogenous Subsets enthalten ist.

**SPSS-Output 4.3-6:** Duncan-Test für Beispiel 4.3

### Homogenous Subsets KF

Duncan<sup>a</sup>

GR	N	Subset for alpha = .05	
		1	2
1	4	91,100	
2	4	91,350	
3	4	91,550	
4	4	91,850	
5	4		92,700
Sig.		,056	1,000

Means for groups in homogenous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.000.

Die Testentscheidung erfolgt nach (4.63). Beim Duncan-Test muss in jedem Schritt eine Mittelwertdifferenz größer sein als  $R_D$  gemäß (4.61), um als signifikant ausgewiesen zu werden. Dabei sind wiederum  $S_I^2 = 0,2247$  und  $r = 4$ .  $q_D(\alpha; p; f)$  mit  $f = n - m = 15$  entnimmt man für das entsprechende  $p$  und  $\alpha = 0,05$  Tabellen der Signifikanzpunkte für den Duncan-Test.

Der Duncan-Test läuft in folgenden Schritten ab:

1. Schritt:

$$\alpha_p = 0,1855; q_D(0,1855; 5; 15) = 3,31; R_D = 0,7845$$

$\bar{x}_5 - \bar{x}_1 = 92,7 - 91,1 = 1,6 > 0,7845$ , die Differenz ist signifikant.

2. Schritt:

$$\alpha_p = 0,1426; q_D(0,1426; 4; 15) = 3,26; R_D = 0,7727$$

$\bar{x}_4 - \bar{x}_1 = 91,85 - 91,1 = 0,75 < 0,7727$ , diese Differenz und damit alle dazwischenliegenden Mittelwerte sind nicht signifikant. Sie ergeben homogenous Subset 1.

$\bar{x}_5 - \bar{x}_2 = 92,7 - 91,35 = 1,35 > 0,7727$ , diese Differenz ist signifikant.

3. Schritt:

$$\alpha_p = 0,0975; q_D(0,0975; 3; 15) = 3,16; R_D = 0,749$$

$\bar{x}_5 - \bar{x}_3 = 92,7 - 91,55 = 1,15 > 0,749$ , diese Differenz ist signifikant.

4. Schritt:

$$\alpha_p = 0,05; q_D(0,05; 2; 15) = 3,02; R_D = 0,7158$$

$\bar{x}_5 - \bar{x}_4 = 92,7 - 91,85 = 0,85 > 0,7158$ , diese Differenz ist signifikant.  $\bar{x}_5$  bildet homogenous Subset 2.  $\mu_5$  ist signifikant verschieden von den Mittelwerten aller anderen Grundgesamtheiten.

In den ersten drei Schritten ist  $q(\alpha_p; p; f)$  und damit  $R_D$  des Duncan-Tests kleiner als  $q(\alpha; p; f)$  und damit  $R_p$  des Student-Newman-Keuls-Test, wodurch der Duncan-Test trennschärfert ist,

#### 4. Parametervergleiche bei unabhängigen Stichproben

denn existierende Mittelwertunterschiede werden eher erkannt.

### Scheffé-Test

Er dient dem paarweisen Vergleich aller Mittelwerte und dem Auffinden von homogenen Untergruppen von Mittelwerten. Die Testentscheidungen sind hier nach (4.68) zu treffen. In der Tabelle der Verteilungsfunktion der F-Verteilung findet man für  $f_1 = 4$ ,  $f_2 = 15$  und  $\alpha = 0,05$  den Wert  $F_{4;15;0,95} = 3,06$ . Mit  $S_I^2 = 0,2247$  und  $n_j = r = 4$  ergibt sich als kritischer Wert auf der rechten Seite von (4.68) 1,1727, mit dem alle Mittelwertdifferenzen zu vergleichen sind. Die sich ergebenden Konfidenzintervalle sind breiter als bei den vorangegangenen Tests, so dass nicht so viele Mittelwertdifferenzen als signifikant ausgewiesen werden, d.h. nur die Mittelwertpaare (1;5) und (2;5). Die zwei Unterteilungen der Gruppen, in denen die Mittelwerte nicht signifikant verschieden sind (homogenous Subsets), sind demzufolge die Gruppen 1, 2, 3, 4 einerseits und die Gruppen 3, 4, 5 andererseits.

**SPSS-Output 4.3-7:** Scheffé-Test für Beispiel 4.3

#### Multiple Comparisons

Dependent Variable: KF

Scheffé

I (GR)	J (GR)	Mean Difference (I - J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-,250	,335	,965	-1,422	,922
	3	-,450	,335	,770	-1,622	,722
	4	-,750	,335	,332	-1,922	,422
	5	-1,600*	,335	,005	-2,772	-,428
2	1	,250	,335	,965	-,922	1,422
	3	-,200	,335	,985	-1,372	,972
	4	-,500	,335	,698	-1,672	,672
	5	-1,350*	,335	,020	-2,522	-,178
3	1	,450	,335	,770	-,722	1,622
	2	,200	,335	,985	-,972	1,372
	4	-,300	,335	,934	-1,472	,872
	5	-1,150	,335	,056	-2,322	2,174E-02
4	1	,750	,335	,332	-,422	1,922
	2	,500	,335	,698	-,672	1,672
	3	,300	,335	,934	-,872	1,472
	5	-,850	,335	,224	-2,022	,322
5	1	1,600*	,335	,005	,428	2,772
	2	1,350*	,335	,020	,178	2,522
	3	1,150	,335	,056	-2,174E-02	2,322
	4	,850	,335	,224	-,322	2,022

\*. The mean difference is significant at the .05 level.

**Homogenous Subsets****KF**Scheffé<sup>a</sup>

GR	N	Subset for alpha = .05	
		1	2
1	4	91,100	
2	4	91,350	
3	4	91,550	91,550
4	4	91,850	91,850
5	4		92,700
Sig.		,332	,056

Means for groups in homogenous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.000.

• Beispiel 4.4:

Es soll die Frage überprüft werden, ob sich das mittlere persönliche Einkommen der Frauen (Variable X) aufgrund unterschiedlichen Schulabschlusses (Faktor) von mittlerer Reife, Fachhochschulreife und Abitur (Faktorstufen) signifikant unterscheidet. Mit der Datei allbus.sav liegen unabhängige Stichproben für jede Faktorstufe vor, wobei die zu analysierende Variable einkomp1, die Faktorvariable schule und die Faktorstufen die Werte 3 bis 5 der Variablen schule sind.

Um nur die relevanten Fälle einzubeziehen, muss bei der Variablen sex eine Einschränkung auf sex = 2 (Frau) und bei der Variablen schule eine Beschränkung auf den Wert 3 (mittlere Reife) oder 4 (Fachhochschulreife) oder 5 (Abitur) erfolgen, was entsprechend Kapitel 1 über Fälle auswählen (siehe Abb. 1.5) erfolgt. Dazu wird im Dialogfeld „Select Cases: If“ (siehe Abb. 1.6) in das freie Textfeld eingegeben:

(sex = 2) &amp; (schule=3|schule=4|schule=5).

Zunächst<sup>46</sup> wird die Voraussetzung der Normalverteilung in jeder der 3 Stichproben auf einem Signifikanzniveau von  $\alpha = 0,05$  überprüft (siehe Abschnitt 3.3.1, Kolmogorov-Smirnov-Test mit Lillefors Korrektur).

<sup>46</sup>Die nachfolgenden Tests müßten auf der Basis verschiedener Stichproben erfolgen, um das jeweilige Signifikanzniveau einzuhalten. Da das an dieser Stelle praktisch nicht durchführbar ist, werden die Tests mit der Stichprobe der Datei allbus.sav demonstriert.

#### 4. Parametervergleiche bei unabhängigen Stichproben

**SPSS-Output 4.4-1:** Test auf Normalverteilung des monatlichen persönlichen Nettoeinkommens nach den Gruppen mittlere Reife, Fachhochschulreife und Abitur

**Test of Normality**

	Allgemeiner Schulabschluß	Kolmogorov-Smirnov <sup>a</sup>		
		Statistic	df	Sig.
Monatl. Nettoeinkommen in DM	Mittlere Reife	,112	81	,013
	Fachhochschulreife	,313	8	,021
	Abitur	,098	29	,200*

\* This is lower bound of the true significance

a. Lillefors Significance Correction.

Die Nullhypothese kann nur in der Gruppe Abitur nicht abgelehnt werden. Diese Voraussetzung für die ANOVA ist somit insgesamt nicht erfüllt.

Beim Aufruf der One-Way ANOVA (siehe Abb. 4.11) wird die Variable einkomp1 in das Feld „Dependent List“ und die Variable schule in das Feld „Factor“ gebracht und unter Options (siehe Abb. 4.12) Descriptive und Homogeneity-of-variance ausgewählt.

**SPSS-Output 4.4-2:** One-Way ANOVA des persönlichen Nettoeinkommens nach den Gruppen mittlere Reife, Fachhochschule und Abitur

#### Descriptives

Monatliches Nettoeinkommen in DM

	N	Mean	Std. Deviation	Std. Error	95% Confidence Intervall for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Mittlere Reife	81	1358,42	777,68	86,41	1186,46	1530,38	240	5000
Fachhochschulreife	8	965,00	602,57	213,04	461,24	1468,76	400	2000
Abitur	29	1806,90	835,03	155,06	1489,27	2124,52	500	3700
Total	118	1441,97	809,64	74,53	1294,36	1589,58	240	5000

#### Test of Homogeneity of Variance

Monatl. Nettoeinkommen in DM

Levene Statistic	df1	df2	Sig.
,514	2	115	,599

#### ANOVA

Monatl. Nettoeinkommen in DM

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6247405,446	2	3123702,723	5,099	,008
Within Groups	70447788,418	115	612589,465		
Total	76695193,864	117			

Die Nullhypothese auf Gleichheit der Varianzen in den drei Grundgesamtheiten wird auf dem 5%-Niveau nach dem Levene-Test nicht abgelehnt. Diese Voraussetzung der ANOVA ist somit erfüllt.

Die Nullhypothese auf Gleichheit von  $\mu_3$ ,  $\mu_4$  und  $\mu_5$  (Indizierung erfolgt entsprechend der Faktorwerte) wird aufgrund der Stichproben auf dem vorgegebenen Signifikanzniveau verworfen. Dieses Ergebnis ist insoweit mit Vorsicht zu verwenden, da die Voraussetzung der Normalverteilung nicht erfüllt war. Bei nochmaligem Aufruf der One-Way ANOVA wird unter „Post hoc...“ der Scheffé-Test angefordert, da er

- relativ robust gegenüber Verletzungen der Voraussetzungen der Normalverteilung ist und
- unterschiedliche Stichprobenumfänge in den drei Gruppen vorliegen.

**SPSS-Output 4.3-7:** Scheffé-Test für das monatliche persönliche Nettoeinkommen nach den Gruppen mittlere Reife, Fachhochschulreife und Abitur

#### Multiple Comparisons

Dependent Variable: Monatl. Nettoeinkommen in DM

Scheffé

I (Allgemeiner Schulabschluss)	J (Allgemeiner Schulabschluss)	Mean Difference (I - J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Mittlere Reife	Fachhochschulreife	393,42	290,06	,402	-325,93	1112,77
	Abitur	-448,48*	169,37	,033	-868,51	-28,44
Fachhochschulreife	Mittlere Reife	-393,42	290,06	,402	-1112,77	325,93
	Abitur	-841,90*	312,57	,030	-1617,05	-66,74
Abitur	Mittlere Reife	448,48*	169,37	,033	28,44	868,51
	Fachhochschulreife	841,90*	312,57	,030	66,74	1617,05

\*. The mean difference is significant at the .05 level.

#### Homogenous Subsets

Monatl. Nettoeinkommen in DM

Scheffé<sup>a,b</sup>

Allgemeiner Schulabschluss	N	Subset for alpha = .05	
		1	2
Fachhochschulreife	8	965,00	
Mittlere Reife	81	1358,42	1358,42
Abitur	29		1806,90
Sig.		,335	,243

Means for groups in homogenous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 17,459.

b. The group are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Das mittlere monatliche Nettoeinkommen beim Schulabschluss Abitur  $\mu_5$  erweist sich als signifikant verschieden vom mittleren monatlichen Nettoeinkommen beim Schulabschluss Mittlere Reife  $\mu_3$  und vom mittleren persönlichen Nettoeinkommen beim Schulabschluss Fachhochschulreife  $\mu_4$ , während zwischen  $\mu_3$  und  $\mu_4$  kein signifikanter Unterschied aufgrund dieser Stichproben festgestellt werden konnte.

Bezüglich der weiteren unter SPSS verfügbaren Tests sei auf die SPSS-Handbücher verwiesen.

4. Parametervergleiche bei unabhängigen Stichproben

# Anhang A

## Testentscheidung unter Verwendung statistischer Software

Die allgemeine Vorgehensweise bei der Hypothesenprüfung ist wie folgt:

- Formulierung der Null- und Alternativhypothese
- Konstruktion der Teststatistik  $V$  als Funktion der Stichprobenvariablen  $V = V(X_1, \dots, X_n)$ .  
Die Verteilungsfunktion der Teststatistik  $V$  muss unter der Annahme, dass die Nullhypothese wahr ist, zumindest approximativ bekannt sein.
- Vorgabe eines Signifikanzniveaus  $\alpha$  ( $0 < \alpha < 1$ )
- Bestimmung des Ablehnungsbereiches der Nullhypothese im Wertebereich der Teststatistik  $V$ , so dass die Wahrscheinlichkeit dafür, dass  $V$  Werte aus diesem Ablehnungsbereich annimmt, nicht größer als  $\alpha$  ausfällt, falls die Nullhypothese wahr ist. Der Wert, der den Nichtablehnungsbereich der Nullhypothese vom Ablehnungsbereich trennt, heißt kritischer Wert und kann für das vorgegebene Signifikanzniveau  $\alpha$  aus der Verteilungsfunktion von  $V$  bestimmt werden. Beim zweiseitigen Test erhält man zwei kritische Werte und zwei Teilsegmente des Ablehnungsbereiches der Nullhypothese.
- Ziehen einer Zufallsstichprobe vom Umfang  $n$  und Berechnung der Realisation  $v$  (Testwert) der Teststatistik  $V$
- Testentscheidung: Die Nullhypothese wird auf dem vorgegebenen Signifikanzniveau  $\alpha$  abgelehnt, wenn der aus der Stichprobe berechnete Wert  $v$  der Teststatistik  $V$  ein Element des Ablehnungsbereiches ist, andernfalls besteht keine Veranlassung die Nullhypothese zu verwerfen.

Das vorgegebene Signifikanzniveau  $\alpha$  entspricht dabei der Wahrscheinlichkeit eines Fehlers 1. Art, d.h. der Wahrscheinlichkeit, die Nullhypothese  $H_0$  abzulehnen, obwohl sie wahr ist.

Zur Veranschaulichung sei angenommen, dass

## Anhang A

- ein rechtsseitiger Test für einen Parameter  $\vartheta$  durchgeführt wird:

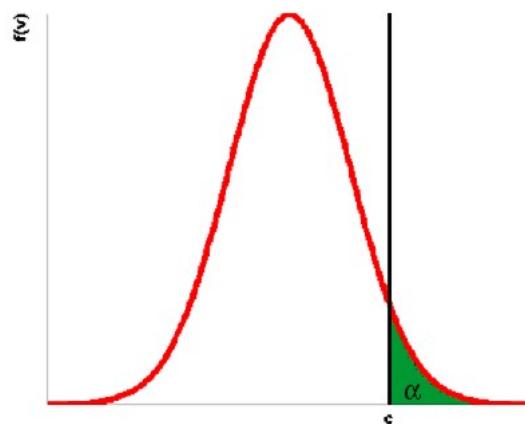
$$H_0 : \vartheta \leq \vartheta_0 \quad \text{und} \quad H_1 : \vartheta > \vartheta_0,$$

- die Teststatistik  $V$  bei Gültigkeit der Nullhypothese standardnormalverteilt ist:

$$V \sim N(0; 1).$$

Der Ablehnungsbereich der  $H_0$  wird dann durch alle Werte der Teststatistik gebildet, für die  $\{v | v > c\}$  gilt. Die Wahrscheinlichkeit, eine Realisation aus dem Ablehnungsbereich der Nullhypothese  $H_0$  zu erhalten, entspricht dem vorgegebenen Signifikanzniveau  $\alpha = P(V > c | \vartheta_0)$  und ist in der folgenden Abb. A.1 durch die markierte (grüne) Fläche gekennzeichnet.

Abbildung A.1.: Signifikanzniveau  $\alpha$  und Entscheidungsbereiche beim rechtsseitigen Test



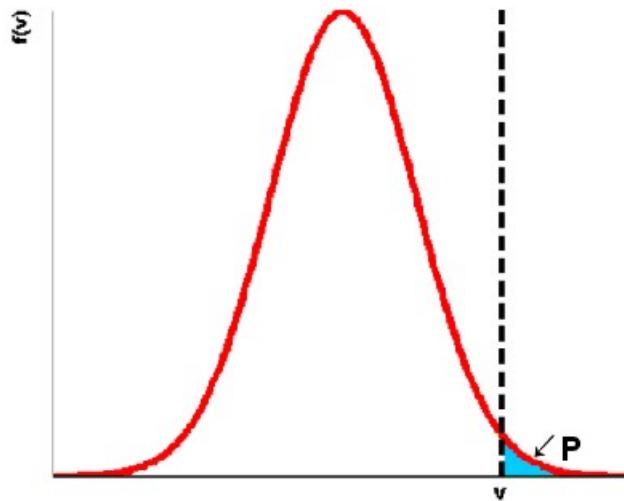
Nichtablehnungsbereich der  $H_0$  | Ablehnungsbereich der  $H_0$

Die Testentscheidung ist wie folgt: Ist der aus der Stichprobe berechnete Testwert  $v$  ein Element des Ablehnungsbereichs der  $H_0$ , so wird die Nullhypothese auf dem vorgegebenen Signifikanzniveau  $\alpha$  und basierend auf der Zufallsstichprobe vom Umfang  $n$  verworfen. Andernfalls besteht keine Veranlassung,  $H_0$  abzulehnen. Die Testentscheidung basiert somit auf einem Vergleich des Testwertes  $v$  mit den Entscheidungsbereichen.

Bei Verwendung statistischer Software (z.B. SPSS) wird ebenfalls der Testwert  $v$  auf der Grundlage der Stichprobe berechnet und im Output ausgewiesen. Zusätzlich wird die Überschreitungswahrscheinlichkeit dieses Testwertes  $v$  ausgegeben, d.h. die Wahrscheinlichkeit  $P(V > v | \vartheta_0)$ , dass die Teststatistik  $V$  einen Wert annimmt, der größer als dieser berechnete Testwert  $v$  ist (bei Gültigkeit der Nullhypothese  $H_0$ ). Diese Überschreitungswahrscheinlichkeit wird im Output statistischer Software sehr unterschiedlich bezeichnet (z.B. als Significance, P-value, 1-tailed P bzw. 1-tailed Sig. beim einseitigen Test bzw. 2-tailed P bzw. 2 - tailed Sig beim zweiseitigen Test). Hier sei das Symbol  $P$  verwendet, so dass  $P = P(V > v | \vartheta_0)$  gilt.

Abb. A.2 veranschaulicht diese Überschreitungswahrscheinlichkeit durch die markierte (himmelblaue) Fläche.

Abbildung A.2.: Überschreitungswahrscheinlichkeit  $P = P(V > v|\vartheta_0)$  bei Gültigkeit der  $H_0$



Der Nutzer der Software braucht nicht erst zu Tabellen der entsprechenden Verteilung der Teststatistik  $V$  greifen, um den bzw. die kritischen Werte und damit die Entscheidungsbereiche des Tests zu ermitteln. Im Output sind alle notwendigen Informationen für die Testentscheidung enthalten, die nunmehr auf dem Vergleich des vorgegebenen Signifikanzniveaus  $\alpha$  und der Überschreitungswahrscheinlichkeit  $P$  beruht.

Dies sei wie folgt gezeigt.

**a) Ablehnung der Nullhypothese  $H_0$**

Ergibt sich aufgrund einer konkreten Stichprobe ein Testwert  $v$ , der weit von  $\vartheta_0$  entfernt liegt, dann ist die Überschreitungswahrscheinlichkeit  $P = P(V > v|\vartheta_0)$  unter der Verteilung von  $H_0$  sehr klein.  $v$  ist ein für die Gültigkeit der Nullhypothese extremer Wert und die Nullhypothese erscheint unplausibel. Ein solcher Wert  $v$  kommt eher unter der Alternativhypothese zustande, so dass auf einen signifikanten Unterschied zwischen  $\vartheta_0$  und  $\vartheta$  geschlossen wird, d.h. die Nullhypothese abgelehnt wird.

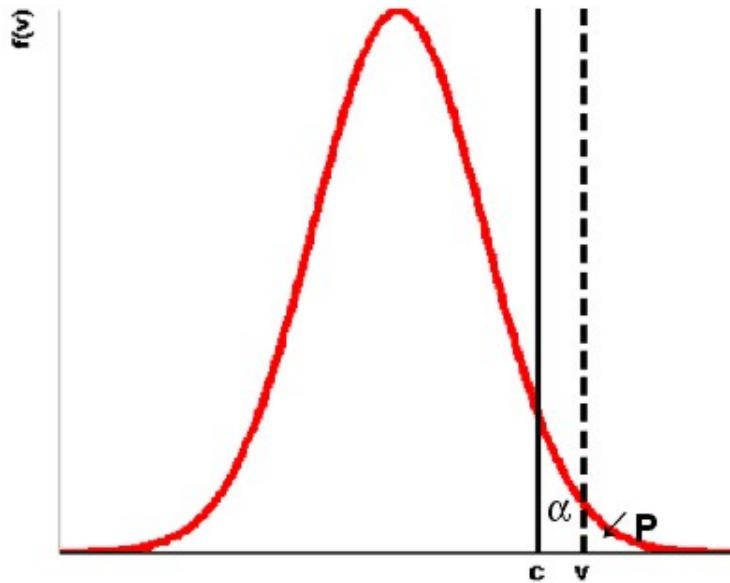
Entscheidungsregel:

Ist die im Output der Software ausgegebene Überschreitungswahrscheinlichkeit  $P$  kleiner als das vorgegebene Signifikanzniveau  $\alpha$  ( $P < \alpha$ ), so impliziert dies, dass der Testwert  $v$  ein Element des Ablehnbereiches der  $H_0$  zum vorgegebenen Signifikanzniveau  $\alpha$  ist. Die Nullhypothese wird abgelehnt.

## Anhang A

Bei dem hier demonstrierten rechtsseitigen Test wird diese Entscheidungsregel in der Abb. A.3 deutlich.

Abbildung A.3.: Signifikanzniveau  $\alpha = P(V > c|\vartheta_0)$  und Überschreitungswahrscheinlichkeit  $P = P(V > v|\vartheta_0)$  bei Gültigkeit der Nullhypothese  $H_0$  für einen rechtsseitigen Test



Nichtablehnungsbereich der  $H_0$  | Ablehnungsbereich der  $H_0$

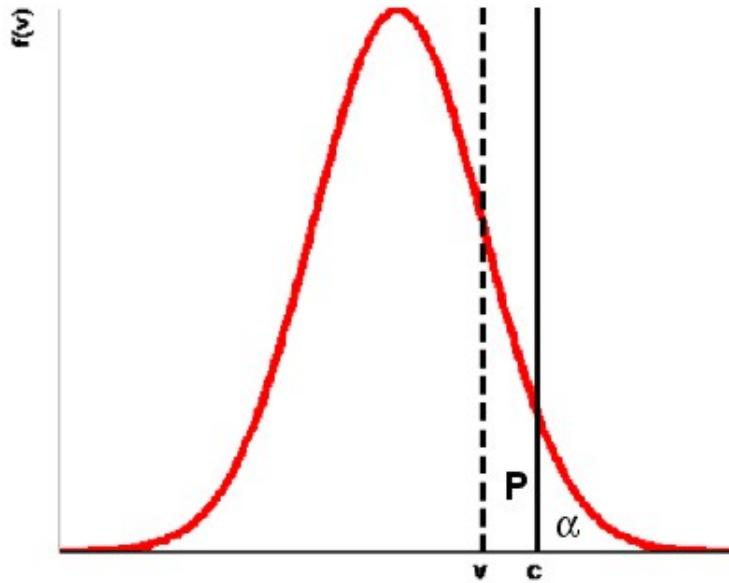
### b) Nichtablehnung der Nullhypothese

Ergibt sich aufgrund einer konkreten Stichprobe ein Testwert  $v$ , der relativ nahe bei  $\vartheta_0$  liegt, dann ist die Überschreitungswahrscheinlichkeit  $P = P(V > v|\vartheta_0)$  unter der Verteilung von  $H_0$  groß.  $v$  ist ein für die Verteilung der Nullhypothese plausibler Wert, die Abweichung zwischen  $v$  und  $\vartheta$  kann als zufällig angesehen werden. Die Nullhypothese wird in diesem Fall nicht abgelehnt.

Entscheidungsregel:

Ist  $P \geq \alpha$ , so impliziert dies, dass der Testwert  $v$  ein Element des Nichtablehnungsbereichs der  $H_0$  ist. Die Nullhypothese wird nicht abgelehnt.

Abbildung A.4.: Signifikanzniveau  $\alpha = P(V > c|\vartheta_0)$  und Überschreitungswahrscheinlichkeit  $P = P(V > v|\vartheta_0)$  bei Gültigkeit der Nullhypothese  $H_0$  für einen rechtsseitigen Test



Nichtablehnungsbereich der  $H_0$  | Ablehnungsbereich der  $H_0$

Mit den gleichen Regeln sind die Testentscheidungen bei einem linksseitigen Test bzw. einem zweiseitigen Test zu treffen.

Wird ein linksseitiger Test für einen Parameter  $\vartheta$  mit  $H_0 : \vartheta \geq \vartheta_0$  und  $H_1 : \vartheta < \vartheta_0$  durchgeführt, dann gilt  $\alpha = P(V < c)$ , wobei  $\alpha$  vorgegeben ist und der kritische Wert  $c$  als Quantil der Ordnung  $\alpha$  aus der Tafel der Standardnormalverteilung aufzusuchen ist. Der Ablehnungsbereich wird durch alle Werte der Teststatistik  $V$  gebildet, für die gilt  $\{v|v < c\}$ . Die im Output der Software ausgegebene Wahrscheinlichkeit beinhaltet nun  $P = P(V < v)$ .

Wird ein zweiseitiger Test für einen Parameter  $\vartheta$  mit  $H_0 : \vartheta = \vartheta_0$  und  $H_1 : \vartheta \neq \vartheta_0$  durchgeführt, dann gilt  $\alpha = P(V < -c) + P(V > c) = \alpha/2 + \alpha/2$ , wobei  $\alpha$  vorgegeben ist und der kritische Wert  $c$  als Quantil der Ordnung  $1-\alpha/2$  aus der Tafel der Standardnormalverteilung aufzusuchen ist. Der Ablehnungsbereich wird durch alle Werte der Teststatistik  $V$  gebildet, für die  $\{v|v < -c \text{ oder } v > c\}$  gilt. Die unter SPSS berechnete Wahrscheinlichkeit beinhaltet nun  $P = P(V < -v) + P(V > v)$ .

In beiden Fällen ist die Testentscheidung wie vorher:

a) Ablehnung der Nullhypothese  $H_0$ :

Ist  $P < \alpha$ , so impliziert dies, dass der berechnete Wert der Teststatistik ein Element des Ablehnungsbereiches der  $H_0$  zum vorgegebenen Signifikanzniveau  $\alpha$  ist. Die Nullhypothese wird

## Anhang A

abgelehnt.

b) Nichtablehnung der Nullhypothese  $H_0$ :

Ist  $P \geq \alpha$ , so impliziert dies, dass der berechnete Wert der Teststatistik ein Element des Nichtablehnungsbereiches der  $H_0$  zur Wahrscheinlichkeit  $1 - \alpha$  ist. Die Nullhypothese wird nicht abgelehnt.

Diese Testentscheidungen gelten entsprechend für nichtparametrische Tests.

# Anhang B

## Kolmogorov-Smirnov-Test: Quantile $d_{n;1-\alpha}$ der Teststatistik $D_n$

n	1 - $\alpha$			
	0,90	0,95	0,98	0,99
1	.95000	.97500	.99000	.99500
2	.77639	.84189	.90000	.92929
3	.63604	.70760	.78456	.82900
4	.56522	.62394	.68887	.73424
5	.50945	.56328	.62718	.66853
6	.46799	.51926	.57741	.61661
7	.43607	.48342	.53844	.57581
8	.40962	.45427	.50654	.54179
9	.38746	.43001	.47960	.51332
10	.36866	.40925	.45662	.48893
11	.35242	.39122	.43670	.46770
12	.33815	.37543	.41918	.44905
13	.32549	.36143	.40362	.43247
14	.31417	.34890	.38970	.41762
15	.30397	.33760	.37713	.40420
16	.29472	.32733	.36571	.39201
17	.28627	.31796	.35528	.38086
18	.27851	.30936	.34569	.37062
19	.27136	.30143	.33685	.36117
20	.26473	.29408	.32866	.35241
22	.25283	.28087	.31394	.33666
24	.24242	.26931	.30104	.32286
26	.23320	.25907	.28962	.31064
28	.22497	.24993	.27942	.29971
30	.21756	.24170	.27023	.28987
35	.20185	.22425	.25073	.26897
40	.18913	.21012	.23494	.25205
45	.17856	.19837	.22181	.23798
50	.16959	.18841	.21068	.22604
60	.15511	.17231	.19267	.20673
70	.14381	.15975	.17863	.19167
80	.13467	.14960	.16728	.17949
90	.12709	.14117	.15786	.16938
100	.12067	.13403	.14987	.16081

Quelle: Rönnz, Strohe (Hrsg.) (1994), S. 186

*Anhang B*

# Anhang C

$\chi^2$ -Verteilung: Quantile  $\chi_{n;1-\alpha}$  der Verteilungsfunktion F für die

Wahrscheinlichkeit  $1 - \alpha : F(\chi_{n,1-\alpha}^2) = P(\chi^2 \leq \chi_{n,1-\alpha}^2) = 1 - \alpha$

n/ $\alpha$	0,10	0,05	0,01	0,001	n
1	2,71	3,841	6,635	10,827	1
2	4,61	5,991	9,210	13,815	2
3	6,25	7,815	11,345	16,268	3
4	7,78	9,488	13,277	18,465	4
5	9,24	11,070	15,086	20,517	5
6	10,6	12,592	16,812	22,457	6
7	12,0	14,067	18,475	24,322	7
8	13,4	15,507	20,090	26,125	8
9	14,7	16,919	21,666	27,877	9
10	16,0	18,307	23,209	29,588	10
11	17,3	19,675	24,725	31,264	11
12	18,5	21,026	26,217	32,909	12
13	19,8	22,362	27,688	34,528	13
14	21,1	23,685	29,141	36,123	14
15	22,3	24,996	30,578	37,697	15
16	23,5	26,296	32,000	39,252	16
17	24,8	27,587	33,409	40,790	17
18	26,0	28,869	34,805	42,312	18
19	27,2	30,144	36,191	43,820	19
20	28,4	31,410	37,566	45,315	20
21	29,6	32,671	38,932	46,797	21
22	30,8	33,924	40,289	48,268	22
23	32,0	35,172	41,638	49,797	23
24	33,2	36,415	42,980	51,179	24
25	34,4	37,652	44,314	52,620	25
26	35,6	38,885	45,642	54,052	26
27	36,7	40,113	46,963	55,476	27
28	37,9	41,337	48,278	56,893	28
29	39,1	42,557	49,588	58,302	29
30	40,3	43,773	50,892	59,703	30
40	51,8	55,8	63,7	73,4	40
50	63,2	67,5	76,2	86,7	50
60	74,4	79,1	88,4	99,6	60
70	85,5	90,5	100,4	112,3	70
80	96,6	101,9	112,3	124,8	80
90	107,6	113,1	124,1	137,2	90
100	118,5	124,3	135,8	149,4	100

(Quelle: Rönnz, B., Strohe, H.G. (Hrsg.) (1994), S. 70)

*Anhang C*

# Anhang D

## Zur Varianzanalyse

Folgende Fakten werden für die weiteren Herleitungen verwendet.

- Aufgrund des Verschiebungssatzes gilt:  $\text{Var}(X_{j,i}) = E(X_{j,i}^2) - [E(X_{j,i})]^2$

$$E(X_{j,i}^2) = \text{Var}(X_{j,i}) + [E(X_{j,i})]^2$$

Wegen der Voraussetzung der Varianzhomogenität ist:  $\text{Var}(X_{j,i}) = \sigma^2$ . Weiterhin gilt:

$$[E(X_{j,i})]^2 = \mu_j^2. \text{ Somit resultiert:}$$

$$E(X_{j,i}^2) = \text{Var}(X_{j,i}) + [E(X_{j,i})]^2 = \sigma^2 + \mu_j^2. \quad (\text{D.1})$$

- Aufgrund des Verschiebungssatzes gilt:  $\text{Var}(\bar{X}_j) = E(\bar{X}_j^2) - [E(\bar{X}_j)]^2$

$$E(\bar{X}_j^2) = \text{Var}(\bar{X}_j) + [E(\bar{X}_j)]^2$$

Wegen der Voraussetzung der Varianzhomogenität ist:  $\text{Var}(\bar{X}_j) = \sigma^2/n_j$ . Weiterhin gilt:

$$[E(\bar{X}_j)]^2 = \mu_j^2. \text{ Somit resultiert:}$$

$$E(\bar{X}_j^2) = \text{Var}(\bar{X}_j) + [E(\bar{X}_j)]^2 = \sigma^2/n_j + \mu_j^2. \quad (\text{D.2})$$

- Aufgrund des Verschiebungssatzes gilt:  $\text{Var}(\bar{X}) = E(\bar{X}^2) - [E(\bar{X})]^2$

$$E(\bar{X}^2) = \text{Var}(\bar{X}) + [E(\bar{X})]^2$$

Es gilt:  $\text{Var}(\bar{X}) = \sigma^2/n$  und  $[E(\bar{X})]^2 = \mu^2$  und damit

$$E(\bar{X}^2) = \text{Var}(\bar{X}) + [E(\bar{X})]^2 = \sigma^2/n + \mu^2. \quad (\text{D.3})$$

### Bestimmung des Erwartungswertes der Summe der Abweichungsquadrat SQG sowie der Varianz $S^2$

$$\begin{aligned} E(SQG) &= E \left( \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,k} - \bar{X})^2 \right) = E \left( \sum_{j=1}^m \sum_{i=1}^{n_j} X_{j,i}^2 - n \bar{X}^2 \right) \\ &= E \left( \sum_{j=1}^m \sum_{i=1}^{n_j} X_{j,i}^2 \right) - E(n \bar{X}^2) \end{aligned}$$

## Anhang D

$$= \sum_{j=1}^m \sum_{i=1}^{n_j} E(X_{j,i}^2) - nE(\bar{X}^2)$$

Einsetzen von (D.1) und (D.3) führt zu:

$$\begin{aligned} E(SQG) &= \sum_{j=1}^m \sum_{i=1}^{n_j} (\sigma^2 + \mu_j^2) - n \left( \frac{\sigma^2}{n} + \mu^2 \right) \\ &= \sum_{j=1}^m n_j \sigma^2 + \sum_{j=1}^m n_j \mu_j^2 - \sigma^2 - n\mu^2 \\ &= \sigma^2 \sum_{j=1}^m n_j - \sigma^2 + \sum_{j=1}^m n_j \mu_j^2 - n\mu^2 \\ &= n\sigma^2 - \sigma^2 + \sum_{j=1}^m n_j (\mu_j - \mu)^2 \\ &= (n-1)\sigma^2 + \sum_{j=1}^m n_j (\mu_j - \mu)^2 \end{aligned} \quad (\text{D.4})$$

$$E\left(\frac{SQG}{n-1}\right) = E(S^2) = \sigma^2 + \frac{1}{n-1} \sum_{j=1}^m n_j (\mu_j - \mu)^2 \quad (\text{D.5})$$

Nur unter  $H_0$  gilt:

$$E\left(\frac{SQG}{n-1}\right) = E(S^2) = \sigma^2 \quad (\text{D.6})$$

### Bestimmung des Erwartungswertes der Summe der Abweichungsquadrat **SQI** sowie der Varianz $S_I^2$

$$\begin{aligned} E(SQI) &= E \left[ \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,i} - \bar{X}_j)^2 \right] = E \left[ \sum_{j=1}^m \sum_{i=1}^{n_j} X_{j,i}^2 - n\bar{X}_j^2 \right] \\ &= \sum_{j=1}^m \sum_{i=1}^{n_j} E[X_{j,i}^2] - \sum_{j=1}^m n_j E[\bar{X}_j^2] \end{aligned}$$

Einsetzen von (D.1) und (D.2) führt zu:

$$\begin{aligned} E(SQI) &= \sum_{j=1}^m \sum_{i=1}^{n_j} (\sigma^2 + \mu_j^2) - \sum_{j=1}^m n_j \left( \frac{\sigma^2}{n_j} + \mu_j^2 \right) \\ &= \sum_{j=1}^m (n_j^2 \sigma^2 + n_j \mu_j^2) - \sum_{j=1}^m (\sigma^2 + n_j \mu_j^2) \\ &= \sigma^2 \sum_{j=1}^m n_j + \sum_{j=1}^m n_j \mu_j^2 - m\sigma^2 - \sum_{j=1}^m n_j \mu_j^2 = \sigma^2 n - m\sigma^2 = (n-m)\sigma^2 \end{aligned} \quad (\text{D.7})$$

$$E\left(\frac{SQI}{n-m}\right) = E(S_I^2) = \sigma^2 \quad (\text{D.8})$$

Dies gilt sowohl unter  $H_1$  als auch unter  $H_0$ .

## Bestimmung des Erwartungswertes der Summe der Abweichungsquadratsumme SQZ sowie der Varianz $S_Z^2$

$$\begin{aligned} E(SQZ) &= E\left(\sum_{j=1}^m n_j(\bar{X}_j - \bar{X})^2\right) = E\left(\sum_{j=1}^m n_j\bar{X}_j^2 - n\bar{X}^2\right) \\ &= E\left(\sum_{j=1}^m n_j\bar{X}_j^2\right) - E(n\bar{X}^2) = \sum_{j=1}^m n_jE(\bar{X}_j^2) - nE(\bar{X}^2) \end{aligned}$$

Einsetzen von (D.2) und (D.3) führt zu:

$$\begin{aligned} E(SQZ) &= \sum_{j=1}^m n_j\left(\frac{\sigma^2}{n_j} + \mu_j^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= m\sigma^2 + \sum_{j=1}^m n_j\mu_j^2 - \sigma^2 - n\mu^2 \\ &= m\sigma^2 - \sigma^2 + \sum_{j=1}^m n_j\mu_j^2 - n\mu^2 \\ &= (m-1)\sigma^2 + \sum_{j=1}^m n_j(\mu_j - \mu)^2 \end{aligned} \tag{D.9}$$

$$E\left(\frac{SQZ}{m-1}\right) = \sigma^2 + \frac{1}{m-1} \sum_{j=1}^m n_j(\mu_j - \mu)^2 \tag{D.10}$$

Nur unter  $H_0$  gilt:

$$E\left(\frac{SQZ}{m-1}\right) = \sigma^2 \tag{D.11}$$

## Verteilungsbetrachtungen

$$\begin{aligned} \frac{1}{\sigma^2}SQG &= \frac{1}{\sigma^2} \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,i} - \bar{X})^2 \\ &= \frac{1}{\sigma^2} \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,i} - \mu)^2 - \frac{n}{\sigma^2} (\bar{X} - \mu)^2 \\ &= \sum_{j=1}^m \sum_{i=1}^{n_j} \left(\frac{X_{j,i} - \mu}{\sigma}\right)^2 - n \left(\frac{\bar{X} - \mu}{\sigma}\right)^2 \\ &= \sum_{j=1}^m \sum_{i=1}^{n_j} \left(\frac{X_{j,i} - \mu}{\sigma}\right)^2 - \left(\frac{\bar{X} - \mu}{\sqrt{n}}\right)^2 \end{aligned} \tag{D.12}$$

## Anhang D

Da vorausgesetzt wurde, dass die Zufallsvariablen  $X_{j,i}$  unabhängig und  $N(\mu_j; \sigma^2)$ -verteilt sind, ist die Stichprobenfunktion  $\bar{X}$   $N(\mu_j; \sigma^2/n)$ -verteilt. Der 2. Teil in (D.12) ist somit das Quadrat einer Zufallsvariablen  $(\bar{X} - \mu)(n)^{1/2}/\sigma$ , die  $N(0; 1)$ -verteilt ist. Wenn  $H_0$  gilt (also  $\mu_j = \mu$  für alle  $j$ ), dann enthält der 1. Teil von (D.12) die Summe der Quadrate von  $n$  unabhängigen Zufallsvariablen  $(X_{j,i} - \mu)/\sigma$ , die jeweils  $N(0; 1)$ -verteilt sind. Entsprechend den Eigenschaften der  $\chi^2$ -Verteilung ist unter  $H_0$

$$SQG/\sigma^2 = (n - 1)S^2/\sigma^2 \sim \chi^2(n - 1).$$

Analog ist:

$$\begin{aligned} \frac{1}{\sigma^2} SQI &= \frac{1}{\sigma^2} \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,i} - \bar{X}_j)^2 \\ &= \frac{1}{\sigma^2} \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{j,i} - \mu_j)^2 - \frac{1}{\sigma^2} n_j (\bar{X}_j - \mu_j)^2 \\ &= \sum_{j=1}^m \sum_{i=1}^{n_j} \left( \frac{X_{j,i} - \mu_j}{\sigma} \right)^2 - \sum_{j=1}^m n_j \left( \frac{\bar{X}_j - \mu_j}{\sigma} \right)^2 \\ &= \sum_{j=1}^m \sum_{i=1}^{n_j} \left( \frac{X_{j,i} - \mu_j}{\sigma} \right)^2 - \sum_{j=1}^m \left( \frac{\bar{X}_j - \mu_j}{\sqrt{n_j}} \right)^2 \end{aligned} \quad (D.13)$$

Da vorausgesetzt wurde, dass die Zufallsvariablen  $X_{j,i}$  unabhängig und  $N(\mu_j; \sigma^2)$ -verteilt sind, so sind die Stichprobenfunktionen  $\bar{X}_j$  ebenfalls unabhängig und  $N(\mu_j; \sigma^2/n)$ -verteilt. Der 2. Teil in (D.13) ist somit die Summe der Quadrate von  $m$  unabhängigen Zufallsvariablen  $(\bar{X}_j - \mu_j)(n_j)^{1/2}/\sigma$ , die jeweils  $N(0; 1)$ -verteilt sind.

Der 1. Teil von (D.13) beinhaltet die Summe der Quadrate von  $n$  unabhängigen Zufallsvariablen  $(X_{j,i} - \mu_j)/\sigma$ , die jeweils  $N(0; 1)$ -verteilt sind. Entsprechend den Eigenschaften der  $\chi^2$ -Verteilung ist  $SQI/\sigma^2 = (n - m)S_I^2/\sigma^2 \sim \chi^2(n - m)$ .

Weiterhin ist:

$$\begin{aligned} \frac{1}{\sigma^2} SQZ &= \frac{1}{\sigma^2} \sum_{j=1}^m (\bar{X}_j - \bar{X})^2 \\ &= \frac{1}{\sigma^2} \sum_{j=1}^m n_j (\bar{X}_j - \mu)^2 - \frac{1}{\sigma^2} \sum_{j=1}^m n_j (\bar{X} - \mu)^2 \\ &= \frac{1}{\sigma^2} \sum_{j=1}^m n_j (\bar{X}_j - \mu)^2 - \frac{1}{\sigma^2} n (\bar{X} - \mu)^2 \\ &= \sum_{j=1}^m n_j \left( \frac{\bar{X}_j - \mu}{\sigma} \right)^2 - n \left( \frac{\bar{X} - \mu}{\sigma} \right)^2 \end{aligned} \quad (D.14)$$

$$= \sum_{j=1}^m \left( \frac{\bar{X}_j - \mu}{\sigma / \sqrt{n_j}} \right)^2 - \left( \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right)^2$$

Da vorausgesetzt wurde, dass die Zufallsvariablen  $X_{j,i}$  unabhängig und  $N(\mu_j; \sigma^2)$ -verteilt sind, so sind die Stichprobenfunktionen  $\bar{X}_j$  ebenfalls unabhängig und  $N(\mu_j; \sigma^2/n_j)$ -verteilt und die Stichprobenfunktion  $\bar{X}$  ist  $N(\mu; \sigma^2/n)$ -verteilt. Der 2. Teil in (D.14) ist somit das Quadrat einer Zufallsvariablen  $(\bar{X} - \mu)(n)^{1/2}/\sigma$ , die  $N(0; 1)$ -verteilt ist. Wenn  $H_0$  gilt (also  $\mu_j = \mu$  für alle  $j$ ), dann enthält der 1. Teil von (D.14) die Summe der Quadrate von  $m$  unabhängigen Zufallsvariablen  $(\bar{X}_j - \mu)(n_j)^{1/2}/\sigma$ , die jeweils  $N(0; 1)$ -verteilt sind. Entsprechend den Eigenschaften der  $\chi^2$ -Verteilung ist unter  $H_0$

$$SQZ/\sigma^2 = (m-1)S_Z^2/\sigma^2 \sim \chi^2(m-1).$$

*Anhang D*

# Literaturverzeichnis

- [1] Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, ALLBUS 1980-94, Codebuch, ZA-Nr. 1795, Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln, Zentrum für Umfragen, Methoden und Analysen Mannheim.
- [2] Andrews, D.F. (1972), Plots of high-dimensional data, Biometrics 28, S. 125-136
- [3] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tuckey, J.W. (1972), Robust estimation of location, Princeton University Press, Princeton
- [4] Bailey, J.R. (1977), Tables of the Bonferroni t Statistics, Journal of the American Statistical Association, 72, S. 469 - 478
- [5] Barnett, V., Lewis, T. (1994), Outliers in statistical data, 3rd. Edition, Wiley, New York
- [6] Bortz, J. (1993) Statistik, Springer, Berlin et al.
- [7] Bosch, K. (1992), Statistik-Taschenbuch, Oldenbourg, München, Wien
- [8] Bühl, A., Zöfel, P. (1994), SPSS für Windows Version 6, Addison-Wesley
- [9] Büning, H., Trenkler, G. (1978), Nichtparametrische statistische Methoden, Walter de Gruyter, Berlin, New York
- [10] Büning, H. (1991), Robuste und adaptive Tests, Walter de Gruyter, Berlin, New York
- [11] D'Agostino, R.B., Stephens, M.A. (1986), Goodness-of-fit-Techniques, Dekker, New York
- [12] David, H.A., Hartley, H.O., Pearson, E.S. (1954), The distribution of the ratio, in a single normal sample, of range to standard deviation, Biometrika, vol. 41, p. 482-493
- [13] Dixon, W.J. (1951), Ratios involving extreme values, Annals of Mathematical Statistics, vol. 22, p.68 - 78

## Literaturverzeichnis

- [14] Dunn, O.J., (1961), Multiple Comparisons Among Means, *Journal of the American Statistical Association*, 56, S. 52-64
- [15] du Toit, S.H.C., Steyn, A.G.W., Stumpf, R.H., (1986), *Graphical exploratory data analysis*, Springer, New York et al.
- [16] Eckstein, P.P. (1997), *Angewandte Statistik mit SPSS*, Betriebswirtschaftlicher Verlag Dr. Th. Gabler GmbH, Wiesbaden
- [17] Ferguson, Th.S. (1961), Rules for rejection of outliers, *Rev. Inst. Internat. Statist.*, 29, S. 29-43
- [18] Fox, J., Long, J.S. (1990), *Modern Methodes of Data Analysis*, Sage Publications, London
- [19] Gastwirth, J.L., Cohen, M.L. (1970), Small sample behaviour of some robust linear estimators of location, *Journal of the American Statistical Association*, vol. 65, p. 946 - 973
- [20] Griliches, Z., Intriligator, M.D. (Hrsg.)(1992), *Handbook of Econometrics*, Volume II, North-Holland, Amsterdam et al.
- [21] Grubbs, F.E. (1950), Sample criteria for testing outlying observations, *Annals of Mathematical Statistics*, vol. 21, p. 27-58
- [22] Grubbs, F.E., Beck, G. (1972), Extension of sample sizes and percentage points for significance tests of outlying observations, *Technometrics*, vol. 14, p. 847 - 854
- [23] Härdle, W. (1991), *Smoothing Techniques*, Springer, New York
- [24] Härdle, W., Klinke, S., Turlach, B.A. (1995), *XploRe: An Interactive Statistical Computer Environment*, Springer, New York
- [25] Harter, H.L. (1960), Critical Values for Duncans New Multiple Range Test, *Biometrics*, 16, No. 4
- [26] Hartung, J., Elpelt, B., Klösener, K.-H. (1993), *Statistik*, 9. Auflage, Oldenbourg, München, Wien
- [27] Hawkins, D.M. (1980), *Identifikation of outliers*, Chapman and Hall, London, New York
- [28] Heiler, S., Michels, P. (1994), *Deskriptive und explorative Datenanalyse*, Oldenbourg, München, Wien
- [29] Hoaglin, D.C., Mosteller, F., Tukey, J.W. (1983), *Understanding robust and exploratory data analysis*, Wiley, New York

- [30] Hochstädter, D., Kaiser, U. (1988), Varianz- und Kovarianzanalyse, Harri Deutsch, Frankfurt am Main, Thun
- [31] Huber, P.J. (1981), Robust statistics, Wiley, New York
- [32] Jobson, J.D. (1991), Applied Multivariate Data Analysis, Vol. I: Regression and Experimental Design, Springer, Berlin et. al.
- [33] Kähler, W.-M. (1994), SPSS für Windows, Vieweg Wiesbaden
- [34] Kinnison (1985), Applied extreme value statistics, Macmillan, New York
- [35] Kockelkorn, U. (1995), Lineare Modelle, TU Berlin, unveröffentlichtes Manuskript
- [36] Launer, R.L., Wilkinson, G.N. (Hrsg.)(1979), Robustness in statistics, Academic Press, New York
- [37] Läuter, H., Pincus, R. (1989), Mathematisch-statistische Datenanalyse, Akademie-Verlag, Berlin
- [38] Lillefors, H.W. (1967), On the Kolmogorov-Smirnov test for normality with mean and variance unknown; Journal of the American Statistical Association, vol. 62, S. 399 - 402
- [39] Lillefors, H.W. (1969), On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown, Journal of the American Statistical Association, vol. 64, S. 387 - 389
- [40] Lohse, H., Ludwig, R., Röhr, M. (1982), Statistische Verfahren für Psychologen, Pädagogen und Soziologen, Volk und Wissen, Berlin
- [41] Miller, R.G.jr (1966), Simultaneous Statistical Inference, McGraw-Hill, New York
- [42] Müller, H., Neumann, P., Storm, R. (1973), Tafeln der mathematischen Statistik, Fachbuchverlag, Leipzig
- [43] Müller, P.H. (Hrsg.) (1975), Wahrscheinlichkeitsrechnung und mathematische Statistik - Lexikon der Stochastik., Akademie-Verlag, Berlin
- [44] Pearson, E.S., Hartley, H.O., (1970) Biometrika tables for statisticians I, Cambridge University Press, London  
Pearson, E.S., Hartley, H.O., (1970) Biometrika tables for statisticians II, Cambridge University Press, London
- [45] Rasch, D., Enderlein, G., Herrendörfer, G. (1973), Biometrie, Deutscher Landwirtschaftsverlag, Berlin

## *Literaturverzeichnis*

- [46] Rönz, B., Strohe, H.G. (Hrsg.) (1994), Lexikon Statistik, Gabler-Verlag, Wiesbaden
- [47] Sachs, L. (1992), Angewandte Statistik, Springer Verlag, Berlin et al.
- [48] Schlittgen, R. (1990), Einführung in die Statistik, Oldenbourg, München, Wien
- [49] SPSS for Windows: Base System User's Guide, Release 6.0, SPSS Inc., 1993
- [50] SPSS Base System Syntax Reference Guide 6.0, SPSS Inc.
- [51] SPSS Statistical Algorithms 2nd Ed., SPSS Inc.
- [52] SPSS for Windows: Professional Statistics 6.0, SPSS Inc.
- [53] SPSS for Windows: Advanced Statistics 6.0, SPSS Inc.
- [54] SPSS for Windows: Tables 6.0
- [55] SPSS for Windows: CHAID 6.0
- [56] SPSS Categories
- [57] SPSS for Windows: Trends 6.0
- [58] Staudte, R.G., Sheather, S.J. (1990), Robust estimation and testing, Wiley, New York
- [59] Tukey, J.W. (1977), Exploratory Data Analysis, Addison-Wesley, London, Amsterdam
- [60] Weber, E. (1972) Grundriss der biologischen Statistik, Gustav Fischer, Jena
- [61] Welch, B.L. (1947), The generalization of Students problem when several different population variances are involved, Biometrika 34, 28 - 35.
- [62] Wittenberg, R. (1991), Computergestützte Datenanalyse, Gustav Fischer Verlag Stuttgart, UTB 1603
- [63] Wittenberg, R., Cramer, H. (1992), Datenanalyse mit SPSS, Gustav Fischer Verlag Stuttgart, UTB 1602

# Index

- $\alpha$ -Gastwirth-Cohen-Mittelwert, 47
- $\alpha$ -getrimmte Varianz, 62
- $\alpha$ -getrimmter Mittelwert, 45, 62
- $\alpha$ -winsorisierter Mittelwert, 46
- $\chi^2$ -Verteilung, 203, 208
- Überschreitungswahrscheinlichkeit, 105, 106, 108, 116, 121, 123, 196
- 3-D Rotation, 32
- 3-D Scatterplot, 29
- abhängige Stichproben, 135
- Ablehnungsbereich, 103, 113, 114, 124, 125, 144, 152, 195
- Alternativhypothese, 101, 110, 161, 195
- Analyze, 17, 36, 59, 69, 78, 86, 93, 97, 105, 115, 121, 137, 146, 153, 178
- Andrews wave, 55
- Andrews-Plot, 33
- Anordnungswert, 44, 97
- Anordnungswerte, 36
- ANOVA, 159, 164, 170, 178–180, 182, 183, 185, 192
- Anpassungstests, 100
- Anteilswert, 85, 120
- Approximation, 84, 113, 120, 122, 124, 169
- Arcus-Sinus-Transformation, 129
- Ausreißer, 3, 17, 73, 79, 85, 99, 131, 133
- Ausreißertest, 3, 35
- Automatic Recode, 15
- Balkendiagramm, 66, 117, 136
- Bandbreite, 82–84
- Bar Chart, 66, 69
- Behrens-Fisher-Problem, 152
- Beispiel 2.1, 19
- Beispiel 2.2, 20, 22, 25, 60
- Beispiel 2.3, 27, 29
- Beispiel 2.4, 34
- Beispiel 2.5, 37, 40, 41, 43, 47, 59
- Beispiel 3.1, 71
- Beispiel 3.10, 131
- Beispiel 3.2, 73, 79, 80, 116
- Beispiel 3.3, 82
- Beispiel 3.4, 87, 94, 108, 130
- Beispiel 3.5, 96–98
- Beispiel 3.6, 99
- Beispiel 3.7, 102
- Beispiel 3.8, 119, 124
- Beispiel 3.9, 122
- Beispiel 4.1, 137, 141, 146, 147, 154
- Beispiel 4.2, 156
- Beispiel 4.3, 181
- Beispiel 4.4, 191
- benutzerdefinierte Missings, 5
- Binomial-Test, 4, 120
- Binomialverteilung, 123, 129, 166
- Bonferroni, 171, 185
- Box-Cox-Transformation, 128
- Box-Whisker-Plot, 22

## Index

- Boxplot, 3, 17, 37, 38, 72, 73, 91, 100, 133, 136, 137, 139
- Chart Editor, 24, 27, 29, 30, 32, 78
- Chi-Quadrat-Anpassungstest, 4
- chi-quadrat-verteilt, 112, 144, 151, 164, 177
- Chi-Quadrat-Verteilung, 113, 143, 151
- clustered bar chart, 70
- Compute, 13, 130
- Cut point, 156
- Data Editor, 5, 7, 12, 13, 15, 24, 28
- Datenmodifikation, 4
- Datenselektion, 4
- David-Hartley-Pearson-Test, 42
- Descriptive Statistics, 17, 36, 59, 69, 78, 86, 93, 97, 105, 137, 146
- Descriptives, 36, 40, 43, 45, 46, 62, 93
- deskriptive Statistik, 3, 116
- dichotom, 120, 121, 124
- diskret, 66, 76, 102, 109–112, 135
- Dixon's r-Statistik, 40
- Draftsman-Display, 26
- Duncan-Test, 175, 189
- einfacher Scatterplot, 26
- Error Bar, 139
- Erwartungswert, 61, 112, 140, 142, 143, 149, 150, 159, 163, 205–207
- explorative Datenanalyse, 3, 17, 66
- Explore, 18, 21, 45, 46, 59, 73, 78, 86, 93, 97, 105, 137, 146
- Exponentialverteilung, 104, 106, 107
- Extremwerte, 23, 25, 28, 39, 79
- Exzeß, 92, 93
- F-Test, 4, 142, 144, 145, 170
- F-Verteilung, 142, 144, 145, 164, 177, 183, 190
- Fälle auswählen, 9
- Fälle gewichten, 12
- Faktor, 159
- Faktorstufe, 159, 179, 181, 182
- Faktorvariable, 4, 70, 135, 140, 141, 145, 146, 148, 153, 159, 179, 181, 191
- Fehler 1. Art, 100, 114, 165, 168, 195
- Fehler 2. Art, 100, 114
- Fehlerbalken-Diagramm, 139
- fence, 23, 100
- Filtervariable, 11
- Freiheitsgrad, 84, 105, 112, 113, 116, 117, 143, 163, 168, 177
- Frequencies, 69, 78, 93
- Gewichtsfunktion, 50, 52, 54–58
- Gleichverteilung, 81, 106, 107, 115, 116, 118
- Graphs, 26, 67, 77, 86, 89, 139
- Grubbs-Test, 36
- Grubbs/Beck-Test, 39
- Gruppenvariable, 6
- Gruppiertes Balkendigramm, 70
- Gruppierungsvariable, 70, 73, 154, 155
- Häufigkeitsdichte, 76
- Häufigkeitsverteilung, 3, 12, 22, 65, 81, 92, 95, 97, 100, 111, 115, 126
- Hampel-Schätzer, 52
- Heverage, 93
- Histogramm, 4, 76, 87, 89, 131–133, 136, 137
- Homogenous Subsets, 186, 188–190
- Homoskedastizität, 159
- Huber-k-Schätzer, 50
- induktive Statistik, 3
- Interquartile Range, 23
- interquartile range, 61

- Interquartilsabstand, 23, 25, 61, 80, 91, 93, 146, 148
- kategorial, 68, 69
- Kerndichteschätzung, 81
- kernel function, 82
- Kernfunktion, 82
- Klasse, 6, 9, 78, 110, 113, 115, 116, 118
- Klassenbreite, 76, 78, 80, 137
- Klassengrenze, 76
- Klassenmitte, 76, 79
- Kolmogorov-Smirnov-Test, 4, 101, 130, 191, 201
- Kolmogorov-Verteilung, 102, 104, 106, 108, 109
- Konfidenzintervall, 140, 141, 156, 167, 168, 174, 178, 179, 186, 190
- Konfidenzniveau, 140, 168
- konservativer Test, 101
- Kurtosis, 4, 92, 131, 139
- L-Schätzer, 44
- Least Significant Difference, 169
- letter-values, 96
- Levene-Test, 144, 155, 156, 159, 182, 183, 192
- Lillefors, 104, 108, 131, 191
- lineare Kontraste, 167, 169, 174, 176, 180, 181
- linksschief, 91, 97, 128
- linksseitiger Test, 150, 199
- Log-Transformation, 129
- LSD-Test, 169, 183, 186
- M-Schätzer, 49
- MAD, 50, 61
- MANOVA, 159
- Maskierung, 41
- matched samples, 136
- matched transformation, 129, 132
- Means Plot, 179, 183
- Median, 22, 46, 47, 49, 50, 58, 61, 91, 93, 95–97, 100, 129, 132, 139, 146, 148
- metrisch skaliert, 4, 7, 15, 17, 22, 25, 65, 72, 91, 92, 125
- midmean, 46
- midsummaries, 97
- Missing, 5, 9–11, 13, 18, 68, 138
- Missing-Werte, 5, 95
- Mittelwert, 2, 4, 34, 36, 37, 39, 41, 44, 45, 48, 79, 86, 91, 93, 104, 106, 136, 139–141, 145, 149, 173
- Modus, 91, 93, 95
- multiple Mittelwertvergleiche, 4, 164, 169, 178
- Newton-Raphson-Verfahren, 59
- nicht-skaleninvariante M-Schätzer, 49
- Nichtablehnungsbereich, 113, 124, 125, 152, 172, 195
- nichtparametrischer Test, 100, 105, 121, 200
- nominalskaliert, 2, 15, 66, 109, 115, 155
- Normal P-P Plot, 86, 89
- Normal Q-Q Plot, 85, 86, 88, 90
- Normalkern, 83
- normalverteilt, 35, 42, 61, 65, 85, 100, 142, 143, 150, 154, 156, 159
- Normalverteilung, 4, 35, 39, 40, 51, 65, 78, 79, 81, 83–85, 87, 89, 91, 92, 98, 99, 101, 103–106, 108, 111, 120, 122, 129, 131, 139, 140, 144, 150, 158, 164, 191, 193
- Nullhypothese, 39, 40, 42, 60, 100, 101, 103–105, 108, 111–114, 124, 131, 143–145, 148, 151, 152, 156–158, 161, 164, 165,

## Index

- 168, 170, 171, 173, 177, 181, 183, 192, 195  
numerische Variable, 5, 11, 12, 15, 115, 119, 154  
order statistics, 36, 44  
ordinalskaliert, 6, 15, 66, 72, 91, 109, 135  
Parallelstichproben, 136  
Parametertest, 149  
Parametervergleich, 4, 135  
parametrischer Test, 149  
Perzentile, 93–95, 97  
Point Selection Ikon, 27, 28  
Poisson-Verteilung, 106, 107, 111  
Potenztransformation, 129  
Primärerhebung, 2  
Projektion, 30, 31  
Pseudosigmas, 99  
Quantil, 85, 102, 104, 144, 145, 152, 164, 170, 172, 177, 199, 201, 203  
Quartil, 22, 47, 49, 91, 95, 106, 116, 121  
Quartilsabstand, 23  
Random sample, 10  
random sample, 41  
Rangzahlen, 85  
rechtsschief, 73, 76, 88, 91, 95, 97, 99, 128, 130–132  
rechtsseitiger Test, 150, 196  
Regressionskoeffizient, 147  
robuster Schätzer, 99  
Rotation, 31, 32  
Scatterplot-Matrix, 25, 28  
Schätzfunktion, 151  
Scheffé-Test, 176, 190  
Schiefe, 4, 91–93, 95, 97, 128, 130  
Schlüsselnummer, 14  
Sekundärdaten, 2  
Select Cases, 9–12  
Signifikanzniveau, 36, 40–42, 100, 102, 108, 109, 113, 116, 122, 144, 145, 150, 152, 154, 165, 170, 180, 195  
skaleninvariante M-Schätzer, 50  
Skalenniveau, 4, 5, 65, 91, 101, 109, 136, 159  
Skalierungsfaktor, 77  
Skewness, 92, 139  
Spread-and-level Plot, 146  
spreads, 98  
SPSS for Windows, 1, 7, 15, 18, 24  
Standardabweichung, 36, 42, 44, 50, 61, 80, 92, 104–106, 111, 139, 140, 167–169, 179  
Standardfehler, 140, 153, 156, 179, 181  
standardnormalverteilt, 144, 151, 196  
Standardnormalverteilung, 50, 86, 98, 103, 121, 124, 125, 143, 152, 172, 199  
statistischer Test, 4  
Statuszeile, 7, 12, 13  
stem width, 19, 20, 73, 76  
Stem-and-Leaf Plot, 3, 38, 72, 100, 130, 136, 137  
stetig, 76, 81, 84, 101, 110, 111, 145, 159  
Stetigkeitskorrektur, 121  
Stichprobe, 42  
Stichprobenfunktion, 44, 143, 144, 155, 208, 209  
Stichprobenmittelwert, 49, 140, 143, 150, 161, 169  
Stichprobenvariable, 142, 143, 150, 160, 195  
Stichprobenvarianz, 142–144, 153, 159, 161, 162  
Streuungsdiagramm, 25

- String-Variable, 5, 154
- Student-Newman-Keuls-Test, 174, 188
- studentisierte Variationsbreite, 168, 169, 172, 175
- Symmetrie, 91, 92, 130
- symmetrisch, 4, 44, 91, 92, 97, 99, 126, 130, 131
- Syntax Editor, 14
- SYSTAT, 34
- system missings, 5
- t-Test, 4, 155, 169, 172
- t-Verteilung, 140, 151, 153, 170–172, 183
- Teststatistik, 36, 41, 42, 60, 100, 104, 109, 111, 112, 122, 124, 142, 144, 145, 150–152, 156, 159, 161, 164, 169, 172, 174, 175, 195, 201
- Transform, 13, 14, 87, 130
- Transformation, 4, 13, 89, 103, 125, 147, 148
- Trendbereinigter Normal P-P Plot, 86, 89
- Trendbereinigter Normal Q-Q Plot, 85, 88, 90
- trimean, 47
- Tukey b-Test, 174
- Tukey's biweight, 56, 59, 61
- Tukey-Test, 172, 186, 188
- Umkodierung, 7, 9, 15, 116
- unabhängige Stichproben, 4, 135, 142, 167, 191
- unpooled variance, 153
- Varianz, 40, 79, 92, 112, 126, 136, 142, 156, 158, 163, 168, 170, 173, 174, 177, 178, 183, 192, 205–207
- Varianzanalyse, 4, 145, 159, 161, 163–165, 170, 171, 174, 177, 179, 181, 183, 205
- Varianzanalyse, 173
- Varianzheterogenität, 152, 155
- Varianzhomogenität, 149, 151, 155, 159, 161, 179, 182, 205
- Varianztabelle, 164, 183
- Variationsbreite, 173
- Verschiebungssatz, 205
- Verschmutzungsgrad, 51, 53
- Verteilungform, 65
- Verteilungsannahme, 80, 101, 112
- Verteilungsform, 100, 109
- Verteilungsfunktion, 14, 44, 49, 85, 98, 102–104, 110, 113, 122, 125, 142, 152, 190, 195, 203
- Wölbung, 92
- Wahrscheinlichkeitsplots, 84
- Wahrscheinlichkeitsplot, 4, 104
- Welch, 152
- Welch-Test, 153–155
- Winsorisierung, 46
- Zentraler Grenzwertsatz, 149, 152, 154, 156
- Zufallsstichprobe, 10, 37, 41, 49, 112, 114, 120, 135, 142, 149, 159, 195, 196
- zweiseitiger Test, 120, 156, 170, 195, 196, 199
- Zweistichproben-t-Test, 149, 151, 155, 165, 169
- Zweistichprobentest, 149, 153