IBM SPSS Statistics Base 19

Note: Before using this information and the product it supports, read the general information under Notices auf S. 326.

Vorwort

IBM® SPSS® Statistics ist ein umfassendes System zum Analysieren von Daten. Das optionale Zusatzmodul Base bietet die zusätzlichen Analyseverfahren, die in diesem Handbuch beschrieben sind. Die Prozeduren im Zusatzmodul Base müssen zusammen mit SPSS Statistics Core verwendet werden. Sie sind vollständig in dieses System integriert.

Über SPSS Inc., ein Unternehmen von IBM

SPSS Inc., ein Unternehmen von IBM, ist ein führender globaler Anbieter von Analysesoftware und -lösungen zur Prognoseerstellung. Mit der vollständigen Produktpalette des Unternehmens – Datenerfassung, Statistik, Modellierung und Bereitstellung – werden Einstellungen und Meinungen von Personen erfasst und Ergebnisse von künftigen Interaktionen mit Kunden prognostiziert. Anschließend werden diese Erkenntnisse durch die Einbettung der Analysen in Geschäftsprozesse praktisch umgesetzt. Lösungen von SPSS Inc. sind durch die Konzentration auf die Zusammenführung von Analysefunktionen, IT-Architektur und Geschäftsprozessen für zusammenhängende unternehmensübergreifende Geschäftsziele konzipiert. Kunden aus den Bereichen Wirtschaft, Regierung und Wissenschaft vertrauen weltweit auf die Technologie von SPSS Inc. als Wettbewerbsvorteil, wenn es gilt, Kunden anzuziehen, zu binden und neue Kunden zu gewinnen und dabei Betrugsfälle zu verringern und Risiken zu entschärfen. SPSS Inc. wurde im Oktober 2009 von IBM übernommen. Weitere Informationen erhalten Sie unter *http://www.spss.com*.

Technischer Support

Kunden mit Wartungsvertrag können den Technischen Support in Anspruch nehmen. Kunden können sich an den Technischen Support wenden, wenn sie Hilfe bei der Arbeit mit den Produkten von SPSS Inc. oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Wie Sie den Technischen Support kontaktieren können, entnehmen Sie der Website von SPSS Inc. unter *http://support.spss.com*. Über die Website unter *http://support.spss.com/default.asp?refpage=contactus.asp* können Sie auch nach Ihrem örtlichen Büro suchen. Wenn Sie Hilfe anfordern, halten Sie bitte Informationen bereit, um sich, Ihre Organisation und Ihren Supportvertrag zu identifizieren.

Kundendienst

Wenden Sie sich bei Fragen zur Lieferung oder Ihrem Kundenkonto an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden. Halten Sie bitte stets Ihre Seriennummer bereit.

Ausbildungsseminare

SPSS Inc. bietet öffentliche und unternehmensinterne Seminare an. Alle Seminare beinhalten auch praktische Übungen. Seminare finden in größeren Städten regelmäßig statt. Wenn Sie weitere Informationen zu diesen Seminaren wünschen, wenden Sie sich an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden.

Weitere Veröffentlichungen

Die Handbücher *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion* und *SPSS Statistics: Advanced Statistical Procedures Companion*, die von Marija Norušis geschrieben und von Prentice Hall veröffentlicht wurden, werden als Quelle für Zusatzinformationen empfohlen. Diese Veröffentlichungen enthalten statistische Verfahren in den Modulen "Statistics Base", "Advanced Statistics" und "Regression" von SPSS. Diese Bücher werden Sie dabei unterstützen, die Funktionen und Möglichkeiten von IBM® SPSS® Statistics optimal zu nutzen. Dabei ist es unerheblich, ob Sie ein Neuling im Bereich der Datenanalyse sind oder bereits über umfangreiche Vorkenntnisse verfügen und damit in der Lage sind, auch die erweiterten Anwendungen zu nutzen. Weitere Informationen zu den Inhalten der Veröffentlichungen sowie Auszüge aus den Kapiteln finden Sie auf der folgenden Autoren-Website: *http://www.norusis.com*

Inhalt

Kapitel 1

Codebuch

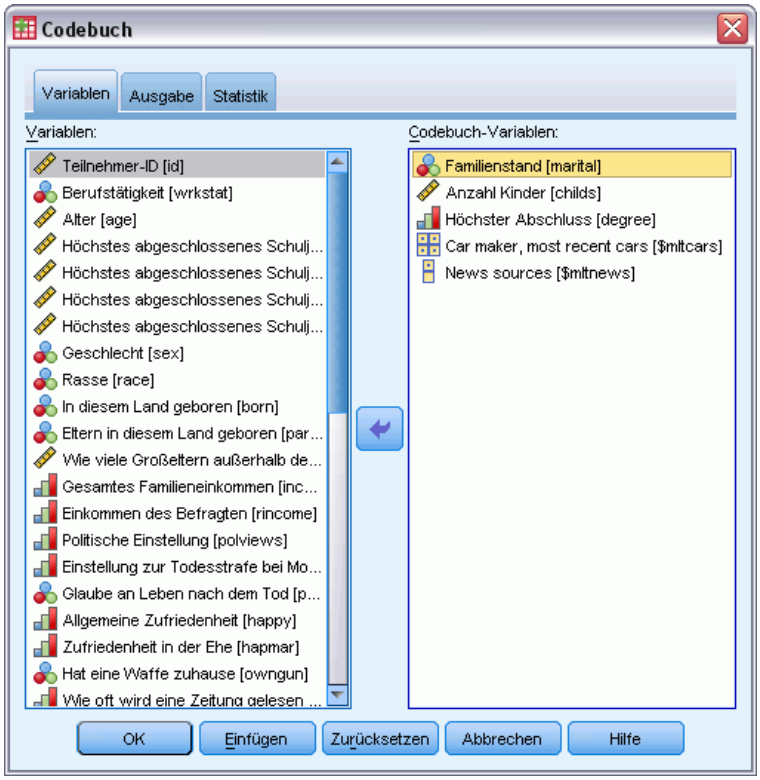
Codebuch meldet die Datenlexikoninformationen – wie Variablennamen, Variablenlabels, Wertlabels, fehlende Werte – und Auswertungsstatistiken für alle oder bestimmte Variablen und Mehrfachantworten-Sets im aktiven Daten-Set. Für nominale und ordinale Variablen und Mehrfachantworten-Sets enthalten die Auswertungsstatistiken Häufigkeiten und Prozentangaben. Für metrische Variablen enthalten die Auswertungsstatistiken Mittelwert, Standardabweichung und Quartile.

Anmerkung: Codebuch ignoriert den Aufteilungsdateistatus. Hierzu gehören Aufteilungsdateigruppen, die für die multiple Imputation von fehlenden Werten erstellt wurden (verfügbar in der Erweiterungsoption “Missing Values”).

So erhalten Sie ein Codebuch

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Berichte > Codebuch
- Klicken Sie auf die Registerkarte “Variablen”.

Abbildung 1-1
Dialog "Codebuch", Registerkarte "Variablen"



▶ Wählen Sie eine(s) oder mehrere Variablen und/oder Mehrfachantworten-Sets.

Die folgenden Optionen sind verfügbar:

- Steuern Sie die angezeigten Variablenbeschreibungen.
- Steuern Sie die angezeigten Statistiken (bzw. schließen Sie alle Auswertungsstatistiken aus).
- Steuern Sie die Reihenfolge, in der Variablen und Mehrfachantworten-Sets angezeigt werden.
- Ändern Sie das Messniveau für Variablen in der Liste der Quellvariablen, um die angezeigten Auswertungsstatistiken zu ändern. Für weitere Informationen siehe Thema Registerkarte "Codebuch-Statistiken" auf S. 5.

Ändern des Messniveaus

Sie können das Messniveau für Variablen temporär ändern. (Das Messniveau für Mehrfachantworten-Sets können Sie nicht ändern. Diese werden stets als nominal behandelt.)

▶ Klicken Sie mit der rechten Maustaste auf eine Variable in der Liste der Quellvariablen.

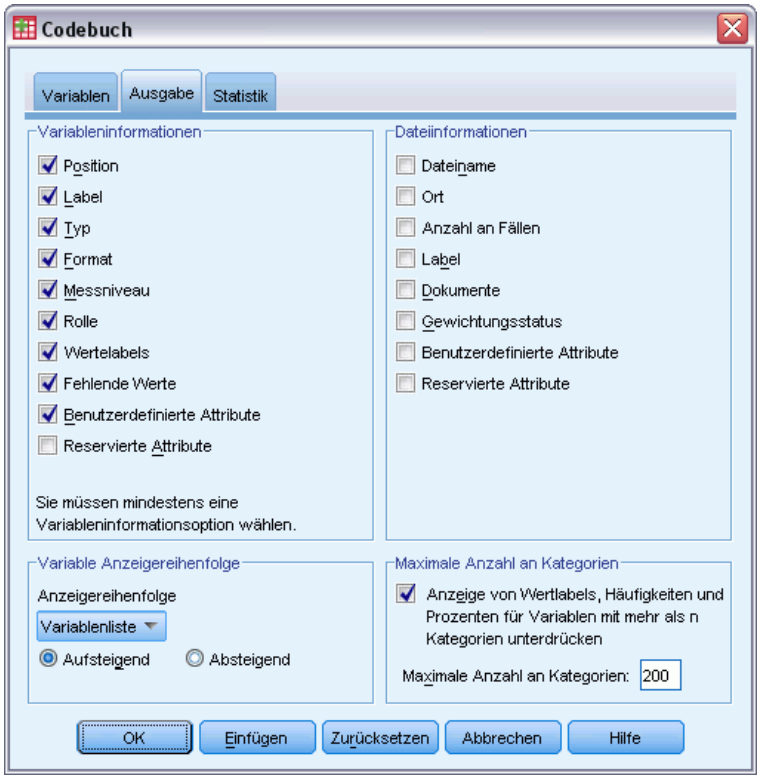
▶ Wählen Sie ein Messniveau im Kontextmenü aus.

Dies ändert das Messniveau temporär. In der Praxis ist das nur für numerische Variablen sinnvoll. Das Messniveau für String-Variablen ist auf nominal und ordinal beschränkt. Beide werden von derselben Codebuch-Prozedur behandelt.

Registerkarte "Codebuch-Ausgabe"

Die Registerkarte "Ausgabe" steuert die Variablenbeschreibungen, die für jede Variable und jedes Mehrfachantworten-Set enthalten sind, die Reihenfolge, in der die Variablen und Mehrfachantworten-Sets angezeigt werden, und den Inhalt der optionalen Dateiinformationstabelle.

Abbildung 1-2
Dialog "Codebuch", Registerkarte "Ausgabe"



Variablenbeschreibung

Dies steuert die für jede Variable angezeigten Datenlexikoninformationen.

Position. Eine Ganzzahl, die die Position der Variablen in Dateireihenfolge darstellt. Dies ist für Mehrfachantworten-Sets nicht verfügbar.

Label. Das deskriptive Label für die Variable oder das Mehrfachantworten-Set.

Typ. Grundlegender Datentyp. Entweder *Numerisch*, *String* oder *Mehrfachantworten-Set*.

Format. Das Anzeigeformat für die Variable wie *A4*, *F8.2* oder *DATE11*. Dies ist für Mehrfachantworten-Sets nicht verfügbar.

Messniveau. Die möglichen Werte sind *Nominal*, *Ordinal*, *Metrisch* und *Unbekannt*. Der angezeigte Wert ist das im Datenlexikon gespeicherte Messniveau und ist nicht von temporären Messniveauänderungen betroffen, die durch das Ändern des Messwerts in der

Quellenvariablenliste auf der Registerkarte “Variablen” angegeben werden. Dies ist für Mehrfachantworten-Sets nicht verfügbar.

Anmerkung: Das Messniveau für numerische Variablen kann vor dem ersten Datendurchlauf “unbekannt” sein, wenn das Messniveau nicht ausdrücklich festgelegt wurde, wie bei eingelesenen Daten aus einer externen Quelle oder neu erstellten Variablen.

Rolle. Einige Dialogfelder unterstützen die Vorauswahl von Variablen für Analysen basierend auf definierten Rollen.

Wertelabels. Deskriptive Labels zu spezifischen Datenwerten.

- Wenn “Häufigkeit” oder “Prozent” auf der Registerkarte “Statistik” ausgewählt ist, werden definierte Wertelabels in die Ausgabe aufgenommen, selbst wenn Sie hier “Wertelabels” nicht auswählen.
- Bei Sets aus dichotomen Variablen sind “Wertelabels” entweder die Variablenlabels für die elementaren Variablen im Set oder die Labels gezählter Werte abhängig von der Definition des Sets.

Fehlende Werte. Benutzerdefinierte fehlende Werte. Wenn “Häufigkeit” oder “Prozent” auf der Registerkarte “Statistik” ausgewählt ist, werden definierte Wertelabels in die Ausgabe aufgenommen, selbst wenn Sie hier “Fehlende Werte” nicht auswählen. Dies ist für Mehrfachantworten-Sets nicht verfügbar.

Benutzerdefinierte Attribute. Benutzerdefinierte Variablenattribute. Die Ausgabe enthält sowohl die Namen als auch die Werte für Attribute von benutzerdefinierten Variablen in Verbindung mit jeder Variable. Dies ist für Mehrfachantworten-Sets nicht verfügbar.

Reservierte Attribute. Reservierte Systemvariablen-Attribute. Sie können die Systemattribute anzeigen, Sie sollten sie aber nicht ändern. Systemattributnamen beginnen mit einem Dollarzeichen ($) . Nichtanzeige-Attribute mit Namen, die mit “@” oder “$@” beginnen, sind nicht enthalten. Die Ausgabe enthält sowohl die Namen als auch die Werte für Systemattribute in Verbindung mit jeder Variable. Dies ist für Mehrfachantworten-Sets nicht verfügbar.

Dateiinformationen

Die optionale Dateiinformationentabelle kann beliebige der folgenden Dateiattribute enthalten:

Dateiname: Name der IBM® SPSS® Statistics-Datendatei. Wenn das Daten-Set nie in SPSS Statistics-Format gespeichert wurde, gibt es keinen Datendateinamen. (Wenn in der Titelleiste des Fensters “Daten-Editor” kein Dateiname angezeigt wird, hat das aktive Daten-Set keinen Dateinamen.)

Ort. Verzeichnis (Ordner) der SPSS Statistics-Datendatei. Wenn das Daten-Set nie in SPSS Statistics-Format gespeichert wurde, gibt es keinen Speicherort.

Anzahl der Fälle. Die Anzahl der Fälle im aktiven Daten-Set. Das ist die Gesamtzahl an Fällen, einschließlich der Fälle, die aufgrund von Filterbedingungen aus Auswertungsstatistiken ausgeschlossen werden können.

Label. Dies ist das Dateilabel (falls vorhanden), definiert durch den Befehl `FILE LABEL`.

Dokumente. Datendatei-Dokumenttext.

Gewichtungsstatus. Bei eingeschalteter Gewichtung wird der Name der Gewichtungsvariablen angezeigt.

Benutzerdefinierte Attribute. Benutzerdefinierte Datendateiattribute. Datendateiattribute, definiert durch den Befehl `DATAFILE ATTRIBUTE`.

Reservierte Attribute. Reservierte Systemdatendateiattribute. Sie können die Systemattribute anzeigen, Sie sollten sie aber nicht ändern. Systemattributnamen beginnen mit einem Dollarzeichen ($) . Nichtanzeige-Attribute mit Namen, die mit "@" oder "$@" beginnen, sind nicht enthalten. Die Ausgabe enthält sowohl die Namen als auch die Werte für Systemdatendateiattribute.

Variable Anzeigereihenfolge

Die folgenden Alternativen stehen zur Verfügung, um die Reihenfolge, in der Variablen und Mehrfachantworten-Sets angezeigt werden, zu steuern.

Alphabetisch. Alphabetische Reihenfolge nach Variablenname.

Datei. Die Reihenfolge, in der die Variablen im Daten-Set erscheinen (die Reihenfolge, in der sie im Daten-Editor angezeigt werden). In aufsteigender Reihenfolge werden Merfachantworten-Sets zuletzt nach allen ausgewählten Variablen angezeigt.

Messniveau. Nach Messniveau sortieren. Erstellt vier Sortiergruppen: nominal, ordinal, metrisch und unbekannt. Mehrfachantworten-Sets werden als nominal behandelt.

Anmerkung: Das Messniveau für numerische Variablen kann vor dem ersten Datendurchlauf "unbekannt" sein, wenn das Messniveau nicht ausdrücklich festgelegt wurde, wie bei eingelesenen Daten aus einer externen Quelle oder neu erstellten Variablen.

Variablenliste. Die Reihenfolge, in der Variablen und Mehrfachantworten-Sets in der ausgewählten Variablenliste in der Registerkarte "Variablen" angezeigt werden.

Benutzerdefinierter Attributname. Die Liste der Sortierfolgeoptionen umfasst ferner die Namen der benutzerdefinierten Variablenattribute. Bei aufsteigender Reihenfolge werden Variablen, die das Attribut nicht besitzen, nach oben sortiert, gefolgt von den Variablen, die das Attribut, aber keinen definierten Wert für das Attribut besitzen, gefolgt von Variablen mit definierten Werten für das Attribut in alphabetischer Reihenfolge der Werte.

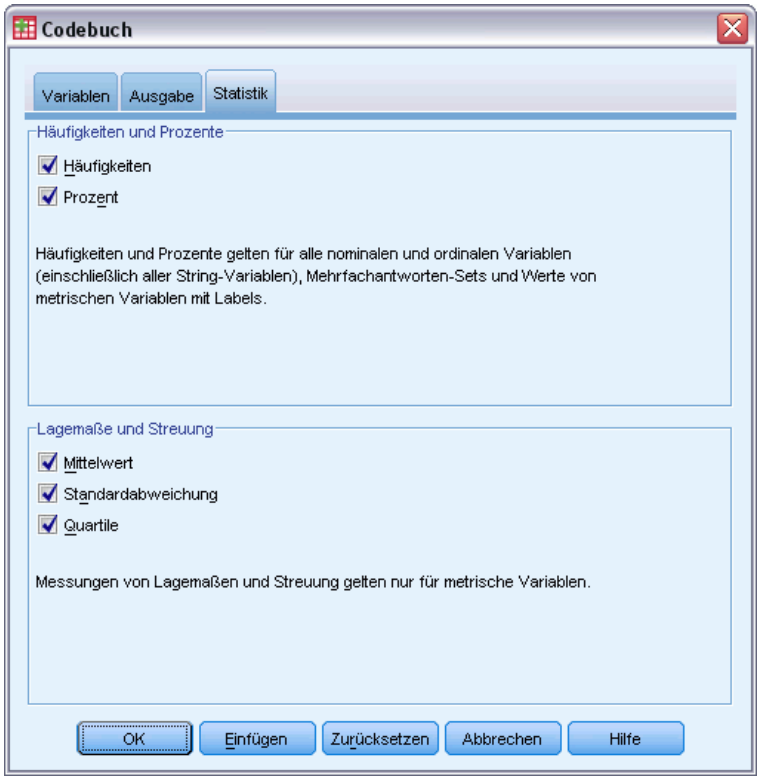
Maximale Anzahl an Kategorien

Wenn die Ausgabe Wertelabels, Häufigkeiten oder Prozentangaben für jeden eindeutigen Wert enthält, können Sie diese Informationen von der Tabelle unterdrücken, wenn die Anzahl an Werten den angegebenen Wert überschreitet. Standardmäßig werden diese Informationen unterdrückt, wenn die Anzahl der eindeutigen Werte für die Variable 200 überschreitet.

Registerkarte "Codebuch-Statistiken"

Über die Registerkarte "Statistik" können Sie die Auswertungsstatistiken steuern, die in die Ausgabe aufgenommen werden, oder die Anzeige von Auswertungsstatistiken komplett unterdrücken.

Abbildung 1-3
Dialog "Codebuch", Registerkarte "Statistik"



Häufigkeiten und Prozente

Für nominale und ordinale Variablen, Mehrfachantworten-Sets und Werte von metrischen Variablen mit Labels sind folgende Statistiken verfügbar:

Anzahl. Die Anzahl der Fälle, die für eine Variable einen bestimmten Wert (oder Wertebereich) aufweisen.

Prozent. Der Prozentsatz der Fälle mit einem bestimmten Wert.

Lagemaße und Streuung

Für metrische Variablen sind folgende Statistiken verfügbar:

Mittelwert. Ein Lagemaß. Die Summe der Ränge, geteilt durch die Zahl der Fälle.

Standardabweichung. Ein Maß für die Streuung um den Mittelwert. In einer Normalverteilung liegen 68% der Fälle innerhalb von einer Standardabweichung des Mittelwerts und 95% der Fälle innerhalb von zwei Standardabweichungen. Wenn beispielsweise für das Alter der Mittelwert 45 und die Standardabweichung 10 beträgt, liegen bei einer Normalverteilung 95 % der Fälle im Bereich zwischen 25 und 65.

Quartile. Zeigt die Werte des 25., 50. und 75. Perzentils an.

Anmerkung: Sie können das Messniveau für eine Variable temporär (und so die für diese Variable angezeigte Auswertungsstatistik) in der Quellenvariablenliste auf der Registerkarte “Variablen” ändern.

Kapitel
2

Häufigkeiten

Die Prozedur “Häufigkeiten” stellt Statistiken und grafische Darstellungen für die Beschreibung vieler Variablentypen zur Verfügung. Die Prozedur “Häufigkeiten” ist ein guter Ausgangspunkt für die Betrachtung Ihrer Daten.

Bei Häufigkeitsberichten und Balkendiagrammen können Sie die unterschiedlichen Werte in aufsteigender oder absteigender Reihenfolge anordnen oder die Kategorien nach deren Häufigkeiten ordnen. Der Häufigkeitsbericht kann unterdrückt werden, wenn für eine Variable viele unterschiedliche Werte vorhanden sind. Sie können Diagramme mit Häufigkeiten (die Standardeinstellung) oder Prozentsätzen beschriften.

Beispiel. Wie sind die Kunden eines Unternehmens nach Industriezweigen verteilt? Sie können aus Ihren Ausgabedaten ersehen, dass 37,5 % Ihrer Kunden zu staatlichen Behörden gehören, 24,9 % zu Unternehmen der freien Wirtschaft, 28,1 % zu akademischen Institutionen und 9,4 % zum Gesundheitswesen. Bei stetigen quantitativen Daten wie Verkaufserlösen könnten Sie beispielsweise ersehen, dass sich der durchschnittliche Produktverkauf auf $3.576 bei einer Standardabweichung von $1.078 beläuft.

Statistiken und Diagramme.Häufigkeiten, Prozentsätze, kumulierte Prozentsätze, Mittelwert, Median, Modalwert, Summe, Standardabweichung, Varianz, Spannweite, Minimum und Maximum, Standardfehler des Mittelwerts, Schiefe und Kurtosis (beide mit Standardfehler), Quartile, benutzerdefinierte Perzentile, Balkendiagramme, Kreisdiagramme und Histogramme.

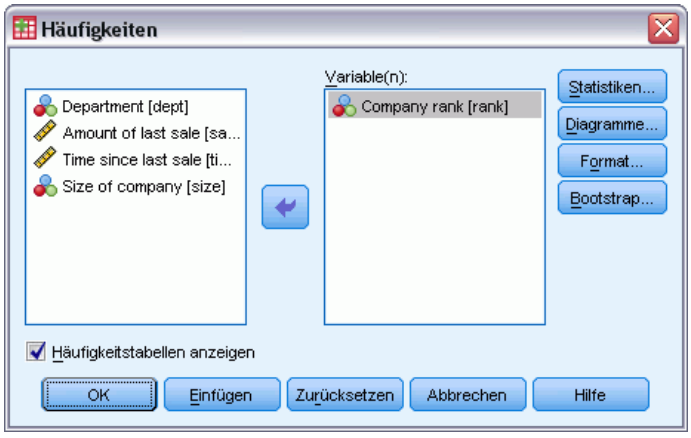
Daten. Verwenden Sie zum Kodieren kategorialer Variablen (nominales oder ordinales Messniveau) numerische Codes oder Strings.

Annahmen. Die Tabellen und Prozentsätze stellen nützliche Beschreibungen für Daten aus allen Verteilungen zur Verfügung, insbesondere für Variablen mit geordneten oder ungeordneten Kategorien. Die meisten der optionalen Auswertungsstatistiken, wie zum Beispiel der Mittelwert und die Standardabweichung, gehen von der Normalverteilung aus und können auf quantitative Variablen mit symmetrischen Verteilungen angewendet werden. Robuste Statistiken, wie zum Beispiel Median, Quartile und Perzentile, sind für quantitative Variablen geeignet, die nur möglicherweise die Annahme erfüllen, dass eine Normalverteilung gilt.

So erstellen Sie Häufigkeitstabellen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Deskriptive Statistiken > Häufigkeiten...

Abbildung 2-1
Hauptdialogfeld von "Häufigkeiten"



- Wählen Sie mindestens eine kategoriale oder quantitative Variable aus.

Die folgenden Optionen sind verfügbar:

- Deskriptive Statistiken für quantitative Variablen erhalten Sie, indem Sie auf Statistik klicken.
- Balkendiagramme, Kreisdiagramme oder Histogramme erhalten Sie, indem Sie auf Diagramme klicken.
- Sie können die Reihenfolge der angezeigten Ergebnisse ändern, indem Sie auf Format klicken.

Häufigkeiten: Statistik

Abbildung 2-2
Dialogfeld "Häufigkeiten: Statistik"

Perzentilwerte. Dies sind Werte einer quantitativen Variablen, welche die geordneten Daten in Gruppen unterteilen, sodass ein bestimmter Prozentsatz darüber und ein bestimmter Prozentsatz darunter liegt. Quartile (die 25., 50. und 75. Perzentile) unterteilen die Beobachtung in vier gleich große Gruppen. Falls Sie eine gleiche Anzahl von Gruppen wünschen, die von vier abweicht, klicken Sie auf Trennen und geben Sie eine Anzahl für “gleiche Gruppen” ein. Sie können auch individuelle Perzentile festlegen (zum Beispiel das 95. Perzentil, also der Wert, unter dem 95 % der Beobachtungen liegen).

Lagemaße. Statistiken, welche die Lage der Verteilung beschreiben, sind Mittelwert, Median, Modalwert und Summe aller Werte.

- **Mittelwert.** Ein Lagemaß. Die Summe der Ränge, geteilt durch die Zahl der Fälle.
- **Median.** Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).
- **Modalwert.** Der am häufigsten auftretende Wert. Wenn mehrere Werte gleichermaßen die größte Häufigkeit aufweisen, ist jeder von ihnen ein Modalwert. Die Prozedur "Häufigkeiten" meldet bei mehreren Modalwerten nur den kleinsten.
- **Summe.** Die Summe der Werte über alle Fälle mit nichtfehlenden Werten.

Streuung. Statistiken, welche die Menge an Variation oder die Streubreite in den Daten messen, sind Standardabweichung, Varianz, Spannweite, Minimum, Maximum und Standardfehler des Mittelwerts.

- **Standardabweichung.** Ein Maß für die Streuung um den Mittelwert. In einer Normalverteilung liegen 68% der Fälle innerhalb von einer Standardabweichung des Mittelwerts und 95% der Fälle innerhalb von zwei Standardabweichungen. Wenn beispielsweise für das Alter der Mittelwert 45 und die Standardabweichung 10 beträgt, liegen bei einer Normalverteilung 95 % der Fälle im Bereich zwischen 25 und 65.
- **Varianz.** Ein Maß der Streuung um den Mittelwert. Es ist gleich dem Quotienten aus der Summe der quadrierten Abweichung vom Mittelwert und der um 1 verringerten Fallanzahl. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.
- **Spannweite.** Die Differenz zwischen den größten und kleinsten Werten einer numerischen Variablen; Maximalwert minus Minimalwert.
- **Minimum.** Der kleinste Wert einer numerischen Variablen.
- **Maximum.** Der größte Wert einer numerischen Variablen.
- **Standardfehler des Mittelwerts.** Ein Maß für die mögliche Variation des Mittelwerts zwischen aus derselben Verteilung stammenden Stichproben. Dieser Wert kann für einen ungefähren Vergleich des beobachteten Mittelwerts mit einem hypothetischen Wert verwendet werden. (Es kann geschlossen werden, dass die beiden Werte unterschiedlich sind, wenn das Verhältnis der Differenz zum Standardfehler kleiner als -2 oder größer als +2 ist.)

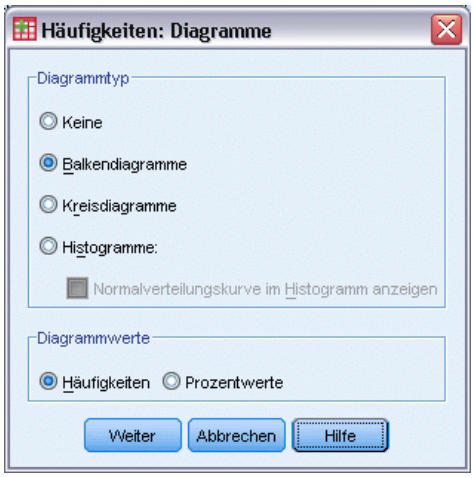
Verteilung. Schiefe und Kurtosis sind Statistiken, die Form und Symmetrie der Verteilung beschreiben. Diese Statistiken werden mit ihren Standardfehlern angezeigt.

- **Schiefe.** Ein Maß für die Asymmetrie einer Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefe-Wert, der mehr als doppelt so groß ist wie sein Standardfehler, für eine Abweichung von der Symmetrie spricht.
- **Kurtosis.** Ein Maß dafür, wie sich die Beobachtungen um einen zentralen Punkt gruppieren. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremverten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.

Werte sind Gruppenmittelpunkte. Falls die Werte in den Daten Gruppenmittelpunkte sind (wenn zum Beispiel das Alter aller Personen in den Dreißigern mit dem Wert 35 kodiert ist), wählen Sie diese Option, um den Median und das Perzentil für die ursprünglichen, nicht gruppierten Daten berechnen zu lassen.

Häufigkeiten: Diagramme

Abbildung 2-3
Dialogfeld "Häufigkeiten: Diagramme"



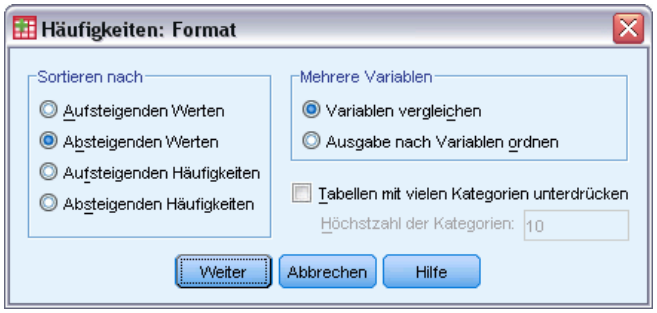
Diagrammtyp. In einem Kreisdiagramm wird der Anteil der Teile an einem Ganzen angezeigt. Jedes Segment eines Kreisdiagramms entspricht einer durch eine einzelne Gruppenvariable definierten Gruppe. In einem Balkendiagramm wird die Anzahl für jeden unterschiedlichen Wert oder jede unterschiedliche Kategorie als separater Balken angezeigt, wodurch Sie Kategorien visuell vergleichen können. Auch Histogramme enthalten Balken, diese sind jedoch an einer Skala mit gleichen Abständen ausgerichtet. Die Höhe jedes Balkens gibt die Anzahl der Werte einer quantitativen Variablen wieder, die innerhalb des Intervalls liegen. In einem Histogramm werden

Form, Mittelpunkt und die Streubreite der Verteilung angezeigt. Eine über das Histogramm gelegte Normalverteilungskurve erleichtert die Beurteilung, ob die Daten normalverteilt sind.

Diagrammwerte. Bei Balkendiagrammen kann die Skalenachse mit Häufigkeiten oder Prozentwerten beschriftet werden.

Häufigkeiten: Format

Abbildung 2-4
Dialogfeld "Häufigkeiten: Format"



Sortieren nach. Die Häufigkeitstabelle kann entsprechend den tatsächlichen Werten der Daten oder entsprechend der Anzahl (Häufigkeit des Vorkommens) dieser Werte geordnet werden. Die Tabelle kann entweder in aufsteigender oder in absteigender Reihenfolge angeordnet werden. Wenn Sie allerdings ein Histogramm oder Perzentile anfordern, wird in der Prozedur "Häufigkeiten" davon ausgegangen, dass die Variable quantitativ ist. Die Werte werden dann in aufsteigender Reihenfolge angezeigt.

Mehrere Variablen. Wenn Sie Statistiktabellen für multiple Variablen erzeugen, können Sie entweder alle Variablen in einer einzigen Tabelle (Variablen vergleichen) oder eine eigene Statistiktabelle für jede Variable (Ausgabe nach Variablen ordnen) anzeigen.

Keine Tabellen mit mehr als n Kategorien. Diese Option verhindert die Anzeige von Tabellen mit mehr als der angegebenen Anzahl von Werten.

Kapitel
3

Deskriptive Statistik

Mit der Prozedur "Deskriptive Statistiken" werden in einer einzelnen Tabelle univariate Auswertungsstatistiken für verschiedene Variablen angezeigt und standardisierte Werte (*Z*-Werte) errechnet. Variablen können folgendermaßen geordnet werden: nach der Größe ihres Mittelwerts (in aufsteigender oder absteigender Reihenfolge), alphabetisch oder in der Reihenfolge, in der sie ausgewählt wurden (dies ist die Standardeinstellung).

Wenn *Z*-Werte gespeichert werden, werden sie zu den Daten im Daten-Editor hinzugefügt und stehen dann für IBM SPSS Statistics-Diagramme, Auflistungen von Daten und Analysen zur Verfügung. Wenn Variablen in verschiedenen Einheiten aufgezeichnet werden (zum Beispiel Bruttoinlandsprodukt pro Kopf der Bevölkerung und Prozentsatz der Alphabetisierung), werden die Variablen durch eine *Z*-Wert-Transformation zur Erleichterung des visuellen Vergleichs auf einer gemeinsamen Skala angeordnet.

Beispiel. Sie zeichnen über mehrere Monate den täglichen Umsatz jedes einzelnen Angestellten der Verkaufsabteilung auf (z. B. ein Eintrag für Herbert, ein Eintrag für Sabine und ein Eintrag für Joachim), sodass jeder Fall in Ihren Daten den täglichen Umsatz jedes Angestellten enthält. Mit der Prozedur "Deskriptive Statistik" wird für Sie jetzt der durchschnittliche Tagesumsatz der einzelnen Angestellten berechnet und das Ergebnis vom höchsten durchschnittlichen Umsatz zum niedrigsten durchschnittlichen Umsatz geordnet.

Statistiken. Stichprobengröße, Mittelwert, Minimum, Maximum, Standardabweichung, Varianz, Spannweite, Summe, Standardfehler des Mittelwerts und Kurtosis und Schiefe mit den Standardfehlern.

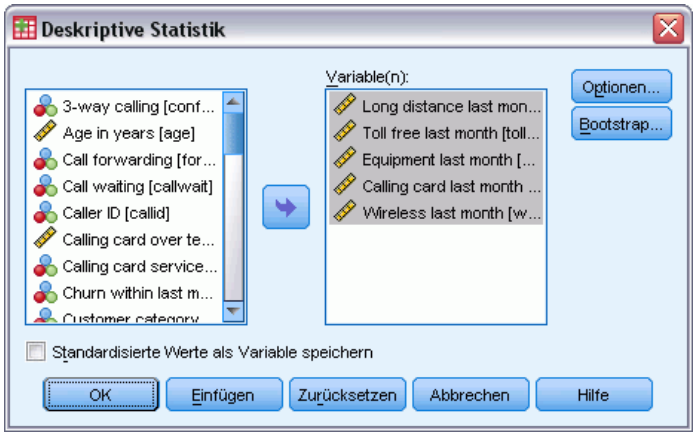
Daten. Verwenden Sie numerische Variablen, nachdem Sie diese im Diagramm auf Aufzeichnungsfehler, Ausreißer und Unregelmäßigkeiten in der Verteilung untersucht haben. Die Prozedur "Deskriptive Statistiken" ist für große Dateien (mit Tausenden von Fällen) besonders effektiv.

Annahmen. Die meisten verfügbaren Statistiken (einschließlich *Z*-Werte) basieren auf der Annahme, dass die Daten normalverteilt sind, und sind für quantitative Variablen (mit Intervall- oder Verhältnis-Messniveau) mit symmetrischen Verteilungen geeignet. Vermeiden Sie Variablen mit ungeordneten Kategorien oder schiefen Verteilungen. Die Verteilung der *Z*-Werte hat dieselbe Form wie die ursprünglichen Daten; daher bietet das Berechnen von *Z*-Werten keine Abhilfe bei problematischen Daten.

So lassen Sie deskriptive Statistiken berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Deskriptive Statistiken > Deskriptive Statistik...

Abbildung 3-1
Dialogfeld "Deskriptive Statistik"



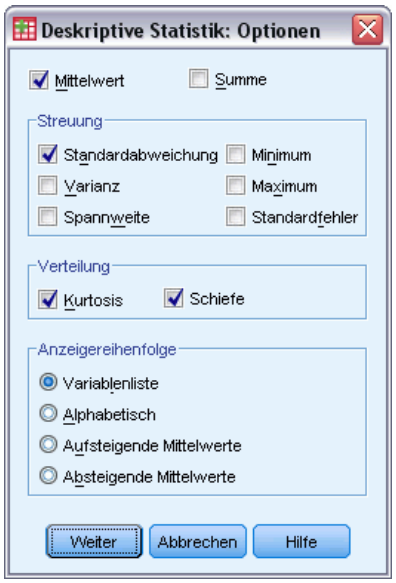
- Wählen Sie mindestens eine Variable aus.

Die folgenden Optionen sind verfügbar:

- Wählen Sie Standardisierte Werte als Variable speichern, um Z-Werte als neue Variablen zu speichern.
- Optionale Statistiken und die Reihenfolge der Anzeige steuern Sie, indem Sie auf Optionen klicken.

Deskriptive Statistik: Optionen

Abbildung 3-2
Dialogfeld "Deskriptive Statistik: Optionen"



Mittelwert und Summe. In der Standardeinstellung wird der Mittelwert bzw. das arithmetische Mittel angezeigt.

Streuung. Zu den Statistiken, welche die Streubreite oder die Variation in den Daten messen, gehören Standardabweichung, Varianz, Spannweite, Minimum, Maximum und Standardfehler des Mittelwerts.

- **Standardabweichung.** Ein Maß für die Streuung um den Mittelwert. In einer Normalverteilung liegen 68% der Fälle innerhalb von einer Standardabweichung des Mittelwerts und 95% der Fälle innerhalb von zwei Standardabweichungen. Wenn beispielsweise für das Alter der Mittelwert 45 und die Standardabweichung 10 beträgt, liegen bei einer Normalverteilung 95 % der Fälle im Bereich zwischen 25 und 65.
- **Varianz.** Ein Maß der Streuung um den Mittelwert. Es ist gleich dem Quotienten aus der Summe der quadrierten Abweichung vom Mittelwert und der um 1 verringerten Fallanzahl. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.
- **Spannweite.** Die Differenz zwischen den größten und kleinsten Werten einer numerischen Variablen; Maximalwert minus Minimalwert.
- **Minimum.** Der kleinste Wert einer numerischen Variablen.
- **Maximum.** Der größte Wert einer numerischen Variablen.
- **Standardfehler.** Ein Maß für die mögliche Variation des Mittelwerts zwischen aus derselben Verteilung stammenden Stichproben. Dieser Wert kann für einen ungefähren Vergleich des beobachteten Mittelwerts mit einem hypothetischen Wert verwendet werden. (Es kann geschlossen werden, dass die beiden Werte unterschiedlich sind, wenn das Verhältnis der Differenz zum Standardfehler kleiner als -2 oder größer als +2 ist.)

Verteilung. Kurtosis und Schiefe sind Statistiken, die Form und Symmetrie der Verteilung charakterisieren. Diese Statistiken werden mit ihren Standardfehlern angezeigt.

- **Kurtosis.** Ein Maß dafür, wie sich die Beobachtungen um einen zentralen Punkt gruppieren. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremverten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.
- **Schiefe.** Ein Maß für die Asymmetrie einer Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefe-Wert, der mehr als doppelt so groß ist wie sein Standardfehler, für eine Abweichung von der Symmetrie spricht.

Anzeigereihenfolge. In der Standardeinstellung werden die Variablen in der Reihenfolge angezeigt, in der sie ausgewählt wurden. Sie können Variablen bei Bedarf in alphabetischer Reihenfolge mit aufsteigend oder absteigend geordneten Mittelwerten anzeigen lassen.

Zusätzliche Funktionen beim Befehl DESCRIPTIVES

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Sie können die standardisierten Werte (*Z*-Werte) selektiv für einige Variablen speichern (mit dem Unterbefehl `VARIABLES`).
- Sie können Namen für die neuen Variablen angeben, die die standardisierte Werte enthalten (mit dem Unterbefehl `VARIABLES`).
- Sie können Fälle mit fehlenden Werten in einer beliebigen Variablen aus der Analyse ausschließen (mit dem Unterbefehl `MISSING`).
- Sie können die Variablen in der Anzeige nach dem Wert einer beliebigen Statistik, nicht nur nach dem Mittelwert sortieren (mit dem Unterbefehl `SORT`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
4

Explorative Datenanalyse

Mit der Prozedur “Explorative Datenanalyse” werden Auswertungsstatistiken und grafische Darstellungen für alle Fälle oder für separate Fallgruppen erzeugt. Es kann viele Gründe für die Verwendung der Prozedur “Explorative Datenanalyse” geben: Sichten von Daten, Erkennen von Ausreißern, Beschreibung, Überprüfung der Annahmen und Charakterisieren der Unterschiede zwischen Teilgrundgesamtheiten (Fallgruppen). Beim Sichten der Daten können Sie ungewöhnliche Werte, Extremwerte, Lücken in den Daten oder andere Auffälligkeiten erkennen. Durch die explorative Datenanalyse können Sie sich vergewissern, ob die für die Datenanalyse vorgesehenen statistischen Methoden geeignet sind. Die Untersuchung kann ergeben, dass Sie die Daten transformieren müssen, falls die Methode eine Normalverteilung erfordert. Sie können sich stattdessen auch für die Verwendung nichtparametrischer Tests entscheiden.

Beispiel. Betrachten Sie die Verteilung der Lernzeiten für Ratten im Labyrinth mit vier verschiedenen Schwierigkeitsgraden. Zu jeder der vier Gruppen können Sie ablesen, ob die Zeiten annähernd normalverteilt und die vier Varianzen gleich sind. Sie können auch die Fälle mit den fünf längsten und den fünf kürzesten Zeiten bestimmen. Sie können die Verteilung der Lernzeiten für jede Gruppe mit Boxplots und Stengel-Blatt-Diagrammen grafisch auswerten.

Statistiken und Diagramme. Mittelwert, Median, 5% getrimmtes Mittel, Standardfehler, Varianz, Standardabweichung, Minimum, Maximum, Spannweite, interquartiler Bereich, Schiefe und Kurtosis und deren Standardfehler, Konfidenzintervall für den Mittelwert (und angegebenes Konfidenzniveau), Perzentile, M-Schätzer nach Huber, Andrew-Wellen-Schätzer, M-Schätzer nach Hampel, Tukey-Biweight-Schätzer, die fünf größten und die fünf kleinsten Werte, die Kolmogorov-Smirnov-Statistik mit Lilliefors-Signifikanzniveau zum Prüfen der Normalverteilung und die Shapiro-Wilk-Statistik. Boxplots, Stengel-Blatt-Diagramme, Histogramme, Normalverteilungsdiagramme und Diagramme der Streubreite gegen das mittlere Niveau mit Levene-Test und Transformationen.

Daten. Die Prozedur “Explorative Datenanalyse” kann für quantitative Variablen (mit Intervall- oder Verhältnis-Messniveau) verwendet werden. Eine Faktorvariable (zum Aufteilen der Daten in Fallgruppen) muss eine sinnvolle Anzahl von unterschiedlichen Werten (Kategorien) enthalten. Diese Werte können kurze Strings oder numerische Werte sein. Die Fallbeschriftungsvariable, die für die Beschriftung von Ausreißern in Boxplots verwendet wird, kann ein kurzer String, ein langer String (die ersten 15 Byte) oder numerisch sein.

Annahmen. Ihre Daten müssen nicht symmetrisch oder normalverteilt sein.

So führen Sie eine explorative Datenanalyse aus:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Deskriptive Statistiken > Explorative Datenanalyse...

Abbildung 4-1
Dialogfeld "Explorative Datenanalyse"

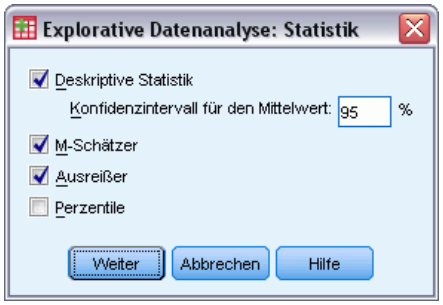
- Wählen Sie eine oder mehrere abhängige Variablen aus.

Die folgenden Optionen sind verfügbar:

- Auswählen einer oder mehrerer Faktorvariablen, mit deren Werten Fallgruppen definiert werden.
- Auswählen einer Identifizierungsvariablen für die Beschriftung von Fällen.
- Zugriff auf robuste Schätzer, Ausreißer, Perzentile und Häufigkeitstabellen erhalten Sie, indem Sie auf Statistik klicken.
- Zugriff auf Histogramme, Normalverteilungsdiagramme und Tests sowie Diagramme der Streubreite gegen das mittlere Niveau mit Levene-Statistik erhalten Sie, indem Sie auf Diagramme klicken.
- Sie können die Behandlung fehlender Werte festlegen, indem Sie auf Optionen klicken.

Explorative Datenanalyse: Statistik

Abbildung 4-2
Dialogfeld "Explorative Datenanalyse: Statistik"



Deskriptive Statistiken. In der Standardeinstellung werden Lage- und Streuungsmaße angezeigt. Mit den Lagemaßen wird die Lage der Verteilung angegeben. Dazu gehören Mittelwert, Median und 5% getrimmtes Mittel. Mit den Maßen für Streuung werden Unähnlichkeiten der Werte

angezeigt. Diese umfassen Standardfehler, Varianz, Standardabweichung, Minimum, Maximum, Spannweite und den Interquartilbereich. Die beschreibenden Statistiken enthalten auch Maße der Verteilungsform. Schiefe und Kurtosis werden mit den jeweiligen Standardfehlern angezeigt. Das 95%-Konfidenzintervall für den Mittelwert wird ebenfalls angezeigt. Sie können auch ein anderes Konfidenzniveau angeben.

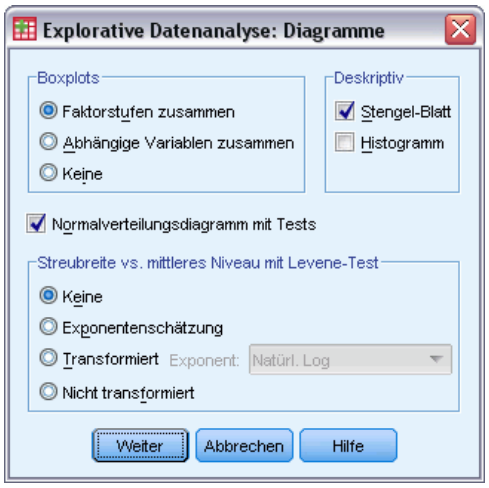
M-Schätzer. Robuste Alternativen zu Mittelwert und Median der Stichprobe zum Schätzen der Lage. Die berechneten Schätzer unterscheiden sich in den Gewichtungen, die sie den Fällen zuweisen. M-Schätzer nach Huber, Andrew-Wellen-Schätzer, M-Schätzer nach Hampel und Tukey-Biweight-Schätzer werden angezeigt.

Ausreißer. Hier werden die fünf größten und die fünf kleinsten Werte mit Fallbeschriftungen angezeigt.

Perzentile. Hier werden die Werte für die 5., 10., 25., 50., 75., 90. und 95. Perzentile angezeigt.

Explorative Datenanalyse: Diagramme

Abbildung 4-3
Dialogfeld "Explorative Datenanalyse: Diagramme"



Boxplots. Mit diesen Optionen legen Sie fest, wie Boxplots bei mehr als einer abhängigen Variablen angezeigt werden. Mit Faktorstufen zusammen wird eine getrennte Anzeige für jede abhängige Variable erzeugt. In einer Anzeige werden Boxplots für alle durch eine Faktorvariable definierten Gruppen angezeigt. Mit Abhängige Variablen zusammen wird für jede durch eine Faktorvariable definierte Gruppe eine getrennte Anzeige erzeugt. In einer Anzeige werden Boxplots für alle abhängigen Variablen in einer Anzeige nebeneinander dargestellt. Diese Anzeige ist insbesondere nützlich, wenn verschiedene Variablen ein einziges, zu unterschiedlichen Zeiten gemessenes Merkmal darstellen.

Deskriptive Statistik. Im Gruppenfeld "Deskriptive Statistik" können Sie Stengel-Blatt-Diagramme und Histogramme auswählen.

Normalverteilungsdiagramme mit Tests. Hier werden Normalverteilungsdiagramme und trendbereinigte Normalverteilungsdiagramme angezeigt. Die Kolmogorov-Smirnov-Statistik mit einem Signifikanzniveau nach Lilliefors für den Test auf Normalverteilung wird angezeigt.

Bei Angabe von nichtganzzahligen Gewichtungen wird die Shapiro-Wilk-Statistik berechnet, wenn die gewichtete Stichprobengröße zwischen 3 und 50 liegt. Bei keinen oder ganzzahligen Gewichtungen wird die Statistik berechnet, wenn die gewichtete Stichprobengröße zwischen 3 und 5,000 liegt.

Streubreite vs. mittleres Niveau mit Levene-Test. Hiermit legen Sie fest, wie Daten für Diagramme der Streubreite versus mittleres Niveau transformiert werden. Für alle Diagramme der Streubreite versus mittleres Niveau werden die Steigung der Regressionsgeraden und der Levene-Test auf Homogenität der Varianz angezeigt. Wenn Sie eine Transformation auswählen, liegen dem Levene-Test die transformierten Daten zugrunde. Wenn keine Faktorvariable ausgewählt wurde, werden keine Diagramme der Streubreite versus mittleres Niveau erstellt. Mit der Exponentenschätzung wird ein Diagramm der natürlichen Logarithmen des Interquartilbereichs über die natürlichen Logarithmen des Medians für alle Zellen sowie eine Schätzung der Potenztransformation zum Erreichen gleicher Varianzen in den Zellen angefordert. Mit Diagrammen der Streubreite versus mittleres Niveau lässt sich der Exponent für Transformationen bestimmen, mit denen über Gruppen hinweg eine höhere Stabilität (höhere Gleichförmigkeit) der Varianzen erreicht wird. Mit Transformiert können Sie einen alternativen Exponenten auswählen, eventuell gemäß der Empfehlung der Exponentenschätzung, und Diagramme der transformierten Daten erzeugen. Der Interquartilbereich und der Median der transformierten Daten werden grafisch dargestellt. Mit Nicht transformiert werden Diagramme der Rohdaten erstellt. Dies entspricht einer Transformation mit einem Exponenten gleich 1.

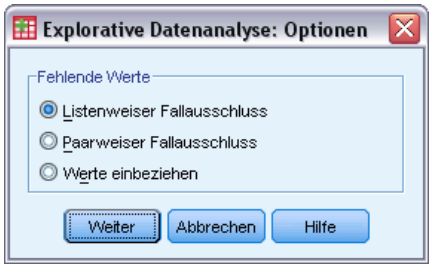
Explorative Datenanalyse: Potenztransformationen

Dies sind die Potenztransformationen für Diagramme der Streubreite versus mittleres Niveau. Für die Transformation von Daten muss ein Exponent ausgewählt werden. Sie können eine der folgenden Möglichkeiten wählen:

- **Natürlicher Logarithmus.**Transformation mit natürlichem Logarithmus. Dies ist die Standardeinstellung.
- **1/Quadratwurzel.**Zu jedem Datenwert wird der reziproke Wert der Quadratwurzel berechnet.
- **Reziprok.**Der reziproke Wert jedes Datenwerts wird berechnet.
- **Quadratwurzel.**Die Quadratwurzel jedes Datenwerts wird berechnet.
- **Quadratisch.**Jeder Datenwert wird quadriert.
- **Kubisch.**Es wird die dritte Potenz jedes Datenwerts errechnet.

Explorative Datenanalyse: Optionen

Abbildung 4-4
Dialogfeld "Explorative Datenanalyse: Optionen"



Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für abhängige Variablen oder Faktorvariablen werden aus allen Analysen ausgeschlossen. Dies ist die Standardeinstellung.
- **Paarweiser Fallausschluss.**Fälle ohne fehlenden Werte für Variablen in einer Gruppe (Zelle) werden in die Analyse dieser Gruppe einbezogen. Der Fall kann fehlende Werte für Variablen enthalten, die in anderen Gruppen verwendet werden.
- **Werte einbeziehen.**Fehlende Werte für Faktorvariablen werden als gesonderte Kategorie behandelt. Die gesamte Ausgabe wird auch für diese zusätzliche Kategorie erstellt. Häufigkeitstabellen enthalten Kategorien für fehlende Werte. Fehlende Werte für Faktorvariablen werden aufgenommen, jedoch als fehlend beschriftet.

Zusätzliche Funktionen beim Befehl EXAMINE

In der Prozedur "Explorative Datenanalyse" wird die Befehlssyntax von `EXAMINE` verwendet. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Anfordern von Ausgaben und Diagrammen für Gesamtsummen neben den Ausgaben und Diagrammen für Gruppen, die durch die Faktorvariablen definiert wurden (mit dem Unterbefehl `TOTAL`).
- Angeben einer gemeinsamen Skala für eine Gruppe von Boxplots (mit dem Unterbefehl `SCALE`).
- Angeben von Interaktionen der Faktorvariablen (mit dem Unterbefehl `VARIABLES`).
- Angeben von anderen Perzentilen als in der Standardeinstellung (mit dem Unterbefehl `PERCENTILES`).
- Berechnen der Perzentile nach fünf Methoden (mit dem Unterbefehl `PERCENTILES`).
- Angeben einer Potenztransformation für Diagramme der Streubreite gegen das mittlere Niveau (mit dem Unterbefehl `PLOT`).
- Angeben der Anzahl von Extremwerten, die angezeigt werden sollen (mit dem Unterbefehl `STATISTICS`).
- Angeben der Parameter für die M-Schätzer, den robusten Schätzern der Lage (mit dem Unterbefehl `MESTIMATORS`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
5

Kreuztabellen

Mit der Prozedur “Kreuztabellen” erzeugen Sie Zweifach- und Mehrfach-Tabellen. Es stehen eine Vielzahl von Tests und Zusammenhangsmaßen für Zweifach-Tabellen zur Verfügung. Welcher Test oder welches Maß verwendet wird, hängt von der Struktur der Tabelle ab und davon, ob die Kategorien geordnet sind.

Statistiken und Zusammenhangsmaße für Kreuztabellen werden nur für Zweifach-Tabellen berechnet. Wenn Sie eine Zeile, eine Spalte und einen Schichtfaktor (Kontroll-Variable) festlegen, wird von der Prozedur “Kreuztabelle” eine separate Ausgabe mit der entsprechenden Statistik sowie den Maßen für jeden Wert des Schichtfaktors (oder eine Kombination der Werte für zwei oder mehrere Kontroll-Variablen) angezeigt. Wenn zum Beispiel *Geschlecht* ein Schichtfaktor für eine Tabelle ist, wobei *verheiratet* (Ja, Nein) gegenüber *Leben* (ist das Leben aufregend, Routine oder langweilig) untersucht wird, werden die Ergebnisse für eine Zweifach-Tabelle für weibliche Personen getrennt von den männlichen berechnet und als aufeinander folgende separate Ausgaben gedruckt.

Beispiel. Wie groß ist die Wahrscheinlichkeit, dass mit den Kunden aus kleineren Unternehmen beim Verkauf von Dienstleistungen (zum Beispiel Weiterbildung und Beratung) ein größerer Gewinn erzielt wird als mit den Kunden aus größeren Unternehmen? Einer Kreuztabelle könnten Sie möglicherweise entnehmen, dass die Mehrheit der kleinen Unternehmen (mit mehr als 500 Angestellten) beim Verkauf von Dienstleistungen einen hohen Gewinn erzielt, während die meisten großen Unternehmen (mit mehr als 2,500 Angestellten) dabei nur niedrige Gewinne erzielen.

Statistiken und Zusammenhangsmaße. Pearson-Chi-Quadrat, Likelihood-Quotienten-Chi-Quadrat, Zusammenhangstest linear-mit-linear, Exakter Test nach Fisher, korrigiertes Chi-Quadrat nach Yates, Pearson-*r*, Spearman-Rho, Kontingenzkoeffizient, Phi, Cramér-*V*, symmetrische und asymmetrische Lambdas, Goodman-und-Kruskal-Tau, Unsicherheitskoeffizient, Gamma, Somer-*d*, Kendall-Tau-*b*, Kendall-Tau-*c*, Eta-Koeffizient, Cohen-Kappa, relativer Risikoschätzer, Quotenverhältnis, McNemar-Test, Cochran- und Mantel-Haenszel-Statistik sowie Spaltenanteilestatistik.

Daten. Um die Kategorien der Tabellenvariablen zu definieren, verwenden Sie Werte einer numerischen Variablen oder einer String-Variablen (maximal 8 Byte). Zum Beispiel können Sie die Daten für *Geschlecht* als 1 und 2 oder als *männlich* und *weiblich* kodieren.

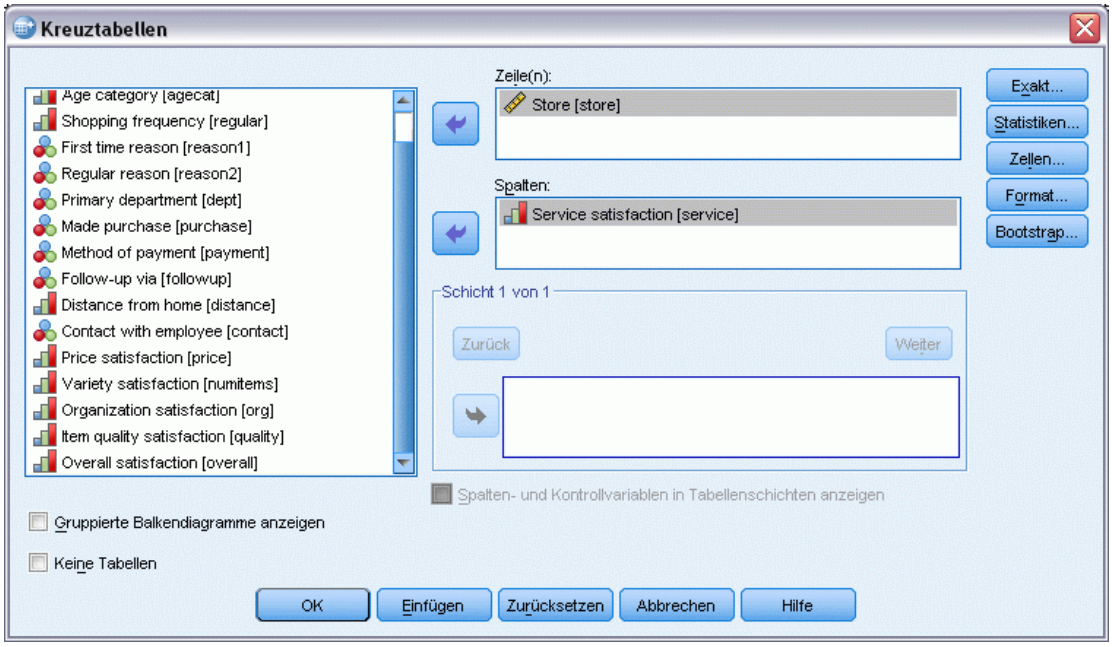
Annahmen. Einige Statistiken und Maße setzen geordnete Kategorien (Ordinal-Daten) oder quantitative Werte (Intervall- oder Verhältnisdaten) voraus, wie bereits im Abschnitt über Statistiken erläutert wurde. Andere sind zulässig, wenn die Tabellenvariablen über ungeordnete Kategorien verfügen (Nominal-Daten). Für Statistiken, die auf Chi-Quadrat basieren (Phi, Cramér-*V*, Kontingenzkoeffizient), sollten die Daten durch eine Zufallsstichprobe aus einer multinomialen Verteilung bezogen werden.

Hinweis: Bei ordinalen Variablen kann es sich um numerische Codes für Kategorien (z. B. 1 = *schwach*, 2 = *mittel*, 3 = *stark*) oder um String-Werte handeln. Die alphabetische Ordnung der String-Werte gibt dabei die Reihenfolge der Kategorien vor. Bei einer String-Variablen mit den Werten *Schwach*, *Mittel* und *Stark* werden die Kategorien beispielsweise in der Reihenfolge *Mittel*, *Schwach*, *Stark* und somit falsch angeordnet. Im allgemeinen ist die Verwendung von numerischem Code für ordinale Daten günstiger.

So lassen Sie Kreuztabellen berechnen:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Deskriptive Statistiken > Kreuztabellen...

Abbildung 5-1
Dialogfeld "Kreuztabellen"



▶ Wählen Sie eine oder mehrere Zeilenvariablen und eine oder mehrere Spaltenvariablen aus.

Die folgenden Optionen sind verfügbar:

- Eine oder mehrere Kontroll-Variablen auswählen.
- Tests und Zusammenhangsmaße der Zweifach-Tabellen oder Untertabellen erhalten Sie, indem Sie auf Statistik klicken.
- Informationen zu beobachteten und erwarteten Werten, Prozentsätzen und Residuen erhalten Sie, indem Sie auf Zellen klicken.
- Durch Klicken auf Format können Sie die Reihenfolge der Kategorien festlegen.

Kreuztabellenschichten

Wenn Sie eine oder mehrere Schichtvariablen auswählen, wird für jede Kategorie jeder Schichtvariablen (Kontroll-Variablen) jeweils eine Kreuztabelle erzeugt. Wenn Sie zum Beispiel über eine Zeilenvariable, eine Spaltenvariable und eine Schichtvariable mit zwei Kategorien verfügen, erhalten Sie eine Zweifach-Tabelle für jede Kategorie der Schichtvariablen. Um eine weitere Schicht von Kontroll-Variablen anzulegen, klicken Sie auf Weiter. Untertabellen werden für jede Kombination von Kategorien für jede Variable der ersten Schicht, jeder Variable der zweiten Schicht und so weiter erzeugt. Wenn Statistiken und Zusammenhangsmaße angefordert werden, treffen diese nur auf Zweifach-Untertabellen zu.

Kreuztabellen: Gruppierte Balkendiagramme

Gruppierte Balkendiagramme anzeigen. Mit einem gruppierten Balkendiagramm können Sie Ihre Daten leichter nach Gruppen von Fällen auswerten. Für jeden Wert der Variablen, der von Ihnen unter Zeilen festgelegt wurde, wird eine Gruppe von Balken erzeugt. Die Balken in jedem Cluster werden durch die unter Spalten angegebene Variable definiert. Für jeden Wert dieser Variablen steht Ihnen ein Set unterschiedlich farbiger oder gemusterter Balken zur Verfügung. Wenn Sie unter Zeilen oder Spalten mehr als eine Variable angeben, wird für jede Kombination von zwei Variablen ein gruppiertes Balkendiagramm erzeugt.

Kreuztabellen: Anzeigen von Schichtvariablen in Tabellenschichten

Anzeigen von Schichtvariablen in Tabellenschichten Sie können festlegen, dass die Schichtvariablen (Kontrollvariablen) als Tabellenschichten in der Kreuztabelle angezeigt werden sollen. Dadurch können Sie Ansichten erstellen, die die Gesamtstatistik für die Zeilen- und Spaltenvariablen anzeigen sowie einen Drilldown für Kategorien der Schichtvariablen gestatten.

Unten finden Sie ein Beispiel, bei dem die Datendatei *demo.sav* () verwendet wird und das wie folgt gewonnen wurde:

- Wählen Sie *Einkommensklassen in Tausend (eink_kl)* als Zeilenvariable aus, *Palm Pilot im Haushalt vorhanden (palm)* als Spaltenvariable und *Schulabschluss (Schulab)* als Schichtvariable.
- Wählen Sie Anzeigen von Schichtvariablen in Tabellenschichten aus.
- Wählen Sie im untergeordneten Dialogfeld “Zellenanzeige” die Option Spalte.
- Führen Sie das Verfahren “Kreuztabellen” aus, doppelklicken Sie auf die Kreuztabelle und wählen Sie in der Dropdown-Liste für das Bildungsniveau die Option Collegeabschluss aus.

Abbildung 5-2
Kreuztabelle mit Schichtvariablen in Tabellenschichten

Einkommensklassen in Tausend * Palm Pilot im Haushalt vorhanden * Schulabschluss Kreuztabelle

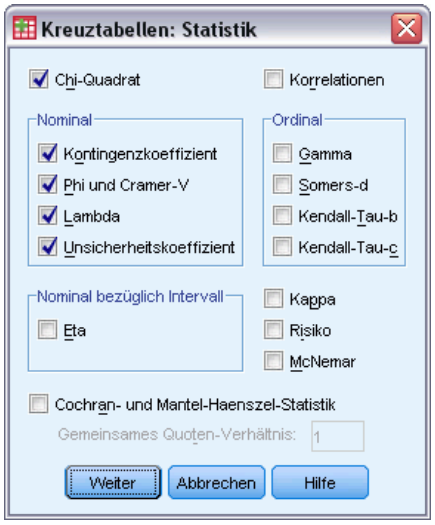
Schulabschluss: Abitur

			Palm Pilot im Haushalt vorhanden		
		Statistik	nein	ja	Gesamt
Einkommensklassen in Tausend	Unter $25	Anzahl	146	50	196
		% innerhalb von Palm Pilot im Haushalt vorhanden	15.8%	11.6%	14.5%
	$25 - $49	Anzahl	335	155	490
		% innerhalb von Palm Pilot im Haushalt vorhanden	36.3%	35.9%	36.2%
	$50 - $74	Anzahl	187	72	259
		% innerhalb von Palm Pilot im Haushalt vorhanden	20.3%	16.7%	19.1%
	$75+	Anzahl	255	155	410
		% innerhalb von Palm Pilot im Haushalt vorhanden	27.6%	35.9%	30.3%
Gesamt		Anzahl	923	432	1355
		% innerhalb von Palm Pilot im Haushalt vorhanden	100.0%	100.0%	100.0%

Die ausgewählte Ansicht der Kreuztabelle zeigt die Statistiken für Befragte mit Collegeabschluss.

Kreuztabellen: Statistik

Abbildung 5-3
Dialogfeld "Kreuztabellen: Statistik"



Chi-Quadrat. Für Tabellen mit zwei Zeilen und zwei Spalten wählen Sie Chi-Quadrat aus, um das Pearson-Chi-Quadrat, das Likelihood-Quotienten-Chi-Quadrat, den exakten Test nach Fisher und das korrigierte Chi-Quadrat nach Yates (Kontinuitätskorrektur) zu berechnen. Für 2 × 2-Tabellen wird der exakte Test nach Fisher berechnet, wenn eine Tabelle, die nicht aus fehlenden Zeilen

oder Spalten einer größeren Tabelle entstanden ist, eine Zelle mit einer erwarteten Häufigkeit von weniger als 5 enthält. Für alle anderen 2 × 2-Tabellen wird das korrigierte Chi-Quadrat nach Yates berechnet. Für Tabellen mit einer beliebigen Anzahl von Zeilen und Spalten wählen Sie Chi-Quadrat aus, um das Pearson-Chi-Quadrat und das Likelihood-Quotienten-Chi-Quadrat zu berechnen. Wenn beide Tabellenvariablen quantitativ sind, ergibt Chi-Quadrat den Zusammenhangstest linear-mit-linear.

Korrelationen. Für Tabellen, in denen sowohl Zeilen als auch Spalten geordnete Werte enthalten, ergeben die Korrelationen den Korrelationskoeffizienten nach Spearman, also Rho (nur numerische Daten). Der Korrelationskoeffizient nach Spearman ist ein Zusammenhangsmaß zwischen den Rangordnungen. Wenn beide Tabellenvariablen (Faktoren) quantitativ sind, ergibt sich unter Korrelationen der Korrelationskoeffizient nach Pearson, *r*, der ein Maß für den linearen Zusammenhang zwischen den Variablen darstellt.

Nominal. Für nominale Daten (ohne implizierte Reihenfolge, wie beispielsweise katholisch, protestantisch, jüdisch) können Sie Kontingenzkoeffizient, Phi (Koeffizient) und Cramér'-V, Lambda (symmetrische und asymmetrische Lambdas sowie Goodman-und-Kruskal-Tau) und Unsicherheitskoeffizient auswählen.

- **Kontingenzkoeffizient.** Ein auf der Chi-Quadrat-Statistik basierendes Zusammenhangsmaß. Dieser Koeffizient liegt immer zwischen 0 und 1, wobei 0 angibt, dass kein Zusammenhang zwischen Zeilen- und Spaltenvariable besteht und Werte nahe 1 auf einen starken Zusammenhang zwischen den Variablen hindeuten. Der maximale Wert hängt von der Anzahl der Zeilen und Spalten in der Tabelle ab.
- **Phi und Cramer-V.** Phi ist ein auf der Chi-Quadrat-Statistik basierendes Zusammenhangsmaß. Es ergibt sich als Wurzel aus dem Quotienten aus Chi-Quadrat und dem Stichprobenumfang. Cramer-V ist ebenfalls ein Zusammenhangsmaß auf der Basis der Chi-Quadrat-Statistik.
- **Lambda.** Ein Zusammenhangsmaß für die proportionale Fehlerreduktion, wenn Werte der unabhängigen Variablen zur Vorhersage von Werten der abhängigen Variablen verwendet werden. Der Wert 1 bedeutet, dass die abhängige Variable durch die unabhängige Variable vollständig vorhergesagt werden kann. Der Wert 0 bedeutet, dass die Vorhersage der abhängigen Variablen durch die unabhängige Variable nicht unterstützt wird.
- **Unsicherheitskoeffizient.** Ein Zusammenhangsmaß, das die proportionale Fehlerreduktion angibt, wenn Werte einer Variablen zur Vorhersage von Werten der anderen Variablen verwendet werden. Ein Wert von 0,83 gibt z. B. an, dass die Kenntnis einer Variablen den Fehler bei der Vorhersage der Werte der anderen Variablen um 83 % reduziert. Das Programm berechnet beide Versionen des Unsicherheitskoeffizienten, die symmetrische und die asymmetrische.

Ordinal.Für Tabellen, in welchen die Zeilen und Spalten geordnete Werte enthalten, wählen Sie Gamma (nullte Ordnung für Zweifach-Tabellen und bedingt für Dreifach- bis Zehnfach-Tabellen), Kendall-Tau-b und Kendall-Tau-c aus. Zur Vorhersage von Spaltenkategorien auf der Grundlage von Zeilenkategorien wählen Sie Somers-d aus.

- **Gamma.** Ein symmetrisches Zusammenhangsmaß für zwei ordinalskalierte Variablen, dessen Wertebereich zwischen -1 und +1 liegt. Werte nahe bei -1 oder +1 weisen auf einen starken Zusammenhang zwischen den Variablen hin. Werte nahe 0 stehen für einen schwachen oder fehlenden Zusammenhang. Zeigt Gamma-Werte nullter Ordnung für Tabellen mit 2 Variablen an. Für Tabellen mit drei oder mehr Variablen werden bedingte Gamma-Werte angezeigt.

- **Somers-d.** Ein Zusammenhangsmaß für zwei ordinale Variablen, dessen Wertebereich zwischen -1 und +1 liegt. Werte, die betragsmäßig nahe bei 1 liegen, geben einen starken Zusammenhang zwischen den beiden Variablen an, Werte nahe 0 einen schwachen oder fehlenden Zusammenhang. Somers-d ist eine asymmetrische Erweiterung von Gamma. Der Unterschied liegt in der Einbeziehung der Anzahl von Paaren, die keine Bindungen in der unabhängigen Variablen aufweisen. Eine symmetrische Version dieser Statistik wird ebenfalls berechnet.
- **Kendall-Tau-b.** Ein nichtparametrisches Korrelationsmaß für ordinale Variablen oder Ränge, das Bindungen berücksichtigt. Das Vorzeichen des Koeffizienten gibt die Richtung des Zusammenhangs an und sein Betrag die Stärke; dabei entsprechen betragsmäßig größere Werte einem stärkeren Zusammenhang. Die möglichen Werte liegen im Bereich von -1 und 1, ein Wert von -1 oder +1 ergibt sich jedoch nur aus quadratischen Tabellen.
- **Kendall-Tau-c.** Ein nichtparametrisches Zusammenhangsmaß für ordinale Variablen, das Bindungen ignoriert. Das Vorzeichen des Koeffizienten gibt die Richtung des Zusammenhangs an und sein Betrag die Stärke; dabei entsprechen betragsmäßig größere Werte einem stärkeren Zusammenhang. Die möglichen Werte liegen im Bereich von -1 und 1, ein Wert von -1 oder +1 ergibt sich jedoch nur aus quadratischen Tabellen.

Nominal bezüglich Intervall. Wenn eine Variable kategorial und eine andere quantitativ ist, wählen Sie Eta aus. Die kategoriale Variable muss numerisch kodiert sein.

- **Eta.** Ein Zusammenhangsmaß, das zwischen 0 und 1 liegt; dabei steht 0 für fehlenden Zusammenhang zwischen den Zeilen- und Spaltenvariablen und Werte nahe bei 1 geben einen starken Zusammenhang an. Eta ist geeignet für eine intervallskalierte abhängige Variable (z. B. Einkommen) und eine unabhängige Variable mit einer begrenzten Anzahl von Kategorien (z. B. Geschlecht). Es werden zwei Eta-Werte berechnet: der eine behandelt die Zeilenvariablen und der andere die Spaltenvariable als intervallskalierte Variable.

Kappa. Der Cohen-Kappa-Koeffizient misst die Übereinstimmung zwischen den Beurteilungen zweier Prüfer, wenn beide dasselbe Objekt bewerten. Der Wert 1 bedeutet perfekte Übereinstimmung. Der Wert 0 bedeutet, dass die Übereinstimmung nicht über das zufallsbedingte Maß hinausgeht. Kappa ist nur für Tabellen verfügbar, in denen beide Variablen die gleiche Anzahl von Kategorien und gleiche Kategorienwerte (Ausprägungen) aufweisen.

Risiko. Ein Maß, das bei 2 x 2-Tabellen die Stärke des Zusammenhangs zwischen dem Vorhandensein eines Faktors und dem Auftreten eines Ereignisses misst. Wenn das Konfidenzintervall für die Statistik den Wert 1 enthält, ist nicht anzunehmen, dass zwischen Faktor und Ereignis ein Zusammenhang besteht. Das Quotenverhältnis (Odds Ratio) kann als Schätzer für das relative Risiko verwendet werden, wenn der Faktor selten auftritt.

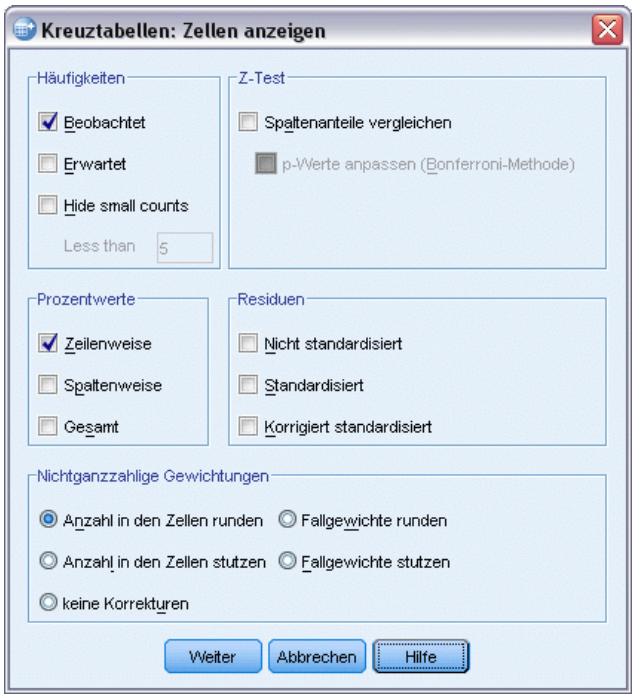
McNemar. Ein nichtparametrischer Test für zwei verbundene dichotome Variablen. Prüft unter Verwendung der Chi-Quadrat-Verteilung, ob Änderungen bei den Antworten vorliegen. Dieser Test ist für das Erkennen von Änderungen bei Antworten nützlich, die durch experimentelle Einflussnahme in so genannten "Vorher-und-nachher-Designs" entstanden sind. Bei größeren quadratischen Tabellen wird der McNemar-Bowker-Test auf Symmetrie ausgegeben.

Cochran- und Mantel-Haenszel-Statistik. Die Cochran- und die Mantel-Haenszel-Statistik können verwendet werden, um auf Unabhängigkeit zwischen einer dichotomen Faktorvariablen und einer dichotomen Response-Variablen zu testen, und zwar in Abhängigkeit von einem Kovariatenmuster, das durch eine oder mehrere Schichtvariablen (Kontrollvariablen) definiert

wird. Beachten Sie, dass andere Statistiken schichtenweise berechnet werden, die Cochran- und die Mantel-Haenszel-Statistik dagegen einmal für alle Schichten berechnet werden.

Kreuztabellen: Zellenanzeige

Abbildung 5-4
Dialogfeld "Kreuztabellen: Zellenanzeige"



Um Sie beim Erkennen von Mustern in den Daten zu unterstützen, die zu einem signifikanten Chi-Quadrat-Test beitragen, zeigt die Prozedur "Kreuztabellen" die erwarteten Häufigkeiten und drei Typen von Residuen (Abweichungen) an, welche die Differenz zwischen beobachteten und erwarteten Häufigkeiten messen. Jede Zelle der Tabelle kann jede Kombination von ausgewählten Häufigkeiten, Prozentzahlen und Residuen enthalten.

Häufigkeiten. Die Anzahl der Fälle, die tatsächlich beobachtet, und die Anzahl der Fälle, die erwartet werden, wenn die Zeilen- und Spaltenvariablen voneinander unabhängig sind.

Spaltenanteile vergleichen. Mit dieser Option werden paarweise Vergleiche von Spaltenanteilen berechnet und es wird angezeigt, welche Spaltenpaare (für eine bestimmte Zeile) sich signifikant unterscheiden. Signifikante Unterschiede werden in der Kreuztabelle mit Formatierung im APA-Stil mit tiefgestellten Buchstaben gekennzeichnet und auf dem 0,05-Signifikanzniveau berechnet.

- **p-Werte anpassen (Bonferroni-Methode).** Bei paarweisen Vergleichen von Spaltenanteilen wird die Bonferroni-Korrektur genutzt, die das beobachtete Signifikanzniveau für Mehrfachvergleiche anpasst.

Prozentwerte. Die Prozentwerte können horizontal in den Zeilen oder vertikal in den Spalten addiert werden. Der prozentuale Anteil der Gesamtanzahl der Fälle, die in einer Tabelle dargestellt werden (eine Schicht), ist ebenfalls verfügbar.

Residuen. Einfache nicht standardisierte Residuen geben die Differenz zwischen den beobachteten und erwarteten Werten wieder. Standardisierte und korrigierte standardisierte Residuen sind ebenfalls verfügbar.

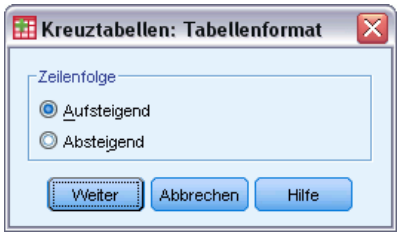
- **Nicht standardisiert.** Die Differenz zwischen einem beobachteten Wert und dem erwarteten Wert. Der erwartete Wert ist die Anzahl von Fällen, die man in einer Zelle erwarten würde, wenn kein Zusammenhang zwischen den beiden Variablen bestünde. Ein positives Residuum zeigt an, dass in der Zelle mehr Fälle vorliegen, als dies der Fall wäre, wenn die Zeilen- und Spaltenvariable unabhängig wären.
- **Standardisiert.** Der Quotient aus dem Residuum und einem Schätzer seiner Standardabweichung. Standardisierte Residuen, auch bekannt als Pearson-Residuen, haben einen Mittelwert von 0 und eine Standardabweichung von 1.
- **Korrigiert standardisiert.** Der Quotient aus dem Residuum einer Zelle (beobachteter Wert minus erwarteter Wert) und dessen geschätztem Standardfehler. Das resultierende standardisierte Residuum wird in Einheiten der Standardabweichung über oder unter dem Mittelwert angegeben.

Nichtganzzahlige Gewichtungen. Bei den Zellhäufigkeiten handelt es sich normalerweise um ganzzahlige Werte, da sie für die Anzahl der Fälle in den einzelnen Zellen stehen. Wenn jedoch die Datendatei derzeit mit einer Gewichtungsvariablen mit Bruchzahlenwerten (z. B. 1,25) gewichtet ist, können die Zellhäufigkeiten ebenfalls Bruchwerte sein. Sie können die Werte vor oder nach der Berechnung der Zellhäufigkeiten abschneiden oder runden oder sowohl für die Tabellenanzeige als auch für statistische Berechnungen gebrochene Zellhäufigkeiten verwenden.

- **Anzahl in den Zellen runden.** Fallgewichte werden verwendet, wie gegeben, aber die akkumulierten Gewichte für die Zellen werden gerundet, bevor Statistiken berechnet werden.
- **Anzahl in den Zellen stutzen.** Fallgewichte werden verwendet, wie gegeben, aber die addierten Gewichte für die Zellen werden auf den ganzzahligen Anteil gestutzt, bevor Statistiken berechnet werden.
- **Fallgewichte runden.** Fallgewichte werden gerundet, bevor sie verwendet werden.
- **Fallgewichte stutzen.** Fallgewichte werden auf den ganzzahligen Anteil gestutzt, bevor sie verwendet werden.
- **keine Korrekturen.** Fallgewichte werden verwendet wie gegeben und auch nicht ganzzahlige Zellhäufigkeiten werden verwendet. Wenn jedoch exakte Statistiken (verfügbar mit dem Modul "Exakte Tests") angefordert werden, dann werden die akkumulierten Gewichte in den Zellen entweder auf den ganzzahligen Anteil gestutzt oder gerundet, bevor die Statistiken für exakte Tests berechnet werden.

Kreuztabellen: Tabellenformat

Abbildung 5-5
Dialogfeld "Kreuztabellen: Tabellenformat"



Sie können Zeilen in aufsteigender oder absteigender Reihenfolge der Werte der Zeilenvariablen anordnen.

Kapitel
6

Zusammenfassen

Mit der Prozedur “Zusammenfassen” werden Untergruppenstatistiken für Variablen innerhalb der Kategorien einer oder mehrerer Gruppenvariablen berechnet. Alle Ebenen der Gruppenvariablen werden in die Kreuztabelle aufgenommen. Sie können wählen, in welcher Reihenfolge die Statistiken angezeigt werden. Außerdem werden Auswertungsstatistiken für jede Variable über alle Kategorien angezeigt. Die Datenwerte jeder Kategorie können aufgelistet oder unterdrückt werden. Bei umfangreichen Daten-Sets haben Sie die Möglichkeit, nur die ersten *n* Fälle aufzulisten.

Beispiel. Wie hoch liegen die durchschnittlichen Verkaufszahlen eines Produkts, gegliedert nach Region und Abnehmer? Möglicherweise stellen Sie fest, dass im Westen im Durchschnitt geringfügig mehr verkauft wird als in anderen Regionen, wobei gewerbliche Kunden in der westlichen Region die wichtigsten Abnehmer sind.

Statistiken. Summe, Anzahl der Fälle, Mittelwert, Median, gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppenvariablen, Variablenwert der letzten Kategorie der Gruppenvariablen, Standardabweichung, Varianz, Kurtosis, Standardfehler der Kurtosis, Schiefe, Standardfehler der Schiefe, Prozent der Gesamtsumme, Prozent der Gesamtanzahl (*N*), Prozent der Summe in, Prozent der Anzahl (*N*) in, geometrisches Mittel und harmonisches Mittel.

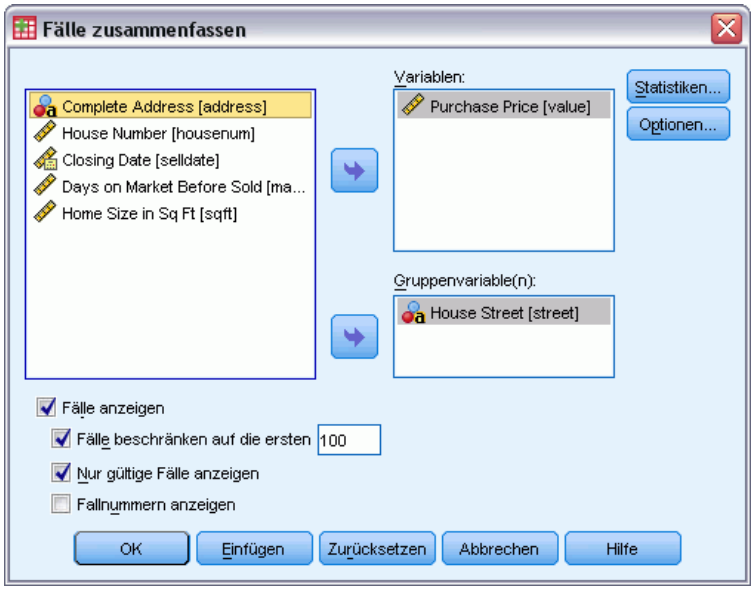
Daten. Die Gruppenvariablen stellen kategoriale Variablen dar, deren Werte numerisch oder Strings sein können. Die Anzahl der Kategorien sollte angemessen klein gehalten werden. Den anderen Variablen müssen Ränge zugeordnet werden können.

Annahmen. Einige der möglichen Untergruppenstatistiken, wie beispielsweise Mittelwert und Standardabweichung, basieren auf der Annahme, dass eine Normalverteilung vorliegt, und sind für Variablen mit symmetrischen Verteilungen geeignet. Robuste Statistiken, wie beispielsweise Median und Spannweite, sind für quantitative Variablen geeignet, die möglicherweise die Annahme einer Normalverteilung erfüllen.

So erstellen Sie Zusammenfassungen von Fällen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Berichte > Fälle zusammenfassen...

Abbildung 6-1
Dialogfeld "Fälle zusammenfassen"



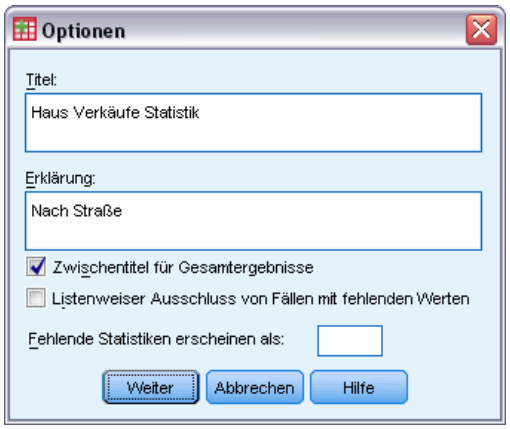
▶ Wählen Sie mindestens eine Variable aus.

Die folgenden Optionen sind verfügbar:

- Sie können eine oder mehrere Gruppenvariablen auswählen, um die Daten in Untergruppen aufzuteilen.
- Klicken Sie auf Optionen, wenn Sie den Ausgabetitel ändern, eine Erklärung unter der Ausgabe hinzufügen oder Fälle mit fehlenden Werten ausschließen möchten.
- Sie können optionale Statistiken anzeigen lassen, indem Sie auf Statistik klicken.
- Wählen Sie Fälle anzeigen, um die Fälle in jeder Untergruppe auflisten zu lassen. In der Standardeinstellung werden nur die ersten 100 Fälle in der Datei aufgelistet. Sie können den Wert für Fälle beschränken auf die ersten*n* erhöhen oder vermindern bzw. diese Option deaktivieren, um alle Fälle auflisten zu lassen.

Zusammenfassen: Optionen

Abbildung 6-2
Dialogfeld "Optionen"



Sie können den Titel der Ausgabe ändern oder eine Erklärung hinzufügen, die unter der Ausgabetabelle angezeigt wird. Sie können den Zeilenumbruch in Titeln und Erklärungen steuern, indem Sie an die Stellen, an denen ein Zeilenumbruch durchgeführt werden soll, die Zeichen \n eingeben.

Außerdem können Sie Untertitel für Gesamtergebnisse ein- oder ausblenden sowie Fälle mit fehlenden Werten für beliebige, in der Analyse verwendete Variablen ein- oder ausschließen. Oft ist es angebracht, fehlende Fälle in der Ausgabe mit einem Punkt oder einem Sternchen zu kennzeichnen. Geben Sie ein Zeichen, eine Wortgruppe oder einen Code ein, der bei einem fehlenden Wert angezeigt werden soll, andernfalls werden fehlende Werte in der Ausgabe nicht besonders verarbeitet.

Zusammenfassung: Statistik

Abbildung 6-3
Dialogfeld "Auswertungsstatistik"

Sie können mindestens eine der folgenden Untergruppen-Statistiken für die Variablen innerhalb jeder Kategorie jeder Gruppenvariablen auswählen: Summe, Anzahl der Fälle, Mittelwert, Median, gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppenvariablen, Variablenwert der letzten Kategorie der Gruppenvariablen, Standardabweichung, Varianz, Kurtosis, Standardfehler der Kurtosis, Schiefe, Standardfehler der Schiefe, Prozent der Gesamtsumme, Prozent der Gesamtanzahl, Prozent der Summe in, Prozent der Anzahl in, geometrisches Mittel und harmonisches Mittel. Die Statistiken werden in der Liste "Zellenstatistik" in derselben Reihenfolge angezeigt, in welcher sie in der Ausgabe angezeigt werden. Außerdem werden die Auswertungsstatistiken für jede Variable über alle Kategorien angezeigt.

Erster. Zeigt den ersten Datenwert in der Datendatei an.

Geometrisches Mittel. Die n-te Wurzel aus dem Produkt der Datenwerte, wobei n der Anzahl der Fälle entspricht.

Gruppierter Median. Der Median für Daten, die in Gruppen kodiert wurden (bei denen also ein Wert für ein ganzes Intervall steht). Wenn z. B. für das Alter jeder Wert in den Dreißigern als 35 kodiert ist, jeder Wert in den Vierzigern als 45 usw., dann wird der gruppierte Median aus den kodierten Daten berechnet.

Harmonisches Mittel. Wird verwendet, um die "mittlere" Gruppengröße zu bestimmen, wenn der Stichprobenumfang in den einzelnen Gruppen unterschiedlich ist. Das harmonische Mittel ist gleich der Gesamtzahl der Stichproben geteilt durch die Summe der Kehrwerte der Stichprobengrößen.

Kurtosis. Ein Maß dafür, wie sich die Beobachtungen um einen zentralen Punkt gruppieren. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremverten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.

Letzter. Hiermit wird der letzte Datenwert in der Datendatei angezeigt.

Maximum. Der größte Wert einer numerischen Variablen.

Mittelwert. Ein Lagemaß. Die Summe der Ränge, geteilt durch die Zahl der Fälle.

Median. Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).

Minimum. Der kleinste Wert einer numerischen Variablen.

N. Die Anzahl der Fälle (Beobachtungen oder Datensätze).

Prozent der Gesamtanzahl. Prozentsatz der Gesamtanzahl von Fällen in jeder Kategorie.

Prozent der Gesamtsumme. Prozentsatz der Gesamtsumme in jeder Kategorie.

Spannweite. Die Differenz zwischen den größten und kleinsten Werten einer numerischen Variablen; Maximalwert minus Minimalwert.

Schiefe. Ein Maß für die Asymmetrie einer Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefe-Wert, der mehr als doppelt so groß ist wie sein Standardfehler, für eine Abweichung von der Symmetrie spricht.

Standardfehler der Kurtosis. Das Verhältnis der Kurtosis zu ihrem Standardfehler kann für einen Test auf Normalverteilung verwendet werden (d. h. die Annahme, dass Normalverteilung vorliegt, kann abgelehnt werden, wenn das Verhältnis kleiner als -2 oder größer als +2 ist). Ein großer positiver Wert für die Kurtosis deutet darauf hin, dass die Flanken der Verteilung länger sind als bei einer Normalverteilung; ein negativer Wert bedeutet, dass sie kürzer sind (etwa wie bei einer kastenförmigen, gleichförmigen Verteilung).

Standardfehler der Schiefe. Das Verhältnis der Schiefe zu ihrem Standardfehler kann für einen Test auf Normalverteilung verwendet werden (d. h. die Annahme, dass Normalverteilung vorliegt, kann abgelehnt werden, wenn das Verhältnis kleiner als -2 oder größer als +2 ist). Ein großer positiver Wert für die Schiefe bedeutet, dass die Verteilung eine lange rechte Flanke hat; ein extremer negativer Wert bedeutet, dass sie eine lange linke Flanke hat.

Summe. Die Summe der Werte über alle Fälle mit nichtfehlenden Werten.

Varianz. Ein Maß der Streuung um den Mittelwert. Es ist gleich dem Quotienten aus der Summe der quadrierten Abweichung vom Mittelwert und der um 1 verringerten Fallanzahl. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.

Kapitel
7

Mittelwerte

Mit der Prozedur “Mittelwerte” werden die Mittelwerte von Untergruppen und verwandte univariate Statistiken für abhängige Variablen innerhalb von Kategorien von mindestens einer unabhängigen Variablen berechnet. Wahlweise können Sie eine einfaktorielle Varianzanalyse, Eta und einen Test auf Linearität berechnen lassen.

Beispiel. Sie messen die mittlere Menge von Fett, die von drei verschiedenen Sorten Speiseöl absorbiert wird. Anschließend führen Sie eine einfaktorielle Varianzanalyse aus, um festzustellen, ob sich die Mittelwerte unterscheiden.

Statistiken. Summe, Anzahl der Fälle, Mittelwert, Median, gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppenvariablen, Variablenwert der letzten Kategorie der Gruppenvariablen, Standardabweichung, Varianz, Kurtosis, Standardfehler der Kurtosis, Schiefe, Standardfehler der Schiefe, Prozent der Gesamtsumme, Prozent der Gesamtanzahl (*N*), Prozent der Summe in, Prozent der Anzahl (*N*) in, geometrisches Mittel und harmonisches Mittel. Unter Optionen stehen außerdem Varianzanalyse, Eta, Eta-Quadrat, *R* und R^2 zur Verfügung.

Daten. Die abhängigen Variablen sind quantitativ, die unabhängigen Variablen kategorial. Die Werte der kategorialen Variablen können numerische Variablen oder String-Variablen sein.

Annahmen. Einige der möglichen Untergruppenstatistiken, wie beispielsweise Mittelwert und Standardabweichung, basieren auf der Annahme, dass eine Normalverteilung vorliegt, und sind für Variablen mit symmetrischen Verteilungen geeignet. Robuste Statistiken, z. B. Median, sind für quantitative Variablen geeignet, die möglicherweise die Annahme einer Normalverteilung erfüllen. Die Varianzanalyse ist gegenüber Abweichungen von der Normalverteilung robust. Allerdings sollten die Daten in jeder Zelle symmetrisch sein. Bei der Varianzanalyse wird außerdem angenommen, dass die Gruppen aus Grundgesamtheiten mit gleichen Varianzen stammen. Zum Testen dieser Annahme können Sie den Levene-Test auf Homogenität der Varianzen verwenden. Dieser Test ist in der Prozedur “Einfaktorielle ANOVA” verfügbar.

So berechnen Sie die Mittelwerte der Untergruppen:

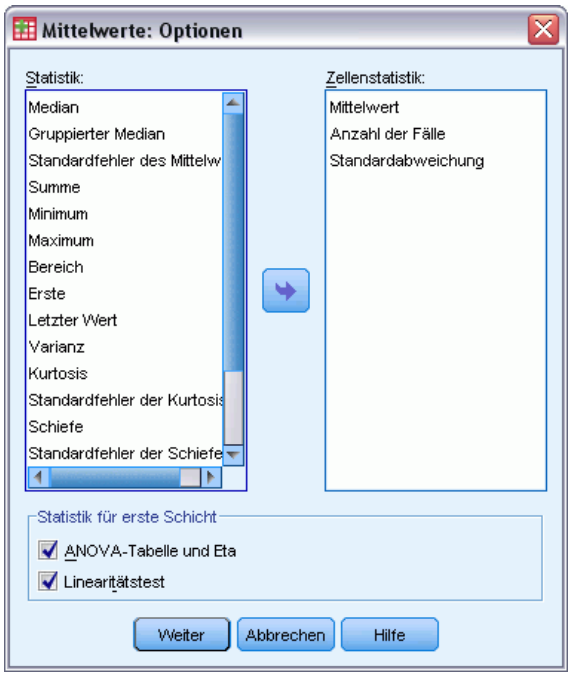
▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Mittelwerte vergleichen > Mittelwerte...

Abbildung 7-1
Dialogfeld "Mittelwerte"

- Wählen Sie eine oder mehrere abhängige Variablen aus.
- Verwenden Sie eine der folgenden Methoden, um die kategorialen unabhängigen Variablen auszuwählen:
  - Wählen Sie mindestens eine unabhängige Variable aus. Für jede unabhängige Variable werden getrennte Ergebnisse angezeigt.
  - Wählen Sie mindestens eine Schicht von unabhängigen Variablen aus. Die Stichprobe wird durch jede Schicht weiter unterteilt. Wenn es eine unabhängige Variable in Schicht 1 und eine unabhängige Variable in Schicht 2 gibt, werden die Ergebnisse nicht in einzelnen Tabellen für die unabhängigen Variablen, sondern in einer Kreuztabelle angezeigt.
- Sie können optionale Statistiken, eine Tabelle für die Varianzanalyse, Eta, Eta-Quadrat, R und R^2 berechnen lassen, indem Sie auf Optionen klicken.

Mittelwerte: Optionen

Abbildung 7-2
Dialogfeld "Mittelwerte: Optionen"



Sie können mindestens eine der folgenden Untergruppen-Statistiken für die Variablen innerhalb jeder Kategorie jeder Gruppenvariablen auswählen: Summe, Anzahl der Fälle, Mittelwert, Median, gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppenvariablen, Variablenwert der letzten Kategorie der Gruppenvariablen, Standardabweichung, Varianz, Kurtosis, Standardfehler der Kurtosis, Schiefe, Standardfehler der Schiefe, Prozent der Gesamtsumme, Prozent der Gesamtanzahl, Prozent der Summe in, Prozent der Anzahl in, geometrisches Mittel und harmonisches Mittel. Sie können die Reihenfolge ändern, in der die Statistiken für die Untergruppen berechnet werden. Die Statistiken werden in der Liste "Zellenstatistik" in derselben Reihenfolge angezeigt, in der sie in der Ausgabe angezeigt werden. Außerdem werden die Auswertungsstatistiken für jede Variable über alle Kategorien angezeigt.

Erster. Zeigt den ersten Datenwert in der Datendatei an.

Geometrisches Mittel. Die n-te Wurzel aus dem Produkt der Datenwerte, wobei n der Anzahl der Fälle entspricht.

Gruppierter Median. Der Median für Daten, die in Gruppen kodiert wurden (bei denen also ein Wert für ein ganzes Intervall steht). Wenn z. B. für das Alter jeder Wert in den Dreißigern als 35 kodiert ist, jeder Wert in den Vierzigern als 45 usw., dann wird der gruppierte Median aus den kodierten Daten berechnet.

Harmonisches Mittel. Wird verwendet, um die "mittlere" Gruppengröße zu bestimmen, wenn der Stichprobenumfang in den einzelnen Gruppen unterschiedlich ist. Das harmonische Mittel ist gleich der Gesamtzahl der Stichproben geteilt durch die Summe der Kehrwerte der Stichprobengrößen.

Kurtosis. Ein Maß dafür, wie sich die Beobachtungen um einen zentralen Punkt gruppieren. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremverten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.

Letzter. Hiermit wird der letzte Datenwert in der Datendatei angezeigt.

Maximum. Der größte Wert einer numerischen Variablen.

Mittelwert. Ein Lagemaß. Die Summe der Ränge, geteilt durch die Zahl der Fälle.

Median. Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).

Minimum. Der kleinste Wert einer numerischen Variablen.

N. Die Anzahl der Fälle (Beobachtungen oder Datensätze).

Prozent der Gesamtanzahl. Prozentsatz der Gesamtanzahl von Fällen in jeder Kategorie.

Prozent der Gesamtsumme. Prozentsatz der Gesamtsumme in jeder Kategorie.

Spannweite. Die Differenz zwischen den größten und kleinsten Werten einer numerischen Variablen; Maximalwert minus Minimalwert.

Schiefe. Ein Maß für die Asymmetrie einer Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefe-Wert, der mehr als doppelt so groß ist wie sein Standardfehler, für eine Abweichung von der Symmetrie spricht.

Standardfehler der Kurtosis. Das Verhältnis der Kurtosis zu ihrem Standardfehler kann für einen Test auf Normalverteilung verwendet werden (d. h. die Annahme, dass Normalverteilung vorliegt, kann abgelehnt werden, wenn das Verhältnis kleiner als -2 oder größer als +2 ist). Ein großer positiver Wert für die Kurtosis deutet darauf hin, dass die Flanken der Verteilung länger sind als bei einer Normalverteilung; ein negativer Wert bedeutet, dass sie kürzer sind (etwa wie bei einer kastenförmigen, gleichförmigen Verteilung).

Standardfehler der Schiefe. Das Verhältnis der Schiefe zu ihrem Standardfehler kann für einen Test auf Normalverteilung verwendet werden (d. h. die Annahme, dass Normalverteilung vorliegt, kann abgelehnt werden, wenn das Verhältnis kleiner als -2 oder größer als +2 ist). Ein großer

positiver Wert für die Schiefe bedeutet, dass die Verteilung eine lange rechte Flanke hat; ein extremer negativer Wert bedeutet, dass sie eine lange linke Flanke hat.

Summe. Die Summe der Werte über alle Fälle mit nichtfehlenden Werten.

Varianz. Ein Maß der Streuung um den Mittelwert. Es ist gleich dem Quotienten aus der Summe der quadrierten Abweichung vom Mittelwert und der um 1 verringerten Fallanzahl. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.

Statistik für erste Schicht

ANOVA-Tabelle und Eta. Zeigt eine Tabelle für eine einfaktorielle Varianzanalyse an und berechnet Eta und Eta-Quadrat (Zusammenhangsmaße) für jede unabhängige Variable in der ersten Schicht.

Linearitätstest. Berechnet für lineare und nichtlineare Komponenten die Quadratsummen, die Freiheitsgrade und das Mittel der Quadrate sowie den F-Wert, R und R-Quadrat. Die Berechnungen für Linearität werden nicht durchgeführt, wenn die unabhängige Variable eine kurze String-Variable ist.

Kapitel
8

OLAP-Würfel

Mit der Prozedur “OLAP-Würfel” (Online Analytical Processing) werden Gesamtwerte, Mittelwerte und andere univariate Statistiken für stetige Auswertungsvariablen innerhalb der Kategorien von mindestens einer kategorialen Gruppenvariablen berechnet. Für jede Kategorie der Gruppenvariablen wird eine separate Schicht erstellt.

Beispiel. Durchschnittlicher und gesamter Umsatz für verschiedene Regionen und Produktlinien innerhalb einer Region.

Statistiken. Summe, Anzahl der Fälle, Mittelwert, Median, Gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppenvariablen, Variablenwert der letzten Kategorie der Gruppenvariablen, Standardabweichung, Varianz, Kurtosis, Standardfehler der Kurtosis, Schiefe, Standardfehler der Schiefe, Prozentsatz der gesamten Fälle, Prozentsatz der Gesamtsumme, Prozentsatz der gesamten Fälle innerhalb der Gruppenvariablen, Prozentsatz der Gesamtsumme innerhalb der Gruppenvariablen, geometrisches Mittel und harmonisches Mittel.

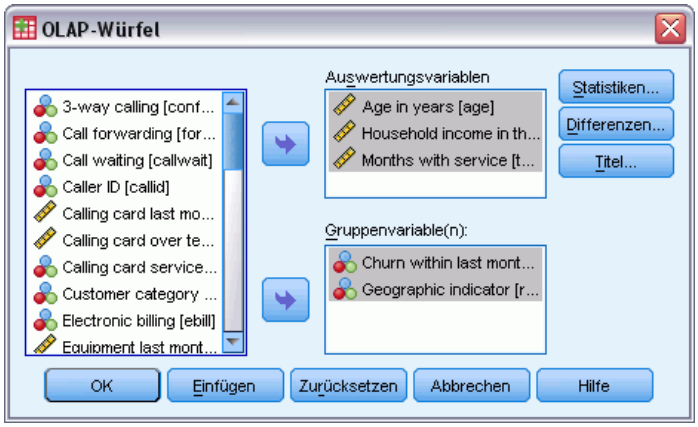
Daten. Die Auswertungsvariablen sind quantitativ (stetige Variablen, die auf einer Intervall- oder Verhältnisskala gemessen werden) und die Gruppenvariablen kategorial. Die Werte der kategorialen Variablen können numerische Variablen oder String-Variablen sein.

Annahmen. Einige der möglichen Untergruppenstatistiken, wie beispielsweise Mittelwert und Standardabweichung, basieren auf der Annahme, dass eine Normalverteilung vorliegt, und sind für Variablen mit symmetrischen Verteilungen geeignet. Robuste Statistiken, wie z. B. Median und Spannweite, sind für quantitative Variablen geeignet, die möglicherweise die Annahme einer Normalverteilung erfüllen.

So erstellen Sie OLAP-Würfel:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Berichte > OLAP-Würfel...

Abbildung 8-1
Dialogfeld "OLAP-Würfel"



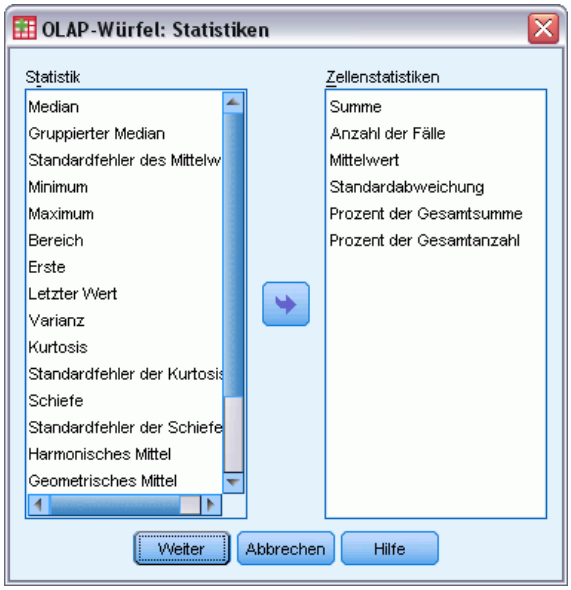
- ▶ Wählen Sie mindestens eine stetige Auswertungsvariable aus.
- ▶ Wählen Sie mindestens eine kategoriale Gruppenvariable aus.

Die folgenden Optionen sind verfügbar:

- Sie können verschiedene Auswertungsstatistiken auswählen, indem Sie auf Statistiken klicken. Sie müssen mindestens eine Gruppenvariable auswählen, bevor Sie die Auswertungsstatistiken auswählen können.
- Sie können die Differenzen zwischen Variablenpaaren und Gruppenpaaren berechnen lassen, die durch eine Gruppenvariable definiert sind, indem Sie auf Differenzen klicken.
- Sie können Titel für benutzerdefinierte Tabellen erstellen, indem Sie auf Titel klicken.

OLAP-Würfel: Statistiken

Abbildung 8-2
Dialogfeld "OLAP-Würfel: Statistiken"



Sie können eine oder mehrere der folgenden Untergruppen-Statistiken für die Auswertungsvariablen in jeder Kategorie aller Gruppenvariablen auswählen: Summe, Anzahl der Fälle, Mittelwert, Median, Gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppenvariablen, Variablenwert der letzten Kategorie der Gruppenvariablen, Standardabweichung, Varianz, Kurtosis, Standardfehler der Kurtosis, Schiefe, Standardfehler der Schiefe, Prozentsatz der gesamten Fälle, Prozentsatz der Gesamtsumme, Prozentsatz der gesamten Fälle innerhalb der Gruppenvariablen, Prozentsatz der Gesamtsumme innerhalb der Gruppenvariablen, geometrisches Mittel und harmonisches Mittel.

Sie können die Reihenfolge ändern, in der die Statistiken für die Untergruppen berechnet werden. Die Statistiken werden in der Liste "Zellenstatistik" in derselben Reihenfolge angezeigt, in der sie in der Ausgabe angezeigt werden. Außerdem werden die Auswertungsstatistiken für jede Variable über alle Kategorien angezeigt.

Erster. Zeigt den ersten Datenwert in der Datendatei an.

Geometrisches Mittel. Die n-te Wurzel aus dem Produkt der Datenwerte, wobei n der Anzahl der Fälle entspricht.

Gruppierter Median. Der Median für Daten, die in Gruppen kodiert wurden (bei denen also ein Wert für ein ganzes Intervall steht). Wenn z. B. für das Alter jeder Wert in den Dreißigern als 35 kodiert ist, jeder Wert in den Vierzigern als 45 usw., dann wird der gruppierte Median aus den kodierten Daten berechnet.

Harmonisches Mittel. Wird verwendet, um die "mittlere" Gruppengröße zu bestimmen, wenn der Stichprobenumfang in den einzelnen Gruppen unterschiedlich ist. Das harmonische Mittel ist gleich der Gesamtzahl der Stichproben geteilt durch die Summe der Kehrwerte der Stichprobengrößen.

Kurtosis. Ein Maß dafür, wie sich die Beobachtungen um einen zentralen Punkt gruppieren. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremverten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.

Letzter. Hiermit wird der letzte Datenwert in der Datendatei angezeigt.

Maximum. Der größte Wert einer numerischen Variablen.

Mittelwert. Ein Lagemaß. Die Summe der Ränge, geteilt durch die Zahl der Fälle.

Median. Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).

Minimum. Der kleinste Wert einer numerischen Variablen.

N. Die Anzahl der Fälle (Beobachtungen oder Datensätze).

Prozent der Anzahl in. Prozentsatz der Gesamtanzahl von Fällen für die angegebene Gruppenvariable in den Kategorien der anderen Gruppenvariablen. Wenn nur eine Gruppenvariable vorhanden ist, ist dieser Wert gleich dem Prozentsatz der Gesamtanzahl von Fällen.

Prozent der Summe in. Prozentsatz der Summe für die angegebene Gruppenvariable in den Kategorien der anderen Gruppenvariablen. Wenn nur eine Gruppenvariable vorhanden ist, ist dieser Wert gleich dem Prozentsatz der Gesamtsumme.

Prozent der Gesamtanzahl. Prozentsatz der Gesamtanzahl von Fällen in jeder Kategorie.

Prozent der Gesamtsumme. Prozentsatz der Gesamtsumme in jeder Kategorie.

Spannweite. Die Differenz zwischen den größten und kleinsten Werten einer numerischen Variablen; Maximalwert minus Minimalwert.

Schiefe. Ein Maß für die Asymmetrie einer Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefe-Wert, der mehr als doppelt so groß ist wie sein Standardfehler, für eine Abweichung von der Symmetrie spricht.

Standardfehler der Kurtosis. Das Verhältnis der Kurtosis zu ihrem Standardfehler kann für einen Test auf Normalverteilung verwendet werden (d. h. die Annahme, dass Normalverteilung vorliegt, kann abgelehnt werden, wenn das Verhältnis kleiner als -2 oder größer als +2 ist). Ein großer positiver Wert für die Kurtosis deutet darauf hin, dass die Flanken der Verteilung länger sind als bei einer Normalverteilung; ein negativer Wert bedeutet, dass sie kürzer sind (etwa wie bei einer kastenförmigen, gleichförmigen Verteilung).

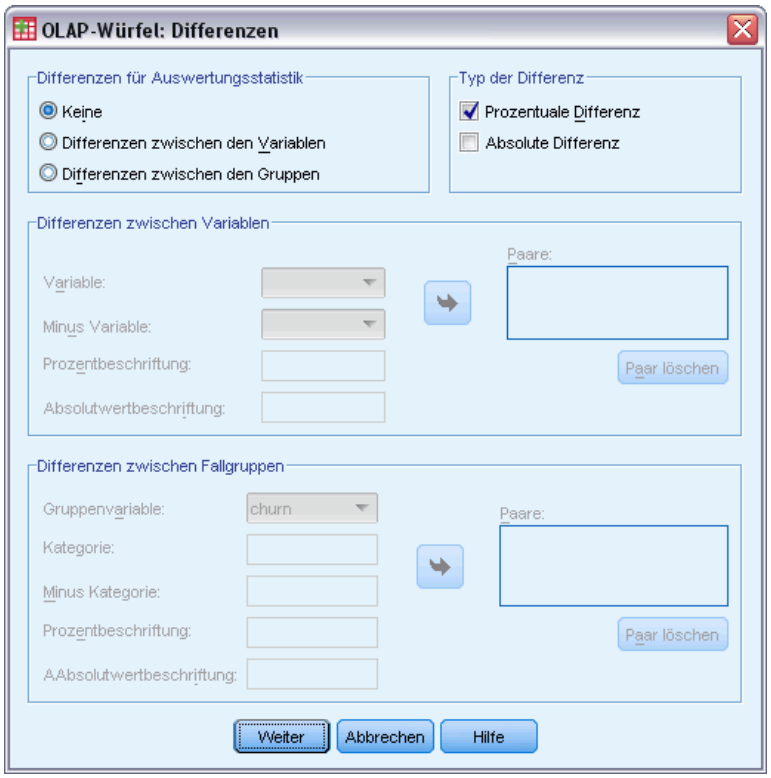
Standardfehler der Schiefe. Das Verhältnis der Schiefe zu ihrem Standardfehler kann für einen Test auf Normalverteilung verwendet werden (d. h. die Annahme, dass Normalverteilung vorliegt, kann abgelehnt werden, wenn das Verhältnis kleiner als -2 oder größer als +2 ist). Ein großer positiver Wert für die Schiefe bedeutet, dass die Verteilung eine lange rechte Flanke hat; ein extremer negativer Wert bedeutet, dass sie eine lange linke Flanke hat.

Summe. Die Summe der Werte über alle Fälle mit nichtfehlenden Werten.

Varianz. Ein Maß der Streuung um den Mittelwert. Es ist gleich dem Quotienten aus der Summe der quadrierten Abweichung vom Mittelwert und der um 1 verringerten Fallanzahl. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.

OLAP-Würfel: Differenzen

Abbildung 8-3
Dialogfeld "OLAP-Würfel: Differenzen"



In diesem Dialogfeld können Sie prozentuale und arithmetische Differenzen zwischen Auswertungsvariablen oder zwischen Gruppen berechnen lassen, die durch eine Gruppenvariable definiert sind. Die Differenzen werden für alle Maße berechnet, die im Dialogfeld "OLAP-Würfel: Statistiken" ausgewählt wurden.

Differenzen zwischen den Variablen. Hiermit werden die Differenzen zwischen Variablenpaaren berechnet. Die Werte der Auswertungsstatistik für die zweite Variable (die Minusvariable) in jedem Paar werden von den Werten der Auswertungsstatistik für die erste Variable im Paar subtrahiert. Bei prozentualen Differenzen wird der Wert der Auswertungsvariable für die

Minusvariable als Nenner verwendet. Sie müssen mindestens zwei Auswertungsvariablen im Hauptdialogfeld auswählen, bevor Sie die Differenzen zwischen den Variablen angeben können.

Differenzen zwischen Fallgruppen. Hiermit werden die Differenzen zwischen Gruppenpaaren berechnet, die durch eine Gruppenvariable definiert sind. Die Werte der Auswertungsstatistik für die zweite Kategorie (die Minuskategorie) in jedem Paar werden von den Werten der Auswertungsstatistik für die erste Kategorie im Paar subtrahiert. Bei prozentualen Differenzen wird der Wert der Auswertungsstatistik für die Minuskategorie als Nenner verwendet. Sie müssen mindestens eine Gruppenvariable im Hauptdialogfeld auswählen, bevor Sie die Differenzen zwischen den Gruppen angeben können.

OLAP-Würfel: Titel

Abbildung 8-4
Dialogfeld "OLAP-Würfel: Titel"

Sie können den Titel der Ausgabe ändern oder eine Erklärung hinzufügen, die unter der Ausgabetabelle angezeigt wird. Sie können auch den Zeilenumbruch in Titeln und Erklärungen selbst bestimmen, indem Sie an der gewünschten Stelle im Text die Zeichenfolge \n eingeben.

Kapitel
9

T-Tests

Es sind drei Typen von *T*-Tests verfügbar:

T-Test bei unabhängigen Stichproben (T-Test bei zwei Stichproben). Vergleicht die Mittelwerte einer Variablen für zwei Fallgruppen. Für jede Gruppe sind beschreibende Statistiken und der Levene-Test auf Gleichheit der Varianzen sowie *t*-Werte für gleiche und verschiedene Varianzen und ein 95%-Konfidenzintervall für die Differenz der Mittelwerte verfügbar.

T-Test bei gepaarten Stichproben (T-Test für abhängige Variablen). Vergleicht den Mittelwert von zwei Variablen für eine einzelne Gruppe. Dieser Test ist auch für Studien mit zugeordneten Paaren oder Fallkontrolle geeignet. Die Ausgabe enthält deskriptive Statistiken für die Testvariablen, die Korrelationen zwischen den Variablen, deskriptive Statistiken für die gepaarten Differenzen, den *T*-Test und ein 95%-Konfidenzintervall.

T-Test bei einer Stichprobe. Vergleicht den Mittelwert einer Variablen mit einem bekannten oder angenommenen Wert. Neben dem *T*-Test werden deskriptive Statistiken für die Testvariablen angezeigt. In der Standardeinstellung wird unter anderem ein 95%-Konfidenzintervall für die Differenz zwischen dem Mittelwert der Testvariablen und dem angenommenen Testwert ausgegeben.

T-Test bei unabhängigen Stichproben

Im T-Test bei unabhängigen Stichproben werden die Mittelwerte von zwei Fallgruppen verglichen. Im Idealfall sollten die Subjekte bei diesem Test zufällig zwei Gruppen zugeordnet werden, sodass Unterschiede bei den Antworten lediglich auf die Behandlung (bzw. Nichtbehandlung) und keine sonstigen Faktoren zurückzuführen sind. Dies ist nicht der Fall, wenn Sie die Durchschnittseinkommen von Männern und Frauen vergleichen. Die jeweiligen Personen sind nicht zufällig auf die Gruppen "männlich" oder "weiblich" verteilt. In solchen Situationen müssen Sie sicherstellen, dass signifikante Differenzen der Mittelwerte nicht durch Abweichungen bei anderen Faktoren verborgen oder verstärkt werden. Unterschiede im Durchschnittseinkommen können auch durch Faktoren wie den Bildungsstand beeinflußt werden (nicht nur durch das Geschlecht).

Beispiel. Patienten mit hohem Blutdruck werden zufällig auf eine Kontrollgruppe und eine Versuchsgruppe verteilt. Die Patienten in der Kontrollgruppe erhalten ein Plazebo. Die Patienten der Versuchsgruppe erhalten ein neues Medikament, dessen blutdrucksenkende Wirkung erprobt werden soll. Nach zweimonatiger Behandlung wird der *T*-Test bei zwei Stichproben angewandt, um den durchschnittlichen Blutdruck der Personen in der Kontrollgruppe mit dem der Personen aus der Versuchsgruppe zu vergleichen. Bei jedem Patienten wird eine Messung vorgenommen, und er gehört zu jeweils einer (1) Gruppe.

Statistiken.Für jede Variable: Stichprobengröße, Mittelwert, Standardabweichung und Standardfehler des Mittelwerts. Für die Differenz der Mittelwerte: Mittelwert, Standardfehler und Konfidenzintervall. (Sie können das Konfidenzniveau bestimmen.) Tests: Levene-Test auf Gleichheit der Varianzen sowie *t*-Tests auf Gleichheit der Mittelwerte bei gemeinsamen und separaten Varianzen.

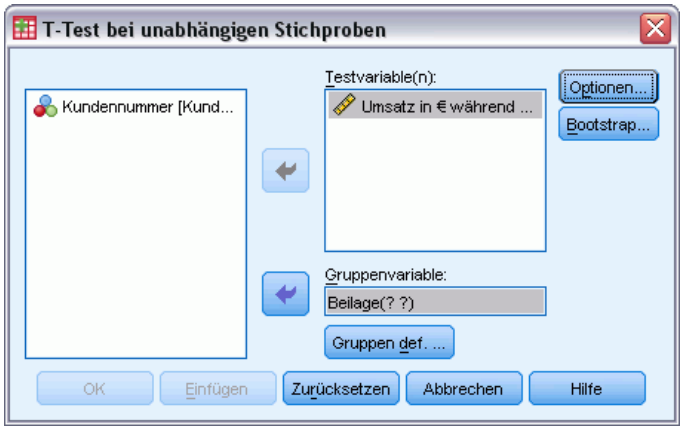
Daten. Die Werte der untersuchten quantitativen Variablen müssen in einer einzelnen Spalte in der Datendatei vorliegen. Die Prozedur verwendet eine Gruppenvariable mit zwei Werten zur Aufteilung der Fälle in zwei Gruppen. Die Gruppenvariable kann numerische Werte (wie zum Beispiel 1 und 2 oder 6,25 und 12,5) oder kurze Strings (beispielsweise *Ja* und *Nein*) enthalten. Alternativ können Sie eine quantitative Variable wie z. B. *Alter* verwenden und die Fälle durch Angabe eines Trennwerts aufteilen (der Trennwert 21 teilt *Alter* in eine Gruppe "unter 21" und eine "21 und darüber").

Annahmen.Für den *T*-Test auf Gleichheit der Varianzen sollten die Beobachtungen unabhängige Zufallsstichproben aus Normalverteilungen mit derselben Varianz der Grundgesamtheit sein. Für den *T*-Test auf Ungleichheit der Varianzen sollten die Beobachtungen unabhängige Zufallsstichproben aus Normalverteilungen sein. Der *T*-Test mit zwei Stichproben ist relativ robust gegenüber Abweichungen von der Normalverteilung. Achten Sie bei der grafischen Überprüfung von Verteilungen darauf, dass diese symmetrisch sind und keine Ausreißer enthalten.

So lassen Sie einen T-Test bei unabhängigen Stichproben berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Mittelwerte vergleichen > T-Test bei unabhängigen Stichproben...

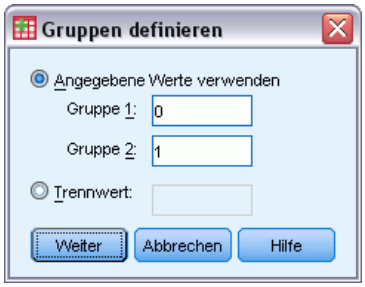
Abbildung 9-1
Dialogfeld "T-Test bei unabhängigen Stichproben"



- Wählen Sie mindestens eine quantitative Testvariable. Für jede Variable wird ein separater *T*-Test berechnet.
- Wählen Sie eine einzelne Gruppenvariable aus und klicken Sie dann auf Gruppen def., um zwei Codes für die zu vergleichenden Gruppen anzugeben.
- Zusätzlich können Sie auf Optionen klicken, um die Behandlung fehlender Daten und das Niveau des Konfidenzintervalls festzulegen.

T-Test bei unabhängigen Stichproben: Gruppen definieren

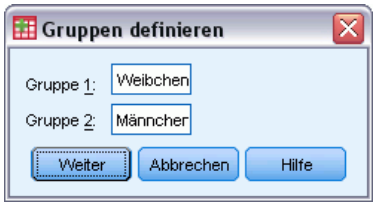
Abbildung 9-2
Dialogfeld "Gruppen definieren" für numerische Variablen



Definieren Sie bei numerischen Gruppenvariablen die zwei Gruppen für den *t*-Test, indem Sie zwei Werte oder einen Trennwert angeben:

- **Angegebene Werte verwenden.** Geben Sie einen Wert für Gruppe 1 und einen weiteren Wert für Gruppe 2 ein. Fälle mit anderen Werten werden aus der Analyse ausgeschlossen. Zahlen müssen nichtganzzahlig sein (so sind beispielsweise 6,25 und 12,5 gültige Werte).
- **Trennwert.** Geben Sie eine Zahl ein, welche die Werte der Gruppenvariablen in zwei Mengen aufteilt. Alle Fälle mit Werten, die kleiner als der Trennwert sind, bilden eine Gruppe. Die Fälle mit Werten größer oder gleich dem Trennwert bilden die andere Gruppe.

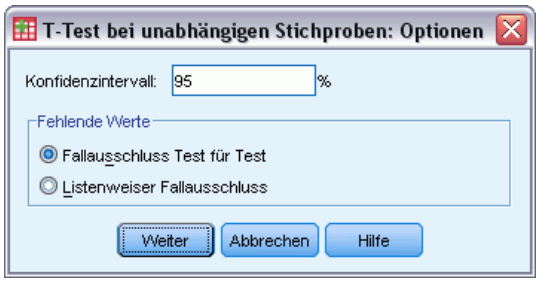
Abbildung 9-3
Dialogfeld "Gruppen definieren" für String-Variablen



Bei String-Gruppenvariablen geben Sie einen String für Gruppe 1 und einen anderen für Gruppe 2 ein, beispielsweise *Ja* und *Nein*. Fälle mit anderen Strings werden von der Analyse ausgeschlossen.

T-Tests bei unabhängigen Stichproben: Optionen

Abbildung 9-4
Dialogfeld "T-Test bei unabhängigen Stichproben: Optionen"



Konfidenzintervall. In der Standardeinstellung wird ein 95%-Konfidenzintervall für die Differenz der Mittelwerte angezeigt. Geben Sie einen Wert zwischen 1 und 99 ein, um ein anderes Konfidenzniveau festzulegen.

Fehlende Werte. Wenn Sie mehrere Variablen testen und bei einer oder mehreren Variablen Daten fehlen, können Sie bestimmen, welche Fälle einzuschließen (oder auszuschließen) sind.

- **Fallausschluss Test für Test.** Bei jedem *T*-Test werden alle Fälle verwendet, für die gültige Daten für die getestete Variable vorliegen. Die Stichprobengröße kann von Test zu Test unterschiedlich ausfallen.
- **Listenweiser Fallausschluss.**Jeder *T*-Test verwendet nur Fälle mit gültigen Daten für alle in den angeforderten *T*-Tests verwendeten Variablen. Die Stichprobengröße bleibt bei allen Tests konstant.

T-Test bei gepaarten Stichproben

Mit der Prozedur "T-Test bei gepaarten Stichproben" werden die Mittelwerte zweier Variablen für eine einzelne Gruppe verglichen. Diese Prozedur berechnet für jeden Fall die Differenzen zwischen den Werten der zwei Variablen und überprüft, ob der Durchschnitt von 0 abweicht.

Beispiel. In einer Studie über Bluthochdruck wird der Blutdruck aller Patienten zu Beginn der Studie und nach der Behandlung gemessen. Daher gibt es für jede Testperson zwei Messwerte, die auch als *Vorher*- und *Nachher*-Messung bezeichnet werden. Dieser Test kann auch bei Studien mit zugeordneten Paaren bzw. mit Fallkontrolle verwendet werden. Hierbei enthält jeder Datensatz der Datendatei die Reaktion des Patienten und die von der zugehörigen Kontroll-Testperson. In einer Blutdruckstudie könnten den Patienten die Kontrollpersonen nach Alter zugeordnet werden (einem 75-jährigen Patienten ein 75-jähriges Mitglied der Kontrollgruppe).

Statistiken.Für jede Variable: Mittelwert, Stichprobengröße, Standardabweichung und Standardfehler des Mittelwerts. Für jedes Variablenpaar: Korrelation, durchschnittliche Differenz der Mittelwerte, *T*-Test und Konfidenzintervall für die Differenz der Mittelwerte. (Sie können das Konfidenzniveau festlegen.) Standardabweichung und Standardfehler der Differenz der Mittelwerte.

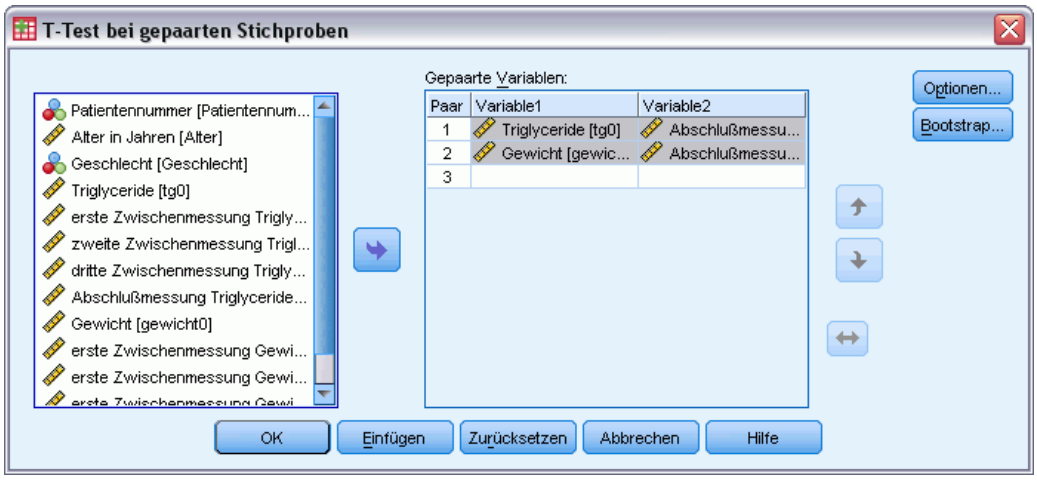
Daten. Legen Sie für jeden gepaarten Test zwei Variablen fest, die auf Intervall-Messniveau oder Verhältnis-Messniveau quantitativ sein müssen. In einer Studie mit zugeordneten Paaren bzw. mit Fallkontrolle müssen die Reaktionen jedes Testsubjektes und dessen zugeordneten Kontrollsubjektes im selben Fall der Datendatei enthalten sein.

Annahmen. Die Beobachtungen für jedes Paar müssen unter gleichen Bedingungen vorgenommen werden. Die Differenzen der Mittelwerte müssen normalverteilt sein. Die Varianzen jeder Variablen können gleich oder ungleich sein.

So lassen Sie einen T-Test bei gepaarten Stichproben berechnen:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Mittelwerte vergleichen > T-Test bei verbundenen Stichproben...

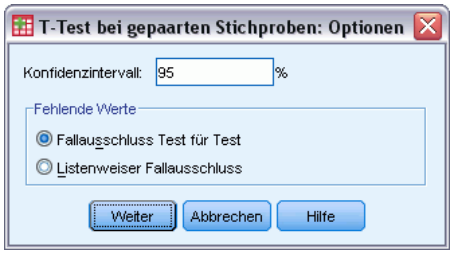
Abbildung 9-5
Dialogfeld "T-Test bei gepaarten Stichproben"



- Wählen Sie mindestens ein Variablenpaar aus
- Zusätzlich können Sie auf Optionen klicken, um die Behandlung fehlender Daten und das Niveau des Konfidenzintervalls festzulegen.

T-Test bei gepaarten Stichproben: Optionen

Abbildung 9-6
Dialogfeld "T-Test bei gepaarten Stichproben: Optionen"



Konfidenzintervall. In der Standardeinstellung wird ein 95%-Konfidenzintervall für die Differenz der Mittelwerte angezeigt. Geben Sie einen Wert zwischen 1 und 99 ein, um ein anderes Konfidenzniveau festzulegen.

Fehlende Werte. Wenn Sie mehrere Variablen testen und bei einer oder mehreren Variablen Daten fehlen, können Sie bestimmen, welche Fälle einzuschließen (oder auszuschließen) sind:

- **Fallausschluss Test für Test.** Bei jedem *T*-Test werden alle Fälle mit gültigen Daten für die getesteten Variablenpaare verwendet. Die Stichprobengröße kann von Test zu Test unterschiedlich ausfallen.
- **Listenweiser Fallausschluss.** Bei jedem *T*-Test werden nur Fälle mit gültigen Daten für alle getesteten Variablenpaare verwendet. Die Stichprobengröße bleibt bei allen Tests konstant.

T-Test bei einer Stichprobe

Die Prozedur “T-Test bei einer Stichprobe” prüft, ob der Mittelwert einer einzelnen Variablen von einer angegebenen Konstanten abweicht.

Beispiele. Ein Forscher könnte testen, ob der durchschnittliche IQ-Wert einer Gruppe von Studenten von 100 abweicht. Ein Hersteller von Getreideprodukten könnte stichprobenartig Packungen aus der Produktion entnehmen und prüfen, ob das Durchschnittsgewicht der Stichproben auf dem 95%-Konfidenzniveau von 500 Gramm abweicht.

Statistiken.Für jede Testvariable: Mittelwert, Standardabweichung und Standardfehler der Differenz der Mittelwerte. Außerdem die durchschnittliche Differenz zwischen jedem Datenwert und dem angenommenen Testwert, ein *T*-Test, der prüft, ob diese Differenz null beträgt, und ein Konfidenzintervall für diese Differenz. (Sie können das Konfidenzniveau festlegen.)

Daten. Um die Werte einer quantitativen Variablen mit einem angenommenen Testwert zu vergleichen, wählen Sie eine quantitative Variable aus und geben Sie einen angenommenen Testwert ein.

Annahmen. Bei diesem Test wird von einer Normalverteilung ausgegangen; er ist jedoch recht robust gegenüber Abweichungen von dieser Verteilung.

So lassen Sie den T-Test bei einer Stichprobe berechnen:

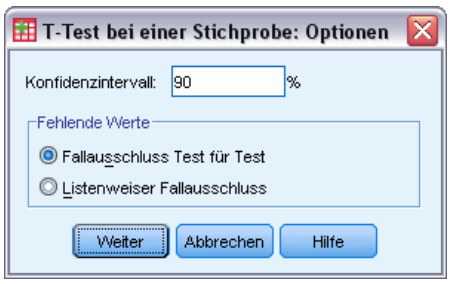
- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Mittelwerte vergleichen > T-Test bei einer Stichprobe...

Abbildung 9-7
Dialogfeld “T-Test bei einer Stichprobe”

- Wählen Sie eine oder mehrere Variablen aus, die mit demselben angenommenen Wert verglichen werden sollen.
- Geben Sie einen numerischen Testwert ein, mit dem jeder Stichprobenmittelwert verglichen werden soll.
- Zusätzlich können Sie auf Optionen klicken, um die Behandlung fehlender Daten und das Niveau des Konfidenzintervalls festzulegen.

T-Test bei einer Stichprobe: Optionen

Abbildung 9-8
Dialogfeld "T-Test bei einer Stichprobe: Optionen"



Konfidenzintervall. In der Standardeinstellung wird ein 95%-Konfidenzintervall für die Differenz zwischen dem Mittelwert und dem angenommenen Testwert angezeigt. Geben Sie einen Wert zwischen 1 und 99 ein, um ein anderes Konfidenzniveau festzulegen.

Fehlende Werte. Wenn Sie mehrere Variablen testen und bei einer oder mehreren Variablen Daten fehlen, können Sie bestimmen, welche Fälle einzuschließen (oder auszuschließen) sind.

- **Fallausschluss Test für Test.** Bei jedem *T*-Test werden alle Fälle verwendet, die gültige Daten für die getestete Variable aufweisen. Die Stichprobengröße kann von Test zu Test unterschiedlich ausfallen.
- **Listenweiser Fallausschluss.** Jeder *T*-Test verwendet nur Fälle, die gültige Daten für alle Variablen aufweisen, die in einem der angeforderten *T*-Tests verwendet werden. Die Stichprobengröße bleibt bei allen Tests konstant.

Zusätzliche Funktionen beim Befehl T-TEST

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Erstellen von T-Tests für eine Stichprobe sowie für unabhängige Stichproben mit einem einzigen Befehl.
- Testen einer Variablen gegen alle Variablen in einer Liste mit einem gepaarten t-Test (mit dem Unterbefehl `PAIRS`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
10

Einfaktorielle ANOVA

Die Prozedur Einfaktorielle ANOVA führt eine einfaktorielle Varianzanalyse für eine quantitative abhängige Variable mit einer einzelnen (unabhängigen) Faktorvariablen durch. Mit der Varianzanalyse wird die Hypothese überprüft, dass mehrere Mittelwerte gleich sind. Dieses Verfahren ist eine Erweiterung des *T*-Tests bei zwei Stichproben.

Sie können zusätzlich zur Feststellung, dass Differenzen zwischen Mittelwerten vorhanden sind, auch bestimmen, welche Mittelwerte abweichen. Für den Vergleich von Mittelwerten gibt es zwei Arten von Tests: A-priori-Kontraste und Post-Hoc-Tests. Kontraste sind Tests, die *vor* der Ausführung des Experiments eingerichtet werden, Post-Hoc-Tests werden *nach* dem Experiment ausgeführt. Sie können auch auf Trends für mehrere Kategorien testen.

Beispiel. Paniertes Fleisch absorbiert beim Fritieren unterschiedliche Mengen an Fett. Ein Experiment wird mit den folgenden drei Fettsorten durchgeführt: Distelöl, Maiskeimöl und Schmalz. Distelöl und Maiskeimöl sind ungesättigte Fette, Schmalz ist ein gesättigtes Fett. Sie können bestimmen, ob die Menge des absorbierten Fetts von der Fettsorte abhängt. Gleichzeitig können Sie einen A-priori-Kontrast einrichten, um zu ermitteln, ob sich die absorbierte Fettmenge bei gesättigten und ungesättigten Fetten unterscheidet.

Statistiken. Für jede Gruppe: Anzahl der Fälle, Mittelwert, Standardabweichung, Standardfehler des Mittelwerts, Minimum, Maximum und 95%-Konfidenzintervall für den Mittelwert. Levene-Test auf Homogenität der Varianzen, Varianzanalyse-Tabellen und zuverlässige Tests auf Gleichheit der Mittelwerte für jede abhängige Variable, benutzerspezifische A-priori-Kontraste, Post-Hoc-Spannweitentests und Mehrfachvergleiche: Bonferroni, Sidak, ehrlich signifikante Differenz nach Tukey, GT2 nach Hochberg, Gabriel, *F*-Test nach Dunnett, Ryan-Einot-Gabriel-Welsch (*F* nach R-E-G-W), Spannweitentest nach Ryan-Einot-Gabriel-Welsch (*Q* nach R-E-G-W), Tamhane-T2, Dunnett-T3, Games-Howell, Dunnett-*C*, Duncans multipler Spannweitentest, Student-Newman-Keuls (S-N-K), Tukey-*b*, Waller-Duncan, Scheffé und geringste signifikante Differenz.

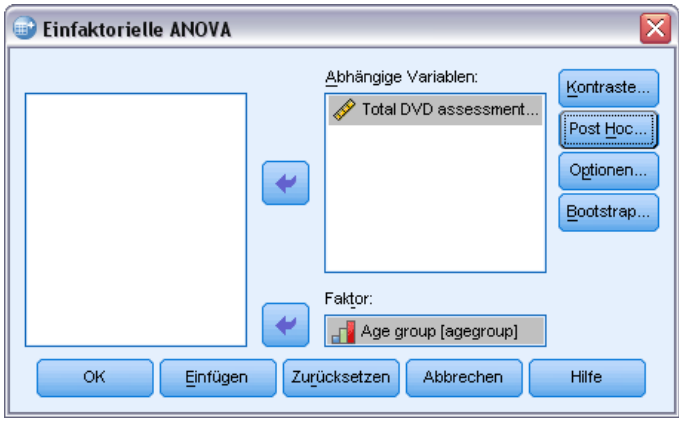
Daten. Die Werte der Faktorvariablen müssen ganzzahlig sein, die abhängige Variable muss quantitativ sein (Messung auf Intervallebene).

Annahmen. Jede Gruppe bildet eine unabhängige zufällige Stichprobe aus einer normalverteilten Grundgesamtheit. Die Varianzanalyse ist unempfindlich gegenüber Abweichungen von der Normalverteilung. Die Daten müssen jedoch symmetrisch verteilt sein. Die Gruppen müssen aus Grundgesamtheiten mit gleichen Varianzen stammen. Sie überprüfen diese Annahme mithilfe des Levene-Tests auf Homogenität der Varianzen.

So lassen Sie eine einfaktorielle ANOVA berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Mittelwerte vergleichen > Einfaktorielle ANOVA...

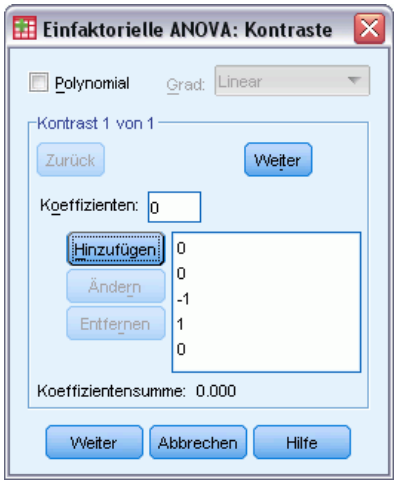
Abbildung 10-1
Dialogfeld "Einfaktorielle ANOVA"



- Wählen Sie eine oder mehrere abhängige Variablen aus.
- Wählen Sie eine unabhängige Faktorvariable aus.

Einfaktorielle ANOVA: Kontraste

Abbildung 10-2
Dialogfeld "Einfaktorielle ANOVA: Kontraste"



Sie können die Quadratsummen zwischen den Gruppen in Trend-Komponenten zerlegen oder A-priori-Kontraste festlegen.

Polynomial. Damit zerlegen Sie die Quadratsummen zwischen den Gruppen in Trend-Komponenten. Sie können die abhängige Variable auf einen Trend über die geordneten Stufen der Faktorvariablen prüfen. Sie können beispielsweise prüfen, ob beim Gehalt über die

geordneten Stufen des höchsten erreichten akademischen Grads ein linearer (steigender oder fallender) Trend vorliegt.

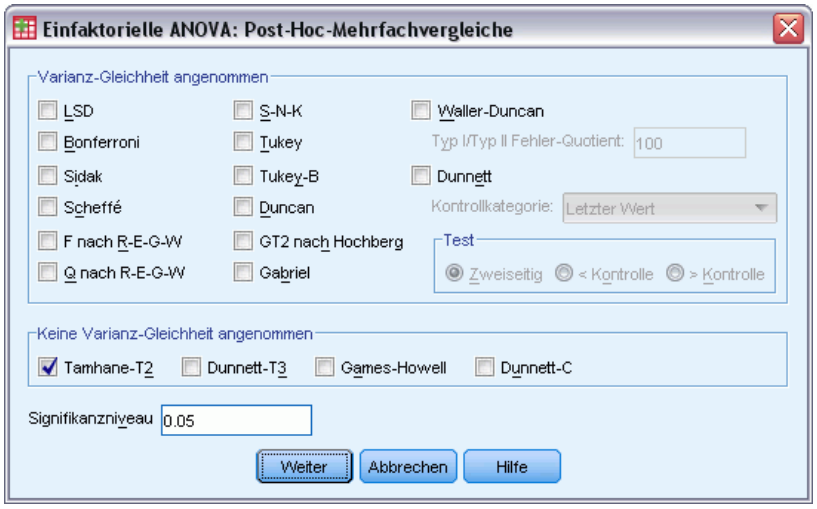
- **Grad.** Sie können Polynome ersten, zweiten, dritten, vierten und fünften Grades auswählen.

Koeffizienten. Mit der *T*-Statistik werden benutzerdefinierte A-priori-Kontraste getestet. Geben Sie für jede Gruppe (Kategorie) der Faktorvariablen einen Koeffizienten ein und klicken Sie nach jeder Eingabe auf Hinzufügen. Jeder neue Wert wird am Ende der Liste der Koeffizienten hinzugefügt. Um zusätzliche Kontrastgruppen festzulegen, klicken Sie auf Weiter. Verwenden Sie Weiter und Zurück, um zwischen den Kontrastgruppen zu wechseln.

Die Reihenfolge der Koeffizienten ist wichtig, weil sie den aufsteigend geordneten Kategoriewerten der Faktorvariablen entspricht. Der erste Koeffizient der Liste entspricht dem kleinsten Gruppenwert der Faktorvariablen, der letzte Koeffizient dem größten Wert. Bei zum Beispiel sechs Kategorien der Faktorvariablen stellen die Koeffizienten –1, 0, 0, 0,5 und 0,5 einen Kontrast zwischen der ersten und der fünften und sechsten Gruppe her. Bei den meisten Anwendungen muss die Summe der Koeffizienten 0 ergeben. Sie können auch Werte benutzen, deren Summe ungleich 0 ist. In diesem Fall wird jedoch eine Warnung angezeigt.

Einfaktorielle ANOVA: Post-Hoc-Mehrfachvergleiche

Abbildung 10-3
Dialogfeld "Einfaktorielle ANOVA: Post-Hoc-Mehrfachvergleiche"



Sobald Sie festgestellt haben, dass es Abweichungen zwischen den Mittelwerten gibt, können Sie mit Post-Hoc-Spannweiten-Tests und paarweisen multiplen Vergleichen untersuchen, welche Mittelwerte sich unterscheiden. Spannweitentests ermitteln homogene Untergruppen von Mittelwerten, die nicht voneinander abweichen. Mit paarweisen Mehrfachvergleichen testen Sie die Differenz zwischen gepaarten Mittelwerten. Die Ergebnisse werden in einer Matrix angezeigt, in der Gruppenmittelwerte, die auf einem Alpha-Niveau von 0,05 signifikant voneinander abweichen, durch Sterne markiert sind.

Varianz-Gleichheit angenommen

Die ehrlich signifikante Differenz nach Tukey, der GT2 nach Hochberg, der Gabriel-Test und der Scheffé-Test sind Tests für Mehrfachvergleiche und Spannweitentests. Andere Spannweitentests sind Tukey-*B*, S-N-K (Student-Newman-Keuls), Duncan, *F* nach R-E-G-W (*F*-Test nach Ryan-Einot-Gabriel-Welsch), *Q* nach R-E-G-W (Spannweitentest nach Ryan-Einot-Gabriel-Welsch) und Waller-Duncan. Verfügbare Tests für Mehrfachvergleiche sind Bonferroni, ehrlich signifikante Differenz nach Tukey, Sidak, Gabriel, Hochberg, Dunnett, Scheffé und LSD (geringste signifikante Differenz).

- **LSD.** Verwendet T-Tests, um alle paarweisen Vergleiche zwischen Gruppenmittelwerten durchzuführen. Es erfolgt keine Korrektur der Fehlerrate bei Mehrfachvergleichen.
- **Bonferroni.** Führt paarweise Vergleiche zwischen Gruppenmittelwerten mit T-Tests aus; regelt dabei jedoch auch die Gesamtfehlerrate, indem die Fehlerrate für jeden Test auf den Quotienten aus der experimentellen Fehlerrate und der Gesamtzahl der Tests gesetzt wird. Dadurch wird das beobachtete Signifikanzniveau für Mehrfachvergleiche angepasst.
- **Sidak.** Ein paarweiser multipler Vergleichstest, basierend auf einer T-Statistik. Beim Sidak-Test wird das Signifikanzniveau für die multiplen Vergleiche korrigiert und es werden engere Grenzen vergeben als bei Bonferroni.
- **Scheffé.** Führt gemeinsame paarweise Vergleiche gleichzeitig für alle möglichen paarweisen Kombinationen der Mittelwerte durch. Verwendet die F-Stichprobenverteilung. Dieser Test kann verwendet werden, um nicht nur paarweise Vergleiche durchzuführen, sondern alle möglichen linearen Kombinationen von Gruppenmittelwerten zu untersuchen.
- **F nach R-E-G-W.** Mehrfaches Rückschrittverfahren nach Ryan-Einot-Gabriel-Welsh, basierend auf einem F-Test.
- **Q nach R-E-G-W.** Mehrfaches Rückschrittverfahren nach Ryan-Einot-Gabriel-Welsh, basierend auf der studentisierten Spannweite.
- **S-N-K.** Führt alle paarweisen Vergleiche zwischen Mittelwerten unter Verwendung der t-Verteilung aus. Bei gleich großen Stichproben werden auch die Mittelwertpaare innerhalb homogener Untergruppen verglichen; dabei wird ein schrittweises Verfahren verwendet. Die Mittelwerte werden in absteigender Reihenfolge (vom größten zum kleinsten Wert) sortiert, extreme Differenzen werden zuerst getestet.
- **Tukey.** Verwendet die Student-Verteilung für alle möglichen paarweisen Vergleiche zwischen den Gruppen. Setzt die Fehlerrate für das Experiment gleich der Fehlerrate für die Gesamtheit aller paarweisen Vergleiche.
- **Tukey-B-Test.** Verwendet die Student-Verteilung für paarweise Vergleiche zwischen Gruppen. Der kritische Wert ist der Durchschnitt des entsprechenden Werts für die ehrlich signifikante Differenz nach Tukey und für Student-Newman-Keuls.
- **Duncan.** Bei diesem Test werden paarweise Vergleiche angestellt, deren schrittweise Reihenfolge identisch ist mit der Reihenfolge, die beim Student-Newman-Keuls-Test verwendet wird. Abweichend wird aber ein Sicherheitsniveau für die Fehlerrate der zusammengefassten Tests statt einer Fehlerrate für die einzelnen Tests festgelegt. Es wird die studentisierte Bereichsstatistik verwendet.
- **GT2 nach Hochberg.** Ein paarweiser Vergleichstest, der auf dem studentisierten Maximalmodul beruht. Ähnelt dem Test auf ehrlich signifikante Differenz nach Tukey.

- **Gabriel.** Ein paarweiser Vergleichstest, der das studentisierte Maximalmodul verwendet. Er ist in der Regel aussagekräftiger als der GT2-Test nach Hochberg, wenn unterschiedliche Zellengrößen vorliegen. Der Gabriel-Test kann ungenau werden, wenn die Zellengrößen stark variieren.
- **Waller-Duncan.** Ein Test für Mehrfachvergleiche auf der Grundlage einer T-Statistik; verwendet eine Bayes-Methode.
- **Dunnett.** Ein paarweiser T-Test für Mehrfachvergleiche, der ein Set von Verarbeitungen mit einem einzelnen Kontrollmittelwert vergleicht. Als Kontrollkategorie ist die letzte Kategorie voreingestellt. Sie können aber auch die erste Kategorie einstellen. Verwenden Sie einen zweiseitigen Test, um zu überprüfen, ob sich der Mittelwert bei jeder Stufe (außer der Kontrollkategorie) des Faktors von dem Mittelwert der Kontrollkategorie unterscheidet. Wählen Sie <>Kontrolle, um zu überprüfen, ob der Mittelwert bei allen Stufen des Faktors kleiner als der Mittelwert der Kontrollkategorie ist. Wählen Sie >Kontrolle, um zu überprüfen, ob der Mittelwert bei allen Stufen des Faktors größer als der Mittelwert der Kontrollkategorie ist.

Keine Varianz-Gleichheit angenommen

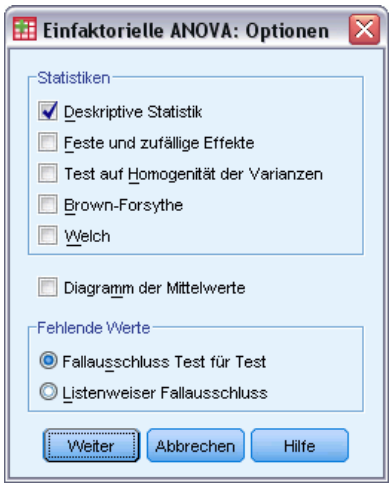
Tests für Mehrfachvergleiche, die keine Varianzgleichheit voraussetzen, sind Tamhane-T2, Dunnett-T3, Games-Howell und Dunnett-*C*.

- **Tamhane-T2.** Konservative, paarweise Vergleichstests auf der Grundlage eines T-Tests. Dieser Test ist für ungleiche Varianzen geeignet.
- **Dunnett-T3.** Ein paarweiser Vergleichstest, der auf dem studentisierten Maximalmodul beruht. Dieser Test ist für ungleiche Varianzen geeignet.
- **Games-Howell.** Ein manchmal schwacher, paarweiser Vergleichstest. Dieser Test ist für ungleiche Varianzen geeignet.
- **Dunnett-C.** Ein paarweiser Vergleichstest, der auf dem studentisierten Bereich beruht. Dieser Test ist für ungleiche Varianzen geeignet.

Hinweis: Die Ausgabe von Post-Hoc-Tests läßt sich oft einfacher interpretieren, wenn Sie im Dialogfeld “Tabelleneigenschaften” die Option Leere Zeilen und Spalten ausblenden deaktivieren. (In einer aktivierten Pivot-Tabelle: Tabelleneigenschaften im Menü “Format”.)

Einfaktorielle ANOVA: Optionen

Abbildung 10-4
Dialogfeld "Einfaktorielle ANOVA: Optionen"



Statistiken. Wählen Sie mindestens eine der folgenden Optionen aus:

- **Deskriptive Statistik.** Hiermit berechnen Sie Anzahl der Fälle, Mittelwert, Standardabweichung, Standardfehler des Mittelwerts, Minimum, Maximum und das 95%-Konfidenzintervall für jede abhängige Variable in jeder Gruppe.
- **Feste und zufällige Effekte.** Hiermit werden die Standardabweichung, der Standardfehler und das 95%-Konfidenzintervall für das Modell mit festen Effekten sowie der Standardfehler, das 95%-Konfidenzintervall und der Schätzer der Varianz zwischen Komponenten für das Modell mit zufälligen Effekten angezeigt.
- **Test auf Homogenität der Varianzen.** Bei dieser Option wird die Levene-Statistik berechnet, mit der Sie die Gruppenvarianzen auf Gleichheit testen können. Dieser Test setzt keine Normalverteilung voraus.
- **Brown-Forsythe.** Bei dieser Option wird die Brown-Forsythe-Statistik berechnet, mit der Sie die Gruppenmittelwerte auf Gleichheit testen können. Diese Statistik ist der *F*-Statistik vorzuziehen, wenn die Annahme gleicher Varianzen sich nicht bestätigt.
- **Welch.** Bei dieser Option wird die Welch-Statistik berechnet, mit der Sie die Gruppenmittelwerte auf Gleichheit testen können. Diese Statistik ist der *F*-Statistik vorzuziehen, wenn die Annahme gleicher Varianzen sich nicht bestätigt.

Diagramm der Mittelwerte. Bei dieser Option wird ein Diagramm für die Mittelwerte der Untergruppen ausgegeben. Dabei handelt es sich um die Mittelwerte für jede Gruppe, die durch die Werte der Faktorvariablen definiert ist.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.** Bei Auswahl dieser Option werden Fälle mit einem fehlenden Wert für die abhängige Variable oder die Faktorvariable in einer bestimmten Analyse in dieser Analyse nicht verwendet. Ein Fall wird außerdem nicht verwendet, wenn er außerhalb des Bereichs liegt, der für die Faktorvariable definiert ist.
- **Listenweiser Fallausschluss.** Fälle mit fehlenden Werten für die Faktorvariable oder eine abhängige Variable, die in der Liste der abhängigen Variablen des Hauptdialogfelds enthalten sind, werden aus allen Analysen ausgeschlossen. Wenn Sie nicht mehrere abhängige Variablen festgelegt haben, hat dies keine Auswirkung.

Zusätzliche Funktionen beim Befehl ONEWAY

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Erstellen von Statistiken mit festen und zufälligen Effekten. Standardabweichung, Standardfehler des Mittelwerts und 95%-Konfidenzintervalle für ein Modell mit festen Effekten. Standardfehler, 95%-Konfidenzintervalle und die Schätzung der Varianz zwischen Komponenten für ein Modell mit zufälligen Effekten (mit `STATISTICS=EFFECTS`).
- Angeben der Alpha-Niveaus für die Tests für Mehrfachvergleiche auf geringste signifikante Differenz sowie nach Bonferroni, Duncan und Scheffé (mit dem Unterbefehl `RANGES`).
- Schreiben einer Matrix der Mittelwerte, Standardabweichungen und Häufigkeiten oder Lesen einer Matrix der Mittelwerte, Häufigkeiten, gemeinsame Varianzen sowie der Freiheitsgrade für die gemeinsamen Varianzen. Diese Matrizen können anstellen der Rohdaten verwendet werden, um eine einfaktorielle Analyse der Varianz durchzuführen (mit dem Unterbefehl `MATRIX`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
11

GLM - Univariat

Mit der Prozedur “GLM - Univariat” können Sie Regressionsanalysen und Varianzanalysen für eine abhängige Variable mit einem oder mehreren Faktoren und/oder Variablen durchführen. Die Faktorvariablen unterteilen die Grundgesamtheit in Gruppen. Unter Verwendung dieser auf einem allgemeinen linearen Modell basierenden Prozedur können Sie Nullhypothesen über die Effekte anderer Variablen auf die Mittelwerte verschiedener Gruppierungen einer einzelnen abhängigen Variablen testen. Sie können die Wechselwirkungen zwischen Faktoren und die Effekte einzelner Faktoren untersuchen, von denen einige zufällig sein können. Außerdem können Sie die Auswirkungen von Kovariaten und Wechselwirkungen zwischen Kovariaten und Faktoren berücksichtigen. Bei der Regressionsanalyse werden die unabhängigen Variablen (Einflußvariablen) als Kovariaten angegeben.

Es können sowohl ausgeglichene als auch nicht ausgeglichene Modelle getestet werden. Ein Design ist ausgeglichen, wenn jede Zelle im Modell dieselbe Anzahl von Fällen enthält. Mit der Prozedur “GLM - Univariat” werden nicht nur Hypothesen getestet, sondern zugleich Parameter geschätzt.

Zum Testen von Hypothesen stehen häufig verwendete a-priori-Kontraste zur Verfügung. Nachdem die Signifikanz mit einem *F*-Gesamttest nachgewiesen wurde, können Sie Post-Hoc-Tests verwenden, um Differenzen zwischen bestimmten Mittelwerten berechnen zu lassen. Geschätzte Randmittel dienen als Schätzer für die vorhergesagten Mittelwerte der Zellen im Modell, und mit Profilplots (Wechselwirkungsdiagrammen) dieser Mittelwerte können Sie einige dieser Beziehungen in einfacher Weise visuell darstellen.

Residuen, Einflußwerte, die Cook-Distanz und Hebelwerte können zum Überprüfen von Annahmen als neue Variablen in der Datendatei gespeichert werden.

Mit der WLS-Gewichtung können Sie eine Variable angeben, um Beobachtungen für eine WLS-Analyse (Weighted Least Squares, deutsch: gewichtete kleinste Quadrate) unterschiedlich zu gewichten. Dies kann notwendig sein, um etwaige Unterschiede in der Präzision von Messungen auszugleichen.

Beispiel. Im Rahmen einer sportwissenschaftlichen Studie beim Berlin-Marathon werden mehrere Jahre lang Daten über einzelne Läufer aufgenommen. Die abhängige Variable ist die Zeit, die jeder Läufer für die Strecke benötigt. Andere berücksichtigte Faktoren sind beispielsweise das Wetter (kalt, angenehm oder heiß), die Anzahl von Trainingsmonaten, die Anzahl der bereits absolvierten Marathons und das Geschlecht. Das Alter der betreffenden Personen wird als Kovariate betrachtet. Ein mögliches Ergebnis wäre, dass das Geschlecht ein signifikanter Effekt und die Wechselwirkung von Geschlecht und Wetter signifikant ist.

Methoden. Zum Überprüfen der verschiedenen Hypothesen können Quadratsummen vom Typ I, Typ II, Typ III und Typ IV verwendet werden. Die Voreinstellung sieht den Typ III vor.

Statistik. Post-Hoc-Spannweitentests und Mehrfachvergleiche: geringste signifikante Differenz, Bonferroni, Sidak, Scheffé, multiples *F* nach Ryan-Einot-Gabriel-Welsch, multiple Spannweite nach Ryan-Einot-Gabriel-Welsch, Student-Newman-Keuls-Test, ehrlich signifikante Differenz nach Tukey, Tukey-*B*, Duncan, GT2 nach Hochberg, Gabriel, Waller-Duncan-*T*-Test, Dunnett (einseitig und zweiseitig), Tamhane-T2, Dunnett-T3, Games-Howell und Dunnett-*C*. Deskriptive Statistiken: beobachtete Mittelwerte, Standardabweichungen und Häufigkeiten aller abhängigen Variablen in allen Zellen. Levene-Test auf Homogenität der Varianzen.

Diagramme. Streubreite gegen mittleres Niveau, Residuen-Diagramme, Profilplots (Wechselwirkung).

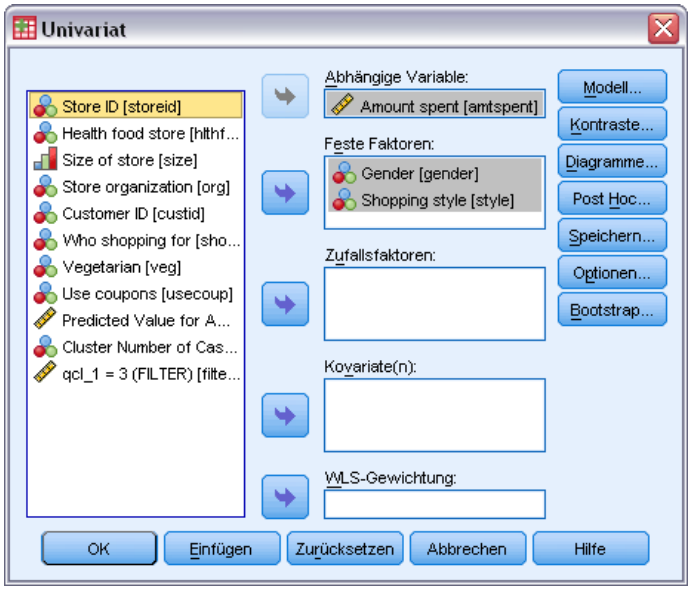
Daten. Die abhängige Variable ist quantitativ. Faktoren sind kategorial. Sie können numerische Werte oder String-Werte von bis zu acht Zeichen Länge annehmen. Kovariaten sind quantitative Variablen, die mit der abhängigen Variablen in Beziehung stehen.

Annahmen. Die Daten sind eine Stichprobe aus einer normalverteilten Grundgesamtheit. In der Grundgesamtheit sind alle Zellenvarianzen gleich. Die Varianzanalyse ist unempfindlich gegenüber Abweichungen von der Normalverteilung. Die Daten müssen jedoch symmetrisch verteilt sein. Zum Überprüfen der Annahmen können Sie Tests auf Homogenität der Varianzen vornehmen und Diagramme der Streubreite gegen das mittlere Niveau ausgeben lassen. Sie können auch die Residuen untersuchen und Residuen-Diagramme anzeigen lassen.

So berechnen Sie eine univariate Analyse der Varianz (GLM):

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Allgemeines lineares Modell > Univariat...

Abbildung 11-1
Dialogfeld "GLM - Univariat"

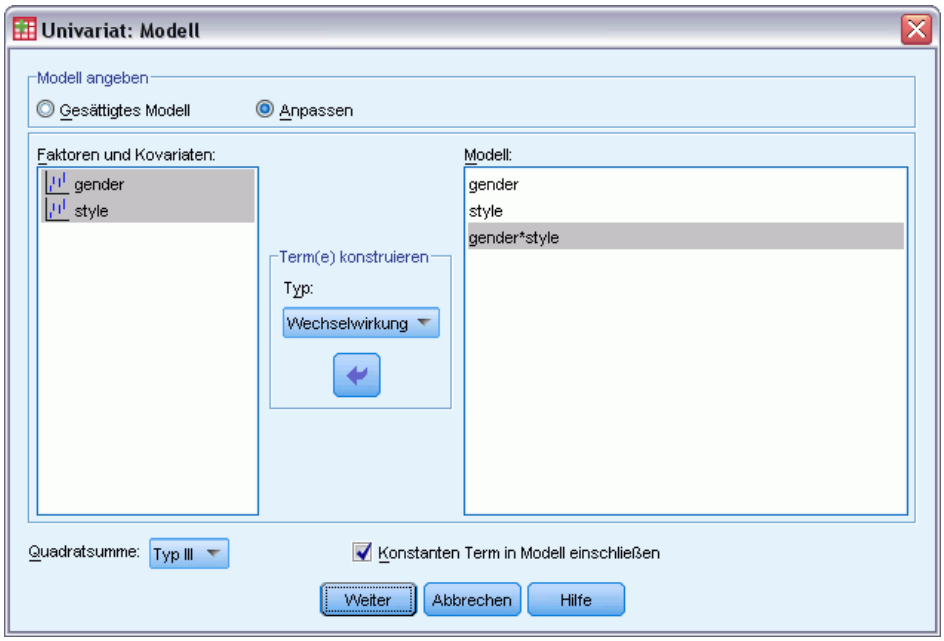


▶ Wählen Sie eine abhängige Variable aus.

- Wählen Sie in Abhängigkeit von den Daten Variablen als feste Faktoren, Zufallsfaktoren und Kovariaten aus.
- Optional können Sie mit der WLS-Gewichtung eine Gewichtungsvariable für WLS-Analyse (Weighted Least Squares, gewichtete kleinste Quadrate) angeben. Wenn der Wert der Gewichtungsvariablen null, negativ oder fehlend ist, wird der Fall aus der Analyse ausgeschlossen. Eine bereits im Model verwendete Variable kann nicht als Gewichtungsvariable verwendet werden.

GLM: Modell

Abbildung 11-2
Dialogfeld "Univariat: Modell"



Modell angeben. Ein gesättigtes Modell enthält alle Faktoren-Haupteffekte, alle Kovariaten-Haupteffekte und alle faktorweisen Wechselwirkungen. Es enthält keine Kovariaten-Wechselwirkungen. Wählen Sie Anpassen aus, um nur eine Teilmenge von Wechselwirkungen oder Wechselwirkungen zwischen Faktoren und Kovariaten festzulegen. Sie müssen alle in das Modell zu übernehmenden Terme angeben.

Faktoren und Kovariaten. Die Faktoren und Kovariaten werden aufgelistet.

Modell. Das Modell ist von der Art Ihrer Daten abhängig. Nach der Auswahl von Anpassen können Sie die Haupteffekte und Wechselwirkungen auswählen, die für Ihre Analyse von Interesse sind.

Quadratsumme. Hier wird die Methode zum Berechnen der Quadratsumme festgelegt. Für ausgeglichene und unausgeglichene Modelle ohne fehlende Zellen wird meistens die Methode mit Quadratsummen vom Typ III angewendet.

Konstanten Term in Modell einschließen. Der konstante Term wird gewöhnlich in das Modell aufgenommen. Falls Sie sicher sind, dass die Daten durch den Koordinatenursprung verlaufen, können Sie den konstanten Term ausschließen.

Terme konstruieren

Für die ausgewählten Faktoren und Kovariaten:

Wechselwirkung. Hiermit wird der Wechselwirkungsterm mit der höchsten Ordnung von allen ausgewählten Variablen erzeugt. Dies ist die Standardeinstellung.

Haupteffekte. Legt einen Haupteffekt-Term für jede ausgewählte Variable an.

Alle 2-Weg. Hiermit werden alle möglichen 2-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 3-Weg. Hiermit werden alle möglichen 3-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 4-Weg. Hiermit werden alle möglichen 4-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 5-Weg. Hiermit werden alle möglichen 5-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Quadratsumme

Für das Modell können Sie einen Typ von Quadratsumme auswählen. Typ III wird am häufigsten verwendet und ist die Standardeinstellung.

Typ I. Diese Methode ist auch als die Methode der hierarchischen Zerlegung der Quadratsummen bekannt. Jeder Term wird nur für den Vorläuferterm im Modell angepaßt. Quadratsummen vom Typ I werden gewöhnlich in den folgenden Situationen verwendet:

- Ein ausgeglichenes ANOVA-Modell, in dem alle Haupteffekte vor den Wechselwirkungseffekten 1. Ordnung festgelegt werden, alle Wechselwirkungseffekte 1. Ordnung wiederum vor den Wechselwirkungseffekten 2. Ordnung festgelegt werden und so weiter.
- Ein polynomiales Regressionsmodell, in dem alle Terme niedrigerer Ordnung vor den Termen höherer Ordnung festgelegt werden.
- Ein rein verschachteltes Modell, in welchem der zuerst bestimmte Effekt in dem als zweiten bestimmten Effekt verschachtelt ist, der zweite Effekt wiederum im dritten und so weiter. (Diese Form der Verschachtelung kann nur durch Verwendung der Befehlssprache erreicht werden.)

Typ II. Bei dieser Methode wird die Quadratsumme eines Effekts im Modell angepaßt an alle anderen "zutreffenden" Effekte berechnet. Ein zutreffender Effekt ist ein Effekt, der mit allen Effekten in Beziehung steht, die den untersuchten Effekt nicht enthalten. Die Methode mit Quadratsummen vom Typ II wird gewöhnlich in den folgenden Fällen verwendet:

- Bei ausgeglichenen ANOVA-Modellen.
- Bei Modellen, die nur Haupteffekte von Faktoren enthalten.
- Bei Regressionsmodellen.
- Bei rein verschachtelten Designs. (Diese Form der Verschachtelung kann durch Verwendung der Befehlssprache erreicht werden.)

Typ III. Voreinstellung. Bei dieser Methode werden die Quadratsummen eines Effekts im Design als Quadratsummen orthogonal zu den Effekten (sofern vorhanden), die den Effekt enthalten, und mit Bereinigung um alle anderen Effekte, die diesen Effekt nicht enthalten, berechnet. Der große Vorteil der Quadratsummen vom Typ III ist, dass sie invariant bezüglich der Zellenhäufigkeiten sind, solange die allgemeine Form der Schätzbarkeit konstant bleibt. Daher wird dieser Typ von Quadratsumme oft für nicht ausgeglichene Modelle ohne fehlende Zellen als geeignet angesehen. In einem faktoriellen Design ohne fehlende Zellen ist diese Methode äquivalent zu der Methode der gewichteten Mittelwertquadrate nach Yates. Die Methode mit Quadratsummen vom Typ III wird gewöhnlich in folgenden Fällen verwendet:

- Alle bei Typ I und Typ II aufgeführten Modelle.
- Alle ausgeglichenen oder unausgeglichenen Modelle ohne leere Zellen.

Typ IV. Diese Methode ist dann geeignet, wenn es keine fehlenden Zellen gibt. Für alle Effekte *F* im Design: Wenn *F* in keinem anderen Effekt enthalten ist, dann gilt: Typ IV = Typ III = Typ II. Wenn *F* in anderen Effekten enthalten ist, werden bei Typ IV die Kontraste zwischen den Parametern in *F* gleichmäßig auf alle Effekte höherer Ordnung verteilt. Die Methode mit Quadratsummen vom Typ IV wird gewöhnlich in folgenden Fällen verwendet:

- Alle bei Typ I und Typ II aufgeführten Modelle.
- Alle ausgeglichenen oder unausgeglichenen Modelle mit leeren Zellen.

GLM: Kontraste

Abbildung 11-3
Dialogfeld "Univariat: Kontraste"

Kontraste werden verwendet, um auf Unterschiede zwischen den Stufen eines Faktors zu testen. Für jeden Faktor im Modell kann ein Kontrast festgelegt werden (in einem Modell mit Messwiederholungen für jeden Zwischensubjektfaktor). Kontraste stellen lineare Kombinationen der Parameter dar.

Das Testen der Hypothesen basiert auf der Nullhypothese **LB** = 0. Dabei ist **L** die Kontrastkoeffizienten-Matrix und **B** der Parametervektor. Wenn ein Kontrast angegeben wird, wird eine **L**-Matrix erstellt. Die Spalten der **L**-Matrix, die dem Faktor entsprechen, stimmen mit dem Kontrast überein. Die verbleibenden Spalten werden so angepaßt, dass die **L**-Matrix schätzbar ist.

Die Ausgabe beinhaltet eine *F*-Statistik für jedes Set von Kontrasten. Für die Kontrastdifferenzen werden außerdem simultane Konfidenzintervalle nach Bonferroni auf der Grundlage der Student-*T*-Verteilung angezeigt.

Verfügbare Kontraste

Als Kontraste sind “Abweichung”, “Einfach”, “Differenz”, “Helmert”, “Wiederholt” und “Polynomial” verfügbar. Bei Abweichungskontrasten und einfachen Kontrasten können Sie wählen, ob die letzte oder die erste Kategorie als Referenzkategorie dient.

Kontrasttypen

Abweichung. Vergleicht den Mittelwert jeder Faktorstufe (außer bei Referenzkategorien) mit dem Mittelwert aller Faktorstufen (Gesamtmittelwert). Die Stufen des Faktors können in beliebiger Ordnung vorliegen.

Einfach. Vergleicht den Mittelwert jeder Faktorstufe mit dem Mittelwert einer angegebenen Faktorstufe. Dieser Kontrasttyp ist nützlich, wenn es eine Kontrollgruppe gibt. Sie können die erste oder die letzte Kategorie als Referenz auswählen.

Differenz. Vergleicht den Mittelwert jeder Faktorstufe (außer der ersten) mit dem Mittelwert der vorhergehenden Faktorstufen. (Dies wird gelegentlich auch als umgekehrter Helmert-Kontrast bezeichnet).

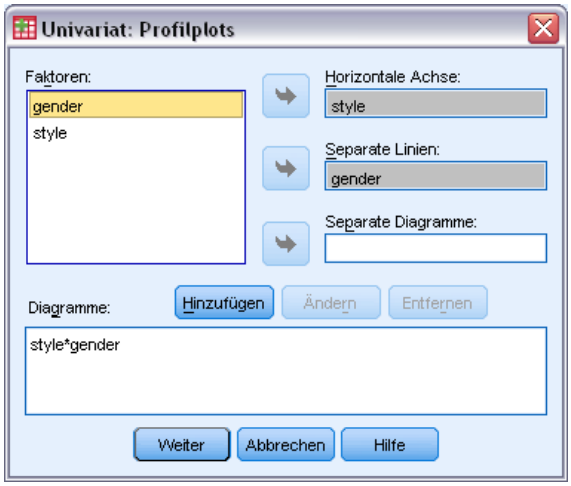
Helmert. Vergleicht den Mittelwert jeder Stufe des Faktors (bis auf die letzte) mit dem Mittelwert der folgenden Stufen.

Wiederholt. Vergleicht den Mittelwert jeder Faktorstufe (außer der letzten) mit dem Mittelwert der folgenden Faktorstufe.

Polynomial. Vergleicht den linearen Effekt, quadratischen Effekt, kubischen Effekt und so weiter. Der erste Freiheitsgrad enthält den linearen Effekt über alle Kategorien; der zweite Freiheitsgrad den quadratischen Effekt und so weiter. Die Kontraste werden oft verwendet, um polynomiale Trends zu schätzen.

GLM: Profilplots

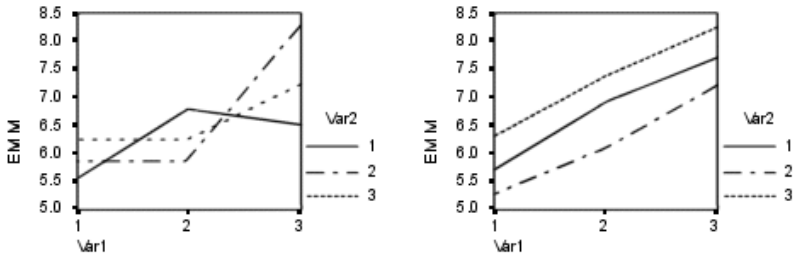
Abbildung 11-4
Dialogfeld "Univariat: Profilplots"



Profilplots (Wechselwirkungsdiagramme) sind hilfreich zum Vergleichen von Randmitteln im Modell. Ein Profilplot ist ein Liniendiagramm, in dem jeder Punkt das geschätzte Randmittel einer abhängigen Variablen (angepaßt an die Kovariaten) bei einer Stufe eines Faktors angibt. Die Stufen eines zweiten Faktors können zum Erzeugen getrennter Linien verwendet werden. Jede Stufe in einem dritten Faktor kann verwendet werden, um ein separates Diagramm zu erzeugen. Alle festen Faktoren und Zufallsfaktoren (sofern vorhanden) sind für Diagramme verfügbar. Bei multivariaten Analysen werden Profilplots für jede abhängige Variable erstellt. Bei einer Analyse mit Messwiederholungen können in Profilplots sowohl Zwischensubjektfaktoren als auch Innersubjektfaktoren verwendet werden. "GLM - Multivariat" und "GLM - Messwiederholungen" sind nur verfügbar, wenn Sie die Option "Advanced Statistics" installiert haben.

Ein Profilplot für einen Faktor zeigt, ob die geschätzten Randmittel mit den Faktorstufen steigen oder fallen. Bei zwei oder mehr Faktoren deuten parallele Linien an, dass es keine Wechselwirkung zwischen den Faktoren gibt. Das heißt, dass Sie die Faktorstufen eines einzelnen Faktors untersuchen können. Nichtparallele Linien deuten auf eine Wechselwirkung hin.

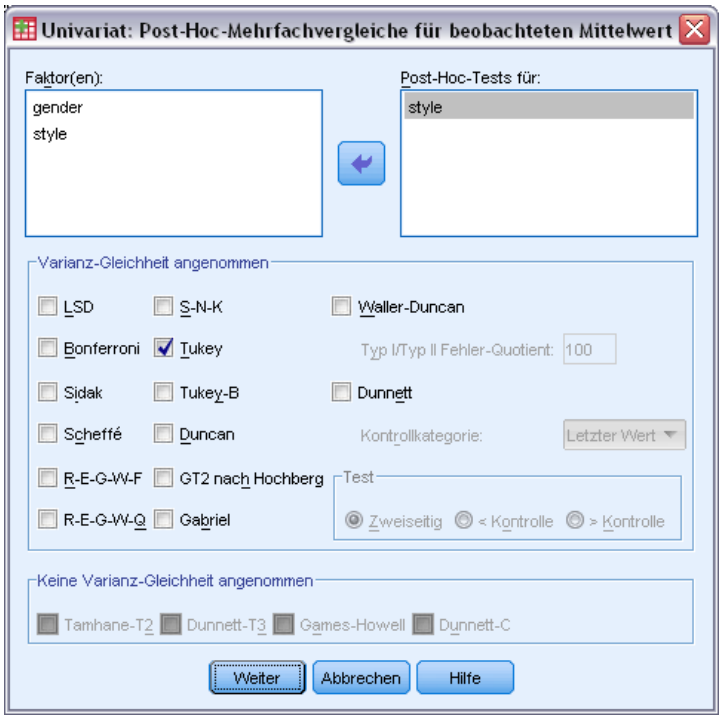
Abbildung 11-5
Nichtparalleles Diagramm (links) und paralleles Diagramm (rechts)



Nachdem ein Diagramm durch Auswahl von Faktoren für die horizontale Achse (und wahlweise von Faktoren für getrennte Linien und getrennte Diagramme) festgelegt wurde, muss das Diagramm der Liste "Diagramme" hinzugefügt werden.

GLM: Post-Hoc-Vergleiche

Abbildung 11-6
Dialogfeld "Post Hoc"



Tests für Post-Hoc-Mehrfachvergleiche. Sobald Sie festgestellt haben, dass es Abweichungen zwischen den Mittelwerten gibt, können Sie mit Post-Hoc-Spannweiten-Tests und paarweisen multiplen Vergleichen untersuchen, welche Mittelwerte sich unterscheiden. Die Vergleiche werden auf der Basis von nicht korrigierten Werten vorgenommen. Diese Tests werden nur für feste Zwischensubjektfaktoren durchgeführt. Bei "GLM - Meßwiederholungen" sind diese Tests nicht verfügbar, wenn es keine Zwischensubjektfaktoren gibt, und die Post-Hoc-Mehrfachvergleiche werden für den Durchschnitt aller Stufen der Innersubjektfaktoren durchgeführt. Bei "GLM - Multivariat" werden für jede abhängige Variable eigene Post-Hoc-Tests durchgeführt. "GLM - Multivariat" und "GLM - Messwiederholungen" sind nur verfügbar, wenn Sie die Option "Advanced Statistics" installiert haben.

Häufig verwendete Mehrfachvergleiche sind der Bonferroni-Test und die ehrlich signifikante Differenz nach Tukey. Der **Bonferroni-Test** auf der Grundlage der studentisierten *T*-Statistik korrigiert das beobachtete Signifikanzniveau unter Berücksichtigung der Tatsache, dass multiple Vergleiche vorgenommen werden. Der **Sidak-T-Test** korrigiert ebenfalls das Signifikanzniveau und liefert engere Grenzen als der Bonferroni-Test. Die **ehrlich signifikante Differenz nach Tukey** verwendet die studentisierte Spannweitenstatistik, um alle paarweisen Vergleiche zwischen den Gruppen vorzunehmen, und setzt die experimentelle Fehlerrate auf die Fehlerrate der Ermittlung aller paarweisen Vergleiche. Beim Testen einer großen Anzahl von Mittelwertpaaren ist der Test auf ehrlich signifikante Differenz nach Tukey leistungsfähiger als der Bonferroni-Test. Bei einer kleinen Anzahl von Paaren ist der Bonferroni-Test leistungsfähiger.

GT2 nach Hochberg ähnelt dem Test auf ehrlich signifikante Differenz nach Tukey, es wird jedoch das studentisierte Maximalmodul verwendet. Meistens ist der Test nach Tukey leistungsfähiger. Der **paarweise Vergleichstest nach Gabriel** verwendet ebenfalls das studentisierte Maximalmodul und zeigt meistens eine größere Schärfe als das GT2 nach Hochberg, wenn die Zellengrößen ungleich sind. Der Test nach Gabriel kann ungenau sein, wenn die Zellengrößen große Abweichungen aufweisen.

Mit dem **paarweisen T-Test für mehrere Vergleiche nach Dunnett** wird ein Set von Verarbeitungen mit einem einzelnen Kontrollmittelwert verglichen. Als Kontrollkategorie ist die letzte Kategorie voreingestellt. Sie können aber auch die erste Kategorie einstellen. Außerdem können Sie einen einseitigen oder zweiseitigen Test wählen. Verwenden Sie einen zweiseitigen Test, um zu überprüfen, ob sich der Mittelwert bei jeder Stufe (außer der Kontrollkategorie) des Faktors von dem Mittelwert der Kontrollkategorie unterscheidet. Wählen Sie < Kontrolle, um zu überprüfen, ob der Mittelwert bei allen Stufen des Faktors kleiner als der Mittelwert der Kontrollkategorie ist. Wählen Sie > Kontrolle, um zu überprüfen, ob der Mittelwert bei allen Stufen des Faktors größer als der Mittelwert bei der Kontrollkategorie ist.

Ryan, Einot, Gabriel und Welsch (R-E-G-W) entwickelten zwei multiple Step-Down-Spannweitentests. Multiple Step-Down-Prozeduren überprüfen zuerst, ob alle Mittelwerte gleich sind. Wenn nicht alle Mittelwerte gleich sind, werden Teilmengen der Mittelwerte auf Gleichheit getestet. Das **F nach R-E-G-W** basiert auf einem *F*-Test, und **Q nach R-E-G-W** basiert auf der studentisierten Spannweite. Diese Tests sind leistungsfähiger als der multiple Spannweitentest nach Duncan und der Student-Newman-Keuls-Test (ebenfalls multiple Step-Down-Prozeduren), aber sie sind bei ungleichen Zellengrößen nicht empfehlenswert.

Bei ungleichen Varianzen verwenden Sie das **Tamhane-T2** (konservativer paarweiser Vergleichstest auf der Grundlage eines *T*-Tests), **Dunnett-T3** (paarweiser Vergleichstest auf der Grundlage des studentisierten Maximalmoduls), den **paarweisen Vergleichstest** nach **Games-Howell** (manchmal ungenau) oder das **Dunnett-C** (paarweiser Vergleichstest auf der Grundlage der studentisierten Spannweite). Beachten Sie, dass diese Tests nicht gültig sind und nicht erzeugt werden, wenn sich mehrere Faktoren im Modell befinden.

Der **multiple Spannweitentest nach Duncan**, Student-Newman-Keuls (**S-N-K**) und **Tukey-B** sind Spannweitentests, mit denen Mittelwerte von Gruppen geordnet und ein Wertebereich berechnet wird. Diese Tests werden nicht so häufig verwendet wie die vorher beschriebenen Tests.

Der **Waller-Duncan-T-Test** verwendet die Bayes-Methode. Dieser Spannweitentest verwendet den harmonischen Mittelwert der Stichprobengröße, wenn die Stichprobengrößen ungleich sind.

Das Signifikanzniveau des **Scheffé**-Tests ist so festgelegt, dass alle möglichen linearen Kombinationen von Gruppenmittelwerten getestet werden können und nicht nur paarweise Vergleiche verfügbar sind, wie bei dieser Funktion der Fall. Das führt dazu, dass der Scheffé-Test oftmals konservativer als andere Tests ist, also für eine Signifikanz eine größere Differenz der Mittelwerte erforderlich ist.

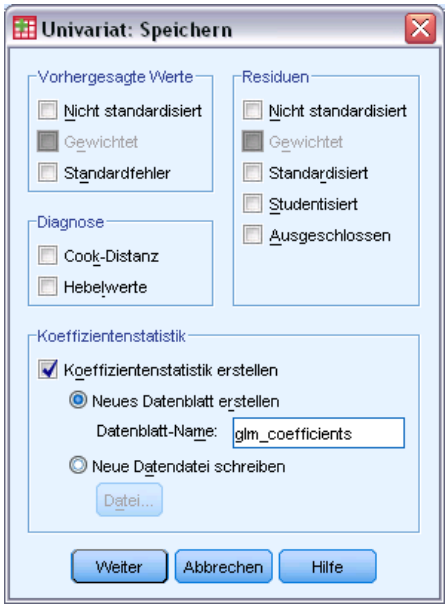
Der paarweise multiple Vergleichstest auf geringste signifikante Differenz (**LSD**) ist äquivalent zu multiplen individuellen *T*-Tests zwischen allen Gruppenpaaren. Der Nachteil bei diesem Test ist, dass kein Versuch unternommen wird, das beobachtete Signifikanzniveau im Hinblick auf multiple Vergleiche zu korrigieren.

Angezeigte Tests. Es werden paarweise Vergleiche für LSD, Sidak, Bonferroni, Games-Howell, T2 und T3 nach Tamhane, Dunnett-*C* und Dunnett-T3 ausgegeben. Homogene Untergruppen für Spannweitentests werden ausgegeben für S-N-K, Tukey-*B*, Duncan, *F* nach R-E-G-W, *Q* nach

R-E-G-W und Waller. Die ehrlich signifikante Differenz nach Tukey, das GT2 nach Hochberg, der Gabriel-Test und der Scheffé-Test sind multiple Vergleiche, zugleich aber auch Spannweitentests.

GLM: Speichern

Abbildung 11-7
Speichern"



Vom Modell vorhergesagte Werte, Residuen und verwandte Maße können als neue Variablen im Daten-Editor gespeichert werden. Viele dieser Variablen können zum Untersuchen von Annahmen über die Daten verwendet werden. Um die Werte zur Verwendung in einer anderen IBM® SPSS® Statistics-Sitzung zu speichern, müssen Sie die aktuelle Datendatei speichern.

Vorhergesagte Werte. Dies sind die Werte, welche das Modell für jeden Fall vorhersagt.

- **Nicht standardisiert (Discriminant Analysis).** Der Wert, den das Modell für die abhängige Variable vorhersagt.
- **Gewichtet.** Gewichtete nichtstandardisierte vorhergesagte Werte. Nur verfügbar, wenn zuvor eine WLS-Variable ausgewählt wurde.
- **Standardfehler.** Ein Schätzer der Standardabweichung des Durchschnittswerts der abhängigen Variablen für die Fälle, die dieselben Werte für die unabhängigen Variablen haben.

Diagnose. Dies sind Maße zum Auffinden von Fällen mit ungewöhnlichen Wertekombinationen bei der unabhängigen Variablen und von Fällen, die einen großen Einfluß auf das Modell haben könnten.

- **Cook-Distanz.** Ein Maß dafür, wie stark sich die Residuen aller Fälle ändern würden, wenn ein spezieller Fall von der Berechnung der Regressionskoeffizienten ausgeschlossen würde. Ein großer Wert der Cook-Distanz zeigt an, dass der Ausschluss eines Falles von der Berechnung der Regressionskoeffizienten die Koeffizienten substanziell verändert.
- **Hebelwerte.** Nicht zentrierte Hebelwerte. Der relative Einfluß einer jeden Beobachtung auf die Anpassungsgüte eines Modells.

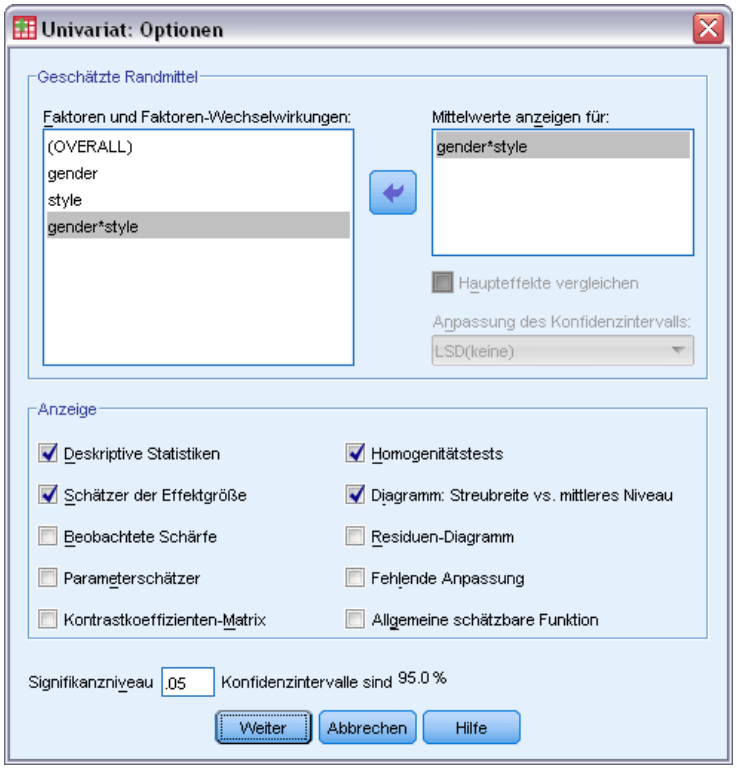
Residuen. Ein nicht standardisiertes Residuum ist der tatsächliche Wert der abhängigen Variablen minus des vom Modell geschätzten Werts. Ebenfalls verfügbar sind standardisierte, studentisierte und ausgeschlossene Residuen. Falls Sie eine WLS-Variable ausgewählt haben, sind auch gewichtete nicht standardisierte Residuen verfügbar.

- **Nicht standardisiert (Discriminant Analysis).** Die Differenz zwischen einem beobachteten Wert und dem durch das Modell vorhergesagten Wert.
- **Gewichtet.** Gewichtete nichtstandardisierte Residuen. Nur verfügbar, wenn zuvor eine WLS-Variable ausgewählt wurde.
- **Standardisiert (Residuals).** Der Quotient aus dem Residuum und einem Schätzer seiner Standardabweichung. Standardisierte Residuen, auch bekannt als Pearson-Residuen, haben einen Mittelwert von 0 und eine Standardabweichung von 1.
- **Studentisiert.** Ein Residuum, das durch seine geschätzte Standardabweichung geteilt wird, die je nach der Distanz zwischen den Werten der unabhängigen Variablen des Falles und dem Mittelwert der unabhängigen Variablen von Fall zu Fall variiert.
- **Löschen (Select Cases).** Das Residuum für einen Fall, wenn dieser Fall nicht in die Berechnung der Regressionskoeffizienten eingegangen ist. Es ist die Differenz zwischen dem Wert der abhängigen Variablen und dem korrigierten Schätzwert.

Koeffizientenstatistik. Hiermit wird eine Varianz-Kovarianz-Matrix der Parameterschätzungen für das Modell in ein neues Daten-Set in der aktuellen Sitzung oder in eine externe Datei im SPSS Statistics-Format geschrieben. Für jede abhängige Variable gibt es weiterhin eine Zeile mit Parameterschätzungen, eine Zeile mit Signifikanzwerten für die *T*-Statistik der betreffenden Parameterschätzungen und eine Zeile mit den Freiheitsgraden der Residuen. Bei multivariaten Modellen gibt es ähnliche Zeilen für jede abhängige Variable. Sie können diese Matrixdatei auch in anderen Prozeduren verwenden, die Matrixdateien einlesen.

GLM-Optionen

Abbildung 11-8
Dialogfeld "Optionen"



In diesem Dialogfeld sind weitere Statistiken verfügbar. Diese werden auf der Grundlage eines Modells mit festen Effekten berechnet.

Geschätzte Randmittel. Wählen Sie die Faktoren und Wechselwirkungen aus, für die Sie Schätzer für die Randmittel der Grundgesamtheit in den Zellen wünschen. Diese Mittel werden gegebenenfalls an die Kovariaten angepaßt.

- **Haupteffekte vergleichen.** Gibt nicht korrigierte paarweise Vergleiche zwischen den geschätzten Randmitteln für alle Haupteffekte im Modell aus, sowohl für Zwischensubjektfaktoren als auch für Innersubjektfaktoren. Diese Option ist nur verfügbar, falls in der Liste "Mittelwerte anzeigen für" Haupteffekte ausgewählt sind.
- **Anpassung des Konfidenzintervalls.** Wählen Sie für das Konfidenzintervall und die Signifikanz entweder die geringste signifikante Differenz (LSD; least significant difference), Bonferroni oder die Anpassung nach Sidak. Diese Option ist nur verfügbar, wenn Haupteffekte vergleichen ausgewählt ist.

Anzeigen. Mit der Option Deskriptive Statistik lassen Sie beobachtete Mittelwerte, Standardabweichungen und Häufigkeiten für alle abhängigen Variablen in allen Zellen berechnen. Die Option Schätzer der Effektgröße liefert einen partiellen Eta-Quadrat-Wert für jeden Effekt und jede Parameterschätzung. Die Eta-Quadrat-Statistik beschreibt den Anteil der Gesamtvariabilität, der einem Faktor zugeschrieben werden kann. Die Option Beobachtete Schärfe liefert die Testschärfe, wenn die alternative Hypothese auf die Basis der beobachteten

Werte eingestellt wurde. Mit Parameterschätzer werden Parameterschätzer, Standardfehler, *T*-Tests, Konfidenzintervalle und die beobachtete Schärfe für jeden Test berechnet. Mit der Option Matrix-Kontrastkoeffizienten wird die **L**-Matrix berechnet.

Mit der Option Homogenitätstest wird der Levene-Test auf Homogenität der Varianzen für alle abhängigen Variablen über alle Kombinationen von Faktorstufen der Zwischensubjektfaktoren durchgeführt (nur für Zwischensubjektfaktoren). Die Optionen für Diagramme der Streubreite gegen das mittlere Niveau und Residuen-Diagramme sind beim Überprüfen von Annahmen über die Daten nützlich. Diese Option ist nur verfügbar, wenn Faktoren vorhanden sind. Wählen Sie Residuen-Diagramm, wenn Sie für jede abhängige Variable ein Residuen-Diagramm (beobachtete über vorhergesagte über standardisierte Werte) erhalten möchten. Diese Diagramme sind beim Überprüfen der Annahme von Gleichheit der Varianzen nützlich. Mit der Option Fehlende Anpassung können Sie überprüfen, ob das Modell die Beziehung zwischen der abhängigen Variablen und der unabhängigen Variablen richtig beschreiben kann. Die Option Allgemeine schätzbare Funktion ermöglicht Ihnen, einen benutzerdefinierten Hypothesentest zu entwickeln, dessen Grundlage die allgemeine schätzbare Funktion ist. Zeilen in einer beliebigen Matrix der Kontrastkoeffizienten sind lineare Kombinationen der allgemeinen schätzbaren Funktion.

Signifikanzniveau. Hier können Sie das in den Post-Hoc-Tests verwendete Signifikanzniveau und das beim Berechnen von Konfidenzintervallen verwendete Konfidenzniveau ändern. Der hier festgelegte Wert wird auch zum Berechnen der beobachteten Schärfe für die Tests verwendet. Wenn Sie ein Signifikanzniveau festlegen, wird das entsprechende Konfidenzniveau im Dialogfeld angezeigt.

Zusätzliche Funktionen beim Befehl UNIANOVA

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl `DESIGN` können Sie verschachtelte Effekte im Design festlegen.
- Mit dem Unterbefehl `TEST` können Sie Tests auf Effekte im Vergleich zu linearen Kombinationen von Effekten oder einem Wert vornehmen.
- Mit dem Unterbefehl `CONTRAST` können Sie multiple Kontraste angeben.
- Mit dem Unterbefehl `MISSING` können Sie benutzerdefinierte fehlende Werte aufnehmen.
- Mit dem Unterbefehl `CRITERIA` können Sie EPS-Kriterien angeben.
- Mit den Unterbefehlen `LMATRIX`, `MMATRIX` und `KMATRIX` können Sie benutzerdefinierte **L**-Matrizen, **M**-Matrizen und **K**-Matrizen erstellen.
- Mit dem Unterbefehl `CONTRAST` können Sie bei einfachen und Abweichungskontrasten eine Referenzkategorie zwischenschalten.
- Mit dem Unterbefehl `CONTRAST` können Sie bei polynomialen Kontrasten Metriken angeben.
- Mit dem Unterbefehl `POSTHOC` können Sie Fehlerterme für Post-Hoc-Vergleiche angeben.
- Mit dem Unterbefehl `EMMEANS` können Sie geschätzte Randmittel für alle Faktoren oder Faktorenwechselwirkungen zwischen den Faktoren in der Faktorenliste berechnen lassen.
- Mit dem Unterbefehl `SAVE` können Sie Namen für temporäre Variablen angeben.
- Mit dem Unterbefehl `OUTFILE` können Sie eine Datendatei mit einer Korrelationsmatrix erstellen.

- Mit dem Unterbefehl OUTFILE können Sie eine Matrix-Datendatei erstellen, die Statistiken aus der Zwischensubjekt-ANOVA-Tabelle enthält.
- Mit dem Unterbefehl OUTFILE können Sie die Design-Matrix in einer neuen Datendatei speichern.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
12

Bivariate Korrelationen

Mit der Prozedur "Bivariate Korrelationen" werden der Korrelationskoeffizient nach Pearson, Spearman-Rho und Kendall-Tau-*b* mit ihren jeweiligen Signifikanzniveaus errechnet. Mit Korrelationen werden die Beziehungen zwischen Variablen oder deren Rängen gemessen. Untersuchen Sie Ihre Daten vor dem Berechnen eines Korrelationskoeffizienten auf Ausreißer, da diese zu irreführenden Ergebnissen führen können. Stellen Sie fest, ob wirklich ein linearer Zusammenhang existiert. Der Korrelationskoeffizient nach Pearson ist ein Maß für den linearen Zusammenhang. Wenn zwei Variablen miteinander in starker Beziehung stehen, der Zusammenhang aber nicht linear ist, ist der Korrelationskoeffizient nach Pearson keine geeignete Statistik zum Messen des Zusammenhangs.

Beispiel. Besteht eine Korrelation zwischen der Anzahl der von einer Basketballmannschaft gewonnenen Spiele und der durchschnittlich pro Spiel erzielten Anzahl von Punkten? Ein Streudiagramm zeigt, dass ein linearer Zusammenhang besteht. Eine Analyse der Daten der NBA-Saison 1994–1995 ergibt, dass der Korrelationskoeffizient nach Pearson (0,581) auf dem Niveau 0,01 signifikant ist. Man könnte vermuten, dass die gegnerischen Mannschaften um so weniger Punkte erreicht haben, je mehr Spiele eine Mannschaft gewann. Zwischen diesen Variablen besteht eine negative Korrelation (–0,401), die auf dem Niveau 0,05 signifikant ist.

Statistiken.Für jede Variable: Anzahl der Fälle mit nichtfehlenden Werten, Mittelwert und Standardabweichung. Für jedes Variablenpaar: Korrelationskoeffizient nach Pearson, Spearman-Rho, Kendall-Tau-*b*, Kreuzprodukt der Abweichungen und Kovarianz.

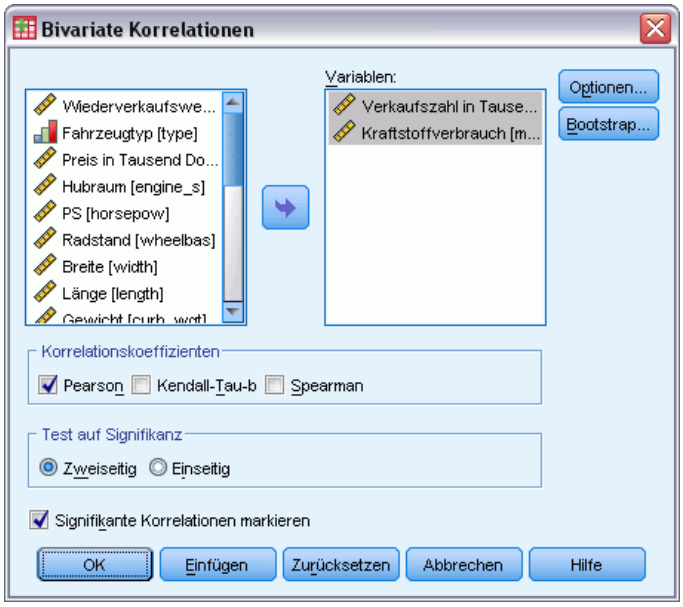
Daten. Verwenden Sie symmetrische quantitative Variablen für den Korrelationskoeffizienten nach Pearson und quantitative Variablen oder Variablen mit ordinalskalierten Kategorien für das Spearman-Rho und Kendall-Tau-*b*.

Annahmen.Für den Korrelationskoeffizient nach Pearson wird angenommen, dass jedes Variablenpaar bivariat normalverteilt ist.

So lassen Sie bivariate Korrelationen berechnen:

Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Korrelation > Bivariat...

Abbildung 12-1
Dialogfeld "Bivariate Korrelationen"



▶ Wählen Sie mindestens zwei numerische Variablen aus.

Außerdem sind folgende Optionen verfügbar:

- **Korrelationskoeffizienten.** Für quantitative, normalverteilte Variablen wählen Sie den Korrelationskoeffizienten nach Pearson. Wenn ihre Daten nicht normalverteilt sind oder mit geordneten Kategorien vorliegen, wählen Sie die Methoden Kendall-Tau-b oder Spearman, mit denen die Beziehungen zwischen Rangordnungen gemessen werden. Der Wertebereich für Korrelationskoeffizienten reicht von –1 (perfekter negativer Zusammenhang) bis +1 (perfekter positiver Zusammenhang). Der Wert 0 bedeutet, dass kein linearer Zusammenhang besteht. Vermeiden Sie bei der Interpretation Ihrer Ergebnisse, Schlüsse über Ursache und Wirkung aufgrund signifikanter Korrelationen zu ziehen.
- **Test auf Signifikanz.** Sie können einseitige oder zweiseitige Wahrscheinlichkeiten wählen. Wenn Ihnen die Richtung des Zusammenhangs im voraus bekannt ist, wählen Sie Einseitig. Wählen Sie anderenfalls Zweiseitig.
- **Signifikante Korrelationen markieren.**Korrelationskoeffizienten, die signifikant auf dem 0,05-Niveau liegen, werden mit einem einfachen Stern angezeigt. Liegen diese signifikant auf dem 0,01-Niveau, werden sie mit zwei Sternen angezeigt.

Bivariate Korrelationen: Optionen

Abbildung 12-2
Dialogfeld "Bivariate Korrelationen: Optionen"

Statistik. Für Pearson-Korrelationen können Sie eine oder auch beide der folgenden Optionen wählen:

- **Mittelwerte und Standardabweichungen.**Diese werden für jede Variable angezeigt. Außerdem wird die Anzahl der Fälle mit nichtfehlenden Werten angezeigt. Fehlende Werte werden Variable für Variable bearbeitet, unabhängig von Ihren Einstellungen für fehlende Werte.
- **Kreuzproduktabweichungen und Kovarianzen.**Werden für jedes Variablenpaar angezeigt. Das Kreuzprodukt der Abweichungen ist gleich der Summe der Produkte mittelwertkorrigierter Variablen. Dies ist der Zähler des Korrelationskoeffizienten nach Pearson. Die Kovarianz ist ein nicht standardisiertes Maß für den Zusammenhang zwischen zwei Variablen und ist gleich der Kreuzproduktabweichung dividiert durch *N*–1.

Fehlende Werte. Sie können eine der folgenden Optionen auswählen:

- **Paarweiser Fallausschluss.**Fälle mit fehlenden Werten für eine oder beide Variablen eines Paares für einen Korrelationskoeffizienten werden von der Analyse ausgeschlossen. Da jeder Koeffizient auf allen Fällen mit gültigen Codes für dieses bestimmte Variablenpaar basiert, werden in allen Berechnungen die maximal zugänglichen Informationen verwendet. Dies kann zu einer Menge von Koeffizienten führen, die auf einer variierenden Anzahl von Fällen basiert.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für Variablen werden von allen Korrelationen ausgeschlossen.

Zusätzliche Funktionen bei den Befehlen CORRELATIONS und NONPAR CORR

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl `MATRIX` kann eine Korrelationsmatrix für Pearson-Korrelationen geschrieben werden. Diese kann anstelle von Rohdaten verwendet werden, um andere Analysen zu berechnen, beispielsweise die Faktorenanalyse.
- Mit dem Schlüsselwort `WITH` im Unterbefehl `VARIABLES` können die Korrelationen zwischen allen Variablen einer Liste und allen Variablen einer zweiten Liste berechnet werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
13

Partielle Korrelationen

Partielle Korrelationskoeffizienten beschreiben die Beziehung zwischen zwei Variablen. Die Prozedur "Partielle Korrelationen" berechnet diese Koeffizienten, wobei die Effekte von einer oder mehr zusätzlichen Variablen überprüft werden. Korrelationen sind Maße für lineare Zusammenhänge. Zwei Variablen können fehlerlos miteinander verbunden sein. Wenn es sich aber nicht um eine lineare Beziehung handelt, ist der Korrelationskoeffizient zur Messung des Zusammenhangs zwischen den beiden Variablen nicht geeignet.

Beispiel. Besteht eine Beziehung zwischen den Ausgaben für das Gesundheitswesen und den Krankheitsraten? Obwohl man annehmen könnte, eine solche Beziehung sei negativ, ergibt eine Studie eine signifikante *positive* Korrelation: mit ansteigenden Ausgaben im Gesundheitswesen scheinen die Krankheitsraten zuzunehmen. Durch die Kontrolle der Effekte aus der Häufigkeit der Besuche bei medizinischem Personal wird die beobachtete positive Korrelation praktisch eliminiert. Die Ausgaben im Gesundheitswesen und die Krankheitsraten scheinen lediglich in einer positiven Beziehung zu stehen, da mit steigender Finanzausstattung mehr Menschen Zugang zu medizinischer Versorgung haben, was zu mehr gemeldeten Krankheiten bei Ärzten und Krankenhäusern führt.

Statistiken.Für jede Variable: Anzahl der Fälle mit nichtfehlenden Werten, Mittelwert und Standardabweichung. Matrizen für partielle Korrelationen und Korrelationen nullter Ordnung mit Freiheitsgraden und Signifikanzniveaus.

Daten. Verwenden Sie symmetrische, quantitative Variablen.

Annahmen. Die Prozedur "Partielle Korrelation" setzt für jedes Variablenpaar eine bivariate Normalverteilung voraus.

So lassen Sie partielle Korrelationen berechnen:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Korrelation > Partiell...

Abbildung 13-1
Dialogfeld "Partielle Korrelationen"

- Wählen Sie mindestens zwei numerische Variablen aus, für die partielle Korrelationen berechnet werden sollen.
- Wählen Sie mindestens eine numerische Kontroll-Variable aus.

Außerdem sind folgende Optionen verfügbar:

- **Test auf Signifikanz.** Sie können einseitige oder zweiseitige Wahrscheinlichkeiten wählen. Wenn Ihnen die Richtung des Zusammenhangs im voraus bekannt ist, wählen Sie Einseitig. Wählen Sie anderenfalls Zweiseitig.
- **Tatsächliches Signifikanzniveau anzeigen.** In der Standardeinstellung werden die Wahrscheinlichkeit sowie die Freiheitsgrade für jeden Korrelationskoeffizienten angezeigt. Wenn Sie diese Option deaktivieren, werden die Koeffizienten mit einem Signifikanzniveau von 0,05 mit einem Sternchen gekennzeichnet. Koeffizienten mit einem Signifikanzniveau von 0,01 werden mit einem doppelten Sternchen gekennzeichnet, und Freiheitsgrade werden unterdrückt. Diese Einstellung beeinflußt sowohl die Matrizen der partiellen Korrelationen als auch die der nullten Ordnung.

Partielle Korrelationen: Optionen

Abbildung 13-2
Dialogfeld "Partielle Korrelationen: Optionen"

Statistik. Sie können eine oder beide der folgenden Möglichkeiten auswählen:

- **Mittelwerte und Standardabweichungen.**Diese werden für jede Variable angezeigt. Außerdem wird die Anzahl der Fälle mit nichtfehlenden Werten angezeigt.
- **Korrelationen nullter Ordnung.** Hiermit wird eine einfache Matrix für Korrelationen zwischen allen Variablen (einschließlich Kontroll-Variablen) angezeigt.

Fehlende Werte. Sie können eine der folgenden Möglichkeiten wählen:

- **Listenweiser Fallausschluss.** Fälle mit fehlenden Werten für Variablen (einschließlich Kontroll-Variablen) werden aus den Berechnungen ausgeschlossen.
- **Paarweiser Fallausschluss.** Bei der Berechnung der Korrelationen nullter Ordnung, die den partiellen Korrelationen zugrunde liegen, werden Fälle mit fehlenden Werten in einer oder beiden Variablen eines Variablenpaars nicht verwendet. Beim paarweisen Löschen wird der größtmögliche Teil der Daten verwendet. Die Anzahl der Fälle kann jedoch von Koeffizient zu Koeffizient variieren. Wenn das paarweise Löschen aktiviert ist, liegt den Freiheitsgraden eines bestimmten partiellen Koeffizienten die niedrigste Anzahl von Fällen zugrunde, die zur Berechnung einer der Korrelationen nullter Ordnung verwendet werden.

Zusätzliche Funktionen beim Befehl PARTIAL CORR

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Sie können eine Korrelationsmatrix nullter Ordnung einlesen und eine Matrix der partiellen Korrelationen schreiben (mit dem Unterbefehl `MATRIX`).
- Sie können partielle Korrelationen zwischen zwei Variablenlisten erstellen (mit dem Schlüsselwort `WITH` im Unterbefehl `VARIABLES`).
- Sie können mehrere Analysen berechnen lassen (mit mehren Unterbefehlen `VARIABLES`).
- Sie können die Ordnung für die Anfrage angeben (z. B. partielle Korrelationen sowohl erster als auch zweiter Ordnung), wenn Sie über zwei Kontrollvariablen verfügen (mit dem Unterbefehl `VARIABLES`).

- Sie können redundante Koeffizienten unterdrücken (mit dem Unterbefehl `FORMAT`).
- Sie können eine Matrix von einfachen Korrelationen anzeigen lassen, wenn einige Koeffizienten nicht berechnet werden können (mit dem Unterbefehl `STATISTICS`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
14

Distanzen

Durch diese Prozedur kann eine Vielzahl von Statistiken berechnet werden, indem Ähnlichkeiten oder Unähnlichkeiten (Distanzen) zwischen Paaren von Variablen oder Fällen gemessen werden. Diese Ähnlichkeits- oder Distanzmaße können dann bei anderen Prozeduren, beispielsweise der Faktorenanalyse, der Cluster-Analyse oder der multidimensionalen Skalierung zur Analyse komplexer Daten-Sets verwendet werden.

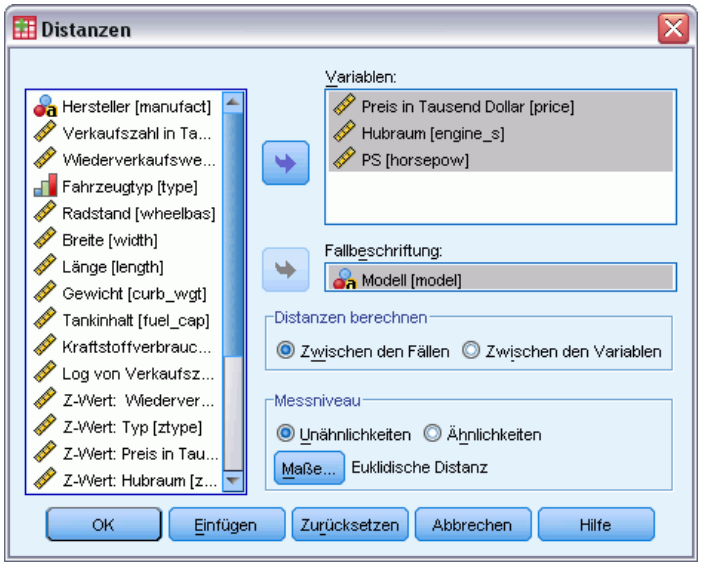
Beispiel. Ist es möglich, Ähnlichkeiten zwischen Paaren von Kraftfahrzeugen anhand bestimmter Merkmale zu messen, z. B. anhand des Hubraums, des Kraftstoffverbrauchs oder der Leistung? Durch die Berechnung von Ähnlichkeiten zwischen Kraftfahrzeugen können Sie besser einordnen, welche Fahrzeuge einander ähneln bzw. welche sich voneinander unterscheiden. Mit einer hierarchischen Cluster-Analyse oder einer multidimensionalen Skalierung auf die Ähnlichkeiten können Sie eine formale Analyse durchführen, um die zugrunde liegende Struktur zu untersuchen.

Statistiken. Unähnlichkeitsmaße (Distanzmaße) für Intervalldaten: Euklidischer Abstand, quadrierter Euklidischer Abstand, Tschebyscheff, Block, Minkowski oder ein benutzerdefiniertes Maß; für Häufigkeiten: Chi-Quadrat-Maß oder Phi-Quadrat-Maß; für Binärdaten: Euklidischer Abstand, quadrierter Euklidischer Abstand, Größendifferenz, Musterdifferenz, Varianz, Form und Distanzmaß nach Lance und Williams. Ähnlichkeitsmaße für Intervalldaten: Pearson-Korrelation oder Kosinus; für Binärdaten: Russel und Rao, einfache Übereinstimmung, Jaccard, Würfel-Ähnlichkeitsmaß, Ähnlichkeitsmaß nach Rogers und Tanimoto, Sokal und Sneath 1, Sokal und Sneath 2, Sokal und Sneath 3, Kulczynski 1, Kulczynski 2, Sokal und Sneath 4, Hamann, Lambda, Anderberg-*D*, Yule-*Y*, Yule-*Q*, Ochiai, Sokal und Sneath 5, Phi-4-Punkt-Korrelation oder Streuung.

So lassen Sie Distanzmatrizen berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Korrelation > Distanzen...

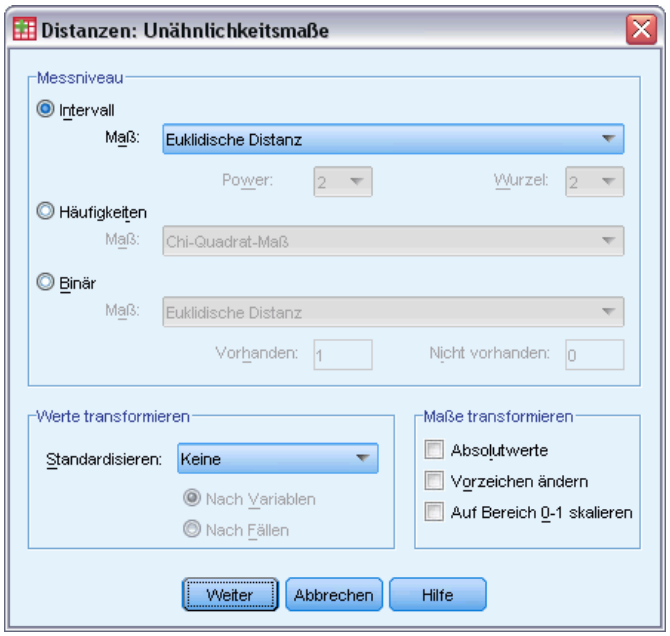
Abbildung 14-1
Dialogfeld "Distanzen"



- Wählen Sie mindestens eine numerische Variable zur Berechnung von Distanzen zwischen Fällen oder wählen Sie mindestens zwei numerische Variablen zur Berechnung von Distanzen zwischen Variablen.
- Wählen Sie im Gruppenfeld "Distanzen berechnen" eine andere Option aus, um Ähnlichkeiten zwischen Fällen oder Variablen zu berechnen.

Unähnlichkeitsmaße für Distanzen

Abbildung 14-2
Dialogfeld "Distanzen: Unähnlichkeitsmaße"



Wählen Sie aus dem Gruppenfeld "Maß" die Option aus, die Ihrem Datentyp entspricht ("Intervall", "Häufigkeiten" oder "Binär"). Wählen Sie dann aus dem Dropdown-Listenfeld ein Maß aus, das diesem Datentyp entspricht. Die folgenden Maße sind je nach Datentyp verfügbar:

- **Intervall.**Euklidischer Abstand, quadrierter Euklidischer Abstand, Tschebyscheff, Block, Minkowski oder ein benutzerdefiniertes Maß.
- **Häufigkeiten.**Chi-Quadrat-Maß oder Phi-Quadrat-Maß.
- **Binär.**Euklidischer Abstand, quadrierter Euklidischer Abstand, Größendifferenz, Musterdifferenz, Varianz, Form und Distanzmaß nach Lance und Williams. (Geben Sie Werte in die Felder "Vorhanden" und "Nicht vorhanden" ein, um anzugeben, welche beiden Werte sinnvoll sind; alle übrigen Werte werden durch die Distanzmaße ignoriert.)

Im Gruppenfeld "Werte transformieren" können Sie festlegen, ob die Datenwerte für Fälle oder Werte *vor* dem Berechnen von Ähnlichkeiten für Fälle oder Variablen standardisiert werden. Diese Transformationen sind nicht auf binäre Daten anwendbar. Die verfügbaren Standardisierungsmethoden sind "Z-Scores", "Bereich –1 bis 1", "Bereich 0 bis 1", "Maximale Größe von 1", "Mittelwert 1" und "Standardabweichung 1".

Im Gruppenfeld "Maße transformieren" können Sie festlegen, ob die durch das Distanzmaß erzeugten Werte transformiert werden. Dies erfolgt, nachdem das Distanzmaß berechnet wurde. Zu den verfügbaren Optionen zählen Absolutwerte, Ändern des Vorzeichens und Skalieren auf den Bereich 0–1.

Ähnlichkeitsmaße für Distanzen

Abbildung 14-3
Dialogfeld "Distanzen: Ähnlichkeitsmaße"

Wählen Sie aus dem Gruppenfeld "Maß" die Option aus, die Ihrem Datentyp entspricht ("Intervall" oder "Binär"). Wählen Sie dann aus dem Dropdown-Listenfeld ein Maß aus, das diesem Datentyp entspricht. Die folgenden Maße sind je nach Datentyp verfügbar:

- **Intervall.** Pearson-Korrelation oder Kosinus
- **Binär.** Russel und Rao, einfache Übereinstimmung, Jaccard, Würfel-Ähnlichkeitsmaß, Ähnlichkeitsmaß nach Rogers und Tanimoto, Ähnlichkeitsmaße nach Sokal und Sneath 1 bis 5, Kulczynski 1, Kulczynski 2, Sokal und Sneath 4, Hamann, Lambda, Anderberg-*D*, Yule-*Y*, Yule-*Q*, Ochiai, Sokal und Sneath 5, Phi-4-Punkt-Korrelation oder Streuung. (Geben Sie Werte in die Felder "Vorhanden" und "Nicht vorhanden" ein, um anzugeben, welche beiden Werte sinnvoll sind; alle übrigen Werte werden durch die Distanzmaße ignoriert.)

Im Gruppenfeld "Werte transformieren" können Sie festlegen, ob die Datenwerte für Fälle oder Variablen vor dem Berechnen von Ähnlichkeiten standardisiert werden. Diese Transformationen sind nicht auf binäre Daten anwendbar. Die verfügbaren Standardisierungsmethoden sind "Z-Scores", "Bereich –1 bis 1", "Bereich 0 bis 1", "Maximale Größe von 1", "Mittelwert 1" und "Standardabweichung 1".

Im Gruppenfeld "Maße transformieren" können Sie festlegen, ob die durch das Distanzmaß erzeugten Werte transformiert werden. Dies erfolgt, nachdem das Distanzmaß berechnet wurde. Zu den verfügbaren Optionen zählen Absolutwerte, Ändern des Vorzeichens und Skalieren auf den Bereich 0–1.

Zusätzliche Funktionen beim Befehl PROXIMITIES

In der Prozedur “Distanzen” wird die Befehlssyntax von `PROXIMITIES` verwendet. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Angeben einer Ganzzahl als Exponent für das Minkowski-Distanzmaß
- Angeben von beliebigen Ganzzahlen als Exponent und Wurzel für ein benutzerdefiniertes Distanzmaß

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
15

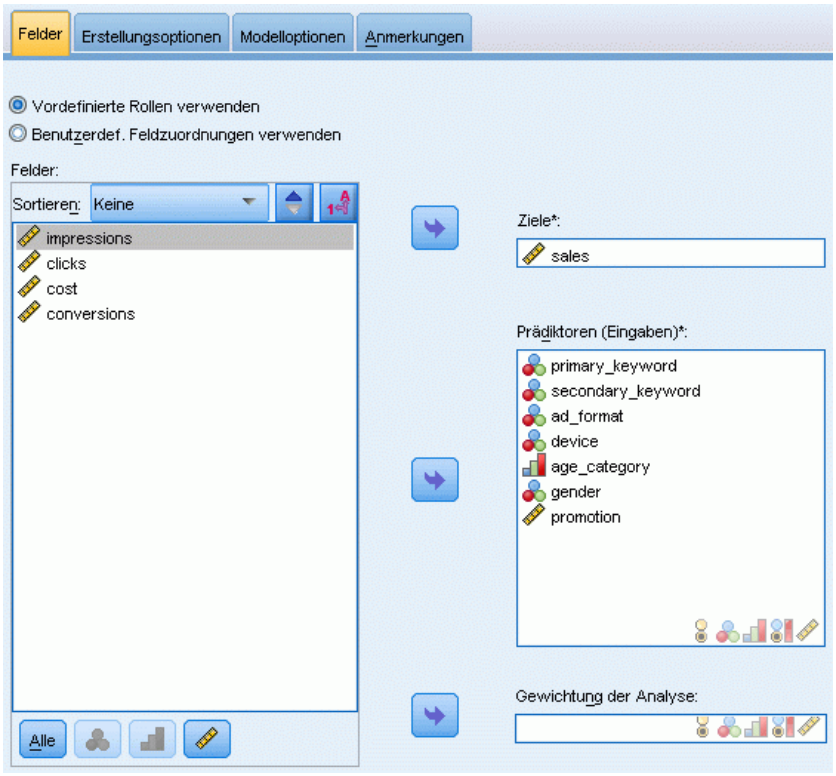
Lineare Modelle

Bei linearen Modellen wird ein stetiges Ziel auf der Basis linearer Beziehungen zwischen dem Ziel und einem oder mehreren Prädiktoren vorhergesagt.

Lineare Modelle sind relativ einfach und bieten eine leicht zu interpretierende mathematische Formel für das Scoring. Die Eigenschaften dieser Modelle sind umfassend bekannt und sie lassen sich üblicherweise sehr schnell im Vergleich zu anderen Modelltypen (beispielsweise neuronale Netze oder Entscheidungsbäume) im selben Datenblatt (Daten-Set) erstellen.

Beispiel. Eine Versicherungsgesellschaft mit beschränkten Ressourcen für die Untersuchung der Versicherungsansprüche von Hauseigentümern möchte ein Modell zur Schätzung der Kosten durch Schadensfälle erstellen. Durch die Bereitstellung dieses Modells in einem Service-Center können Versicherungsvertreter Informationen zu Schadensfällen eingeben, während sie mit einem Kunden telefonieren, und sofort die "erwarteten" Kosten des Schadenfalls auf der Grundlage früherer Daten abrufen.

Abbildung 15-1
Registerkarte "Felder"



Feldanforderungen. Es müssen ein Ziel und mindestens eine Eingabe vorhanden sein. Standardmäßig werden Felder mit den vordefinierten Rollen “Beide” oder “Keines” nicht verwendet. Das Ziel muss stetig (metrisch) sein. Es gibt keine Messniveaubeschränkungen bei Prädiktoren (Eingaben).

So erstellen Sie ein lineares Modell:

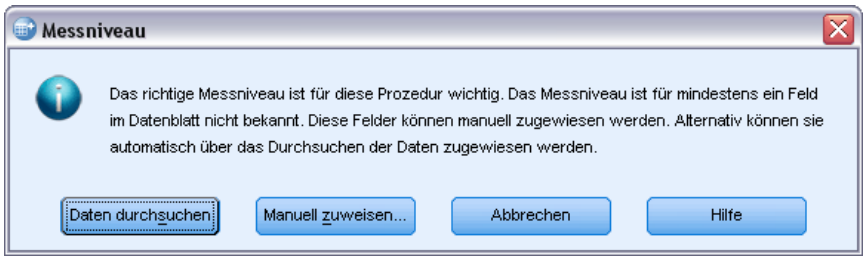
Für diese Funktion ist die Option “Statistics Base” erforderlich.

Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Regression > Automatische lineare Modelle...

- Stellen Sie sicher, dass mindestens ein Ziel und eine Eingabe vorhanden sind.
- Klicken Sie auf Erstellungsoptionen, um optionale Erstellungs- und Modelleinstellungen anzugeben.
- Klicken Sie auf Modelloptionen, um Scores im aktiven Daten-Set zu speichern und das Modell an eine externe Datei zu exportieren.
- Klicken Sie auf Ausführen, um die Prozedur auszuführen und die Modellobjekte zu erstellen.

Die Messniveau-Warnmeldung wird angezeigt, wenn das Messniveau für mindestens eine Variable (ein Feld) im Datenblatt unbekannt ist. Da sich das Messniveau auf die Berechnung der Ergebnisse für diese Prozedur auswirkt, müssen alle Variablen ein definiertes Messniveau aufweisen.

Abbildung 15-2
Messniveau-Warnmeldung



- **Daten durchsuchen.** Liest die Daten im aktiven Datenblatt (Arbeitsdatei) und weist allen Feldern, deren Messniveau zurzeit nicht bekannt ist, das Standardmessniveau zu. Bei großen Datenblättern kann dieser Vorgang einige Zeit in Anspruch nehmen.
- **Manuell zuweisen.** Öffnet ein Dialogfeld, in dem alle Felder mit unbekanntem Messniveau aufgeführt werden. Mit diesem Dialogfeld können Sie diesen Feldern ein Messniveau zuweisen. Außerdem können Sie in der Variablenansicht des Daten-Editors ein Messniveau zuweisen.

Da das Messniveau für diese Prozedur bedeutsam ist, können Sie erst dann auf das Dialogfeld zur Ausführung dieser Prozedur zugreifen, wenn für alle Felder ein Messniveau definiert wurde.

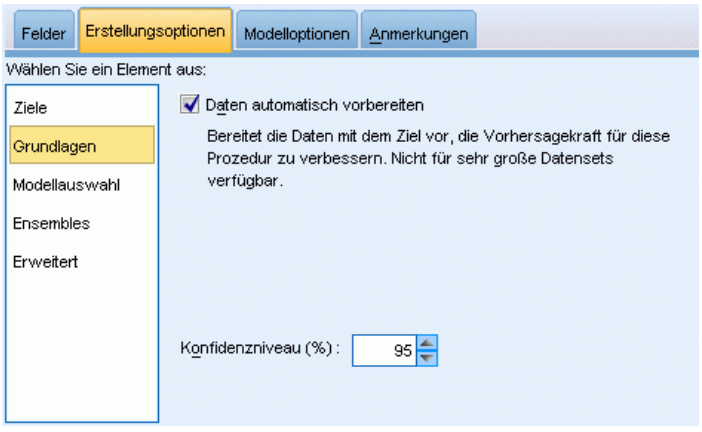
Ziele

Wie lautet Ihr Hauptziel?

- **Standardmodell erstellen.** Bei dieser Methode wird ein einzelnes Modell erstellt, um das Ziel mithilfe der Einflussvariablen vorherzusagen. Allgemein gesagt, sind Standardmodelle einfacher zu interpretieren und schneller zu bewerten als verstärkte Datenblatt-Ensembles, Datenblatt-Ensembles mit Bootstrap-Aggregation oder große Datenblatt-Ensembles.
- **Modellgenauigkeit verbessern (Verstärkung).** Bei dieser Methode wird ein Ensemble mithilfe von Verstärkung (Boosting) erstellt, wobei eine Sequenz von Modellen generiert wird, um genauere Vorhersagen zu erzielen. Bei Ensembles können Erstellung und Bewertung länger dauern als bei Standardmodellen.
- **Modellstabilität verbessern (Bootstrap-Aggregation).** Bei dieser Methode wird ein Ensemble-Modell mithilfe von Bootstrap-Aggregation (Bagging) erstellt, wobei mehrere Modelle generiert werden, um genauere Vorhersagen zu erzielen. Bei Ensembles können Erstellung und Bewertung länger dauern als bei Standardmodellen.
- **Modell für extrem große Datenblätter erstellen (IBM® SPSS® Statistics Server erforderlich).** Bei dieser Methode wird ein Ensemble-Modell durch Aufteilung des Datenblatts in separate Datenblöcke erstellt. Verwenden Sie diese Option, wenn Ihr Datenblatt zu groß ist, um eines der oben genannten Modelle zu erstellen, oder um inkrementelle Modellerstellung durchzuführen. Bei dieser Option kann die Erstellung weniger zeitaufwendig sein, die Bewertung kann jedoch länger dauern als bei Standardmodellen. Für diese Option ist eine Verbindung zum SPSS Statistics Server erforderlich.

Grundeinstellungen

Abbildung 15-3
Grundeinstellungen



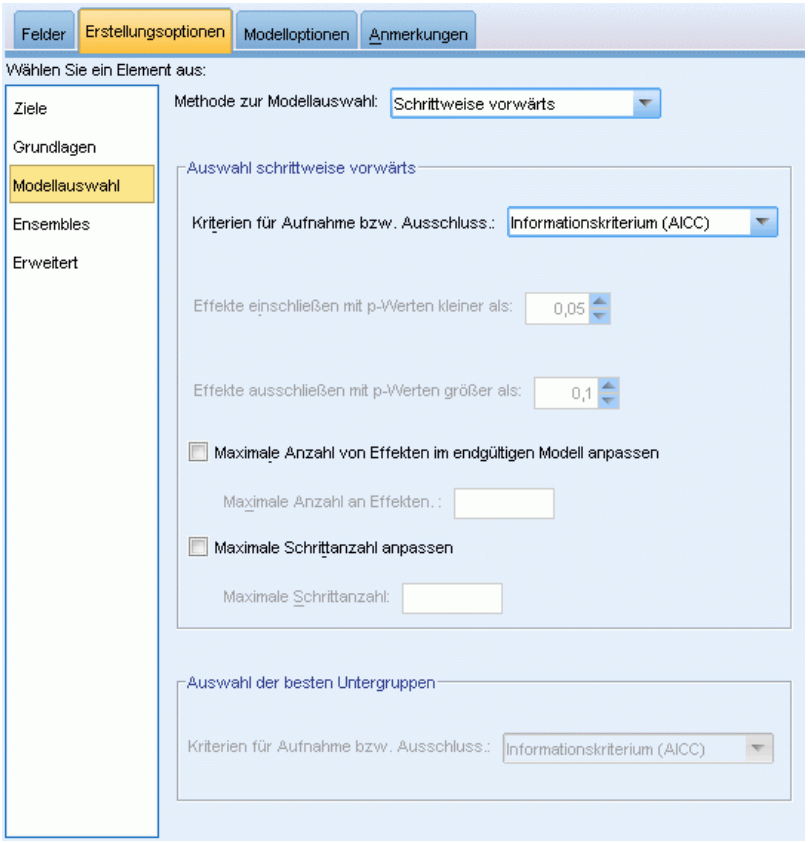
Automatische Datenaufbereitung. Mit dieser Option kann die Prozedur das Ziel und die Prädiktoren intern transformieren, um die Vorhersagekraft des Modells zu maximieren. Etwaige Transformationen werden zusammen mit dem Modell gespeichert und zum Scoren auf neue Daten angewendet. Die Originalversionen der transformierten Felder werden vom Modell ausgeschlossen. Standardmäßig wird folgende automatische Datenaufbereitung durchgeführt.

- **Verarbeitung von Datum und Zeit.** Jeder Datumsprädiktor wird in einen neuen stetigen Prädiktor transformiert, der die Zeit enthält, die seit einem Referenzdatum (1970-01-01) vergangen ist. Jeder Zeitprädiktor wird in einen neuen stetigen Prädiktor transformeirt, der die Zeit enthält, die seit einer Referenzzeit (00:00:00) vergangen ist.
- **Messniveau anpassen.** Stetige Prädiktoren mit weniger als fünf distinkten Werten werden in ordinale Felder umgewandelt. Ordinale Prädiktoren mit mehr als zehn distinkten Werten werden in stetige Prädiktoren umgewandelt.
- **Ausreißer-Behandlung.** Werte stetiger Prädiktoren, die über einem Cutoff-Wert liegen (drei Standardabweichungen vom Mittelwert), werden auf den Cutoff-Wert gesetzt.
- **Behandlung fehlender Werte.** Fehlende Werte nominaler Prädiktoren werden durch den Modus der Trainingspartition ersetzt. Fehlende Werte ordinaler Prädiktoren werden durch den Median der Trainingspartition ersetzt. Fehlende Werte stetiger Prädiktoren werden durch den Mittelwert der Trainingspartition ersetzt.
- **Überwachte Zusammenführung.** Mit dieser Option erstellen Sie ein sparsameres Modell, indem die Anzahl der zu verarbeitenden Felder in Zusammenhang mit dem Ziel reduziert wird. Ähnliche Kategorien werden anhand der Beziehung zwischen der Eingabe und dem Ziel identifiziert. Kategorien, die sich nicht signifikant unterscheiden (d. h. einen p-Wert aufweisen, der größer als 0,1 ist), werden zusammengeführt. Hinweis: Wenn alle Kategorien zu einer verschmolzen werden, werden die originalen und abgeleiteten Versionen des Felds aus dem Modell ausgeschlossen, da sie als Einflussgrößen keinen Wert haben.

Konfidenzniveau. Dieses Konfidenzniveau wird zur Berechnung der Intervallschätzungen der Modellkoeffizienten in der Ansicht Koeffizienten verwendet. Geben Sie einen Wert größer 0 und kleiner 100 ein. Der Standardwert ist 95.

Modellauswahl

Abbildung 15-4
Modellauswahl - Einstellungen



Modellauswahlmethode. Wählen Sie eine der Modellauswahlmethoden (Details unten) oder Keine aus, wodurch einfach alle verfügbaren Prädiktoren als Haupteffekt-Modellterme eingegeben werden. Standardmäßig wird Schrittweise vorwärts verwendet.

Auswahl "Schrittweise vorwärts". Diese Option beginnt ohne Effekte im Modell und nimmt jeweils einen Effekt auf bzw. schließt ihn aus, bis entsprechend den Kriterien bei "Schrittweise vorwärts" keine weiteren Vorgänge möglich sind.

- **Kriterien für Aufnahme/Ausschluss.** Diese Statistik wird zur Bestimmung verwendet, ob ein Effekt im Modell aufgenommen oder aus diesem ausgeschlossen werden soll. Das Informationskriterium (AICC) basiert auf der Wahrscheinlichkeit des Trainings-Sets für das Modell und wird zur Penalisierung übermäßig komplexer Modelle angepasst. Die F-Statistik beruht auf einem statistischen Test der Verbesserung des Modellfehlers. Korrigiertes R-Quadrat beruht auf der Anpassungsgüte des Trainings-Sets und wird zur Penalisierung übermäßig komplexer Modelle angepasst. Das Kriterium zur Verhinderung übermäßiger Anpassung (ASE) basiert auf der Anpassungsgüte (durchschnittliches Fehlerquadrat, Average Squared Error, ASE) des Sets zur Verhinderung übermäßiger Anpassung. Das Set zur Verhinderung von Überanpassung ist eine zufällige Teilstichprobe von ca. 30 % des ursprünglichen Daten-Sets, die nicht zum Trainieren des Modells verwendet wird.

Wenn ein anderes Kriterium als F-Statistik gewählt wird, wird bei jedem Schritt der Effekt im Modell aufgenommen, der dem größten positiven Zuwachs des Kriteriums entspricht. Alle Effekte, die einer Abnahme des Kriteriums entsprechen, werden aus dem Modell ausgeschlossen.

Wenn F-Statistik als Kriterium gewählt wird, wird bei jedem Schritt der Effekt mit dem geringsten *p*-Wert kleiner als der festgelegte Schwellenwert, Einschließen von Effekten mit p-Werten kleiner als , in das Modell aufgenommen. Der Standardwert lautet 0,05. Alle Effekte im Modell mit einem *p*-Wert größer als der festgelegte Schwellenwert, Entfernen von Effekten mit p-Werten größer als werden ausgeschlossen. Der Standardwert lautet 0,10.

- **Anpassen der maximalen Anzahl an Effekten im endgültigen Modell.** Standardmäßig können alle verfügbaren Effekte in das Modell eingegeben werden. Wenn alternativ der schrittweise Algorithmus einen Schritt bei der festgelegten maximalen Anzahl an Effekten beendet, stoppt der Algorithmus beim aktuellen Effekt-Set.
- **Anpassen der maximalen Anzahl an Schritten.** Der schrittweise Algorithmus stoppt nach einer bestimmten Anzahl von Schritten. Standardmäßig ist das dreimal die Anzahl an verfügbaren Effekten. Alternativ kann eine positive Ganzzahl als maximale Anzahl an Schritten angegeben werden.

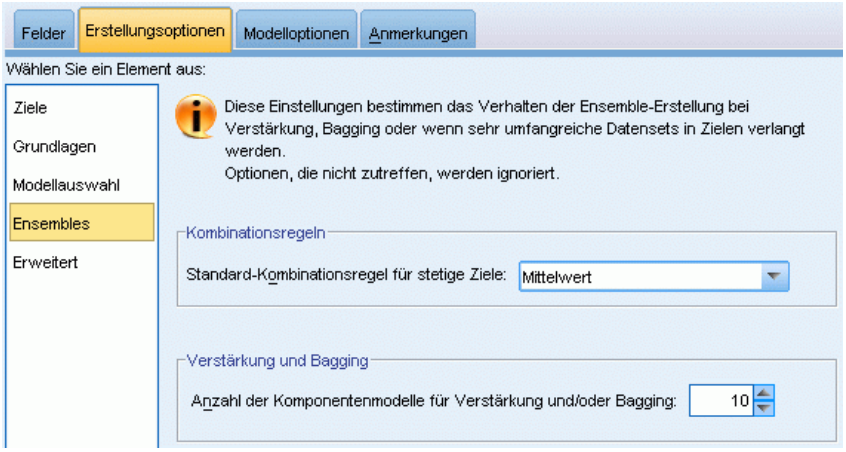
Auswahl "Beste Untergruppen". Diese Option überprüft "alle möglichen" Modelle oder zumindest eine größere Untergruppe der möglichen Modelle als "Schrittweise vorwärts", um die beste Möglichkeit entsprechend dem Kriterium "Beste Untergruppen" auszuwählen. Das Informationskriterium (AICC) basiert auf der Wahrscheinlichkeit des Trainings-Sets für das Modell und wird zur Penalisierung übermäßig komplexer Modelle angepasst. Korrigiertes R-Quadrat beruht auf der Anpassungsgüte des Trainings-Sets und wird zur Penalisierung übermäßig komplexer Modelle angepasst. Das Kriterium zur Verhinderung übermäßiger Anpassung (ASE) basiert auf der Anpassungsgüte (durchschnittliches Fehlerquadrat, Average Squared Error, ASE) des Sets zur Verhinderung übermäßiger Anpassung. Das Set zur Verhinderung von Überanpassung ist eine zufällige Teilstichprobe von ca. 30 % des ursprünglichen Daten-Sets, die nicht zum Trainieren des Modells verwendet wird.

Das Modell mit dem höchsten Wert für das Kriterium wird als das beste Modell ausgewählt.

Anmerkung: Die Auswahl "Beste Untergruppen" ist rechenintensiver als die Auswahl "Schrittweise vorwärts". Wenn "Beste Untergruppen" zusammen mit "Verbesserung", "Verstärkung" oder "Sehr große Daten-Sets" verwendet wird, kann das Erstellen deutlich länger dauern als das Erstellen eines Standard-Modells mithilfe der Auswahl "Schrittweise vorwärts".

Ensembles

Abbildung 15-5
Ensemble-Einstellungen



Diese Einstellungen legen das Verhalten der Ensemblebildung fest, die erfolgt, wenn auf der Registerkarte “Ziele” die Option “Verbesserung”, “Verstärkung” oder “Sehr große Daten-Sets” ausgewählt ist. Optionen, die für das ausgewählte Ziel nicht gelten, werden ignoriert.

Bagging und sehr umfangreiche Daten-Sets. Beim Scoren eines Ensembles wird diese Regel angewendet, um die vorhergesagten Werte aus den Basismodellen für die Berechnung des Score-Werts für das Ensemble zu kombinieren.

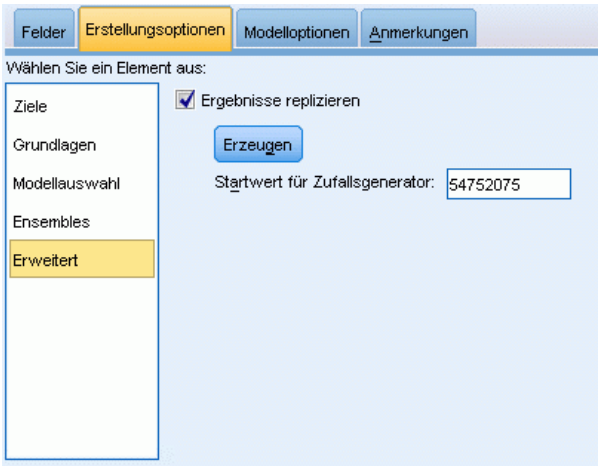
- **Standard-Kombinierungsregel für stetige Ziele.** Ensemble-Vorhersagewerte für stetige Ziele können unter Verwendung des Mittelwerts oder Medians der Vorhersagewerte aus den Basismodellen kombiniert werden.

Hinweis: Wenn als Ziel die Verbesserung der Modellgenauigkeit ausgewählt wurde, wird die Auswahl zum Kombinieren der Regeln ignoriert. Bei der Verbesserung wird für das Scoring der kategorialen Ziele stets eine gewichtete Mehrheit verwendet und für das Scoring stetiger Ziele ein gewichteter Median.

Verbesserung und Verstärkung. Geben Sie die Anzahl der zu erstellenden Basismodelle an, wenn als Ziel die Verbesserung der Modellgenauigkeit oder -stabilität angegeben ist. Im Falle der Verstärkung ist das die Anzahl der Bootstrap-Stichproben. Muss eine positive ganze Zahl sein.

Erweitert

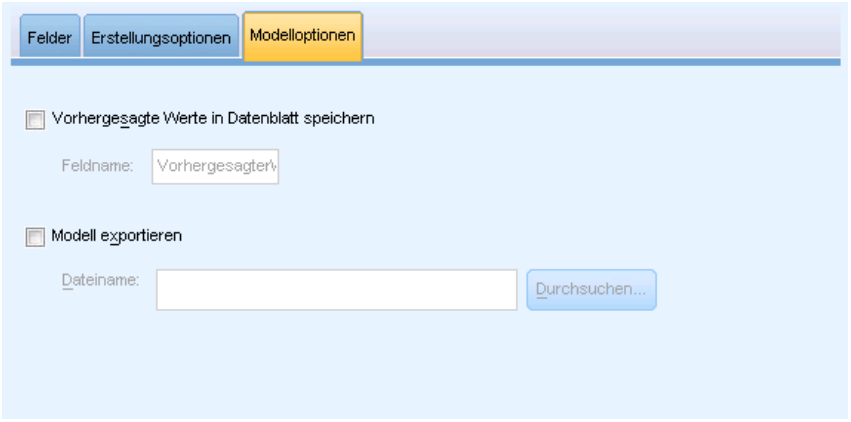
Abbildung 15-6
Erweiterte Einstellungen



Ergebnisse reproduzieren. Durch Einstellen eines Startwerts für den Zufallsgenerator können Analysen reproduziert werden. Der Zufallszahlengenerator wird verwendet, um zu wählen, welche Datensätze sich im Set zur Verhinderung übermäßiger Anpassung befinden. Geben Sie eine ganze Zahl ein oder klicken Sie auf Generieren. Dadurch wird eine pseudozufällige Ganzzahl zwischen 1 und 2147483647 (einschließlich) erzeugt. Der Standardwert lautet 54752075.

Modelloptionen

Abbildung 15-7
Registerkarte "Modelloptionen"

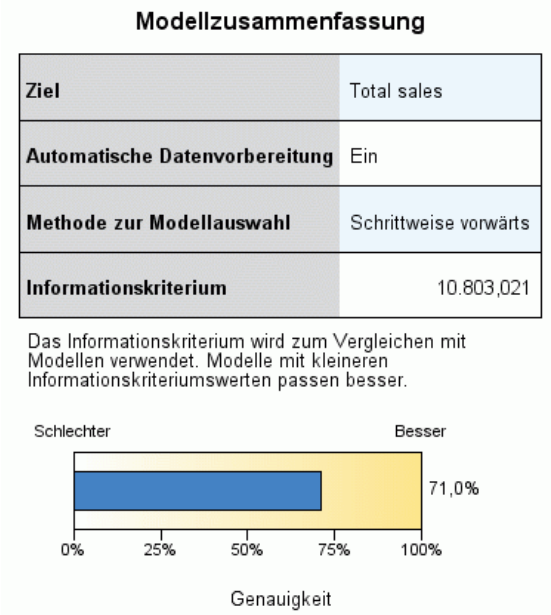


Speichert vorhergesagte Werte im Daten-Set. Der Standard-Variablenname lautet *PredictedValue*.

Modell exportieren. Schreibt das Modell in eine externe *.zip*-Datei. Anhand dieser Modelldatei können Sie die Modellinformationen zu Bewertungszwecken auf andere Datendateien anwenden. Geben Sie einen eindeutigen, gültigen Dateinamen an. Wenn die Dateispezifikation eine bestehende Datei angibt, wird diese Datei überschrieben.

Modellübersicht

Abbildung 15-8
Ansicht "Modellzusammenfassung"



Mit der Ansicht "Modellzusammenfassung" erhalten Sie eine momentane, übersichtliche Zusammenfassung des Modells und seiner Anpassungsgüte.

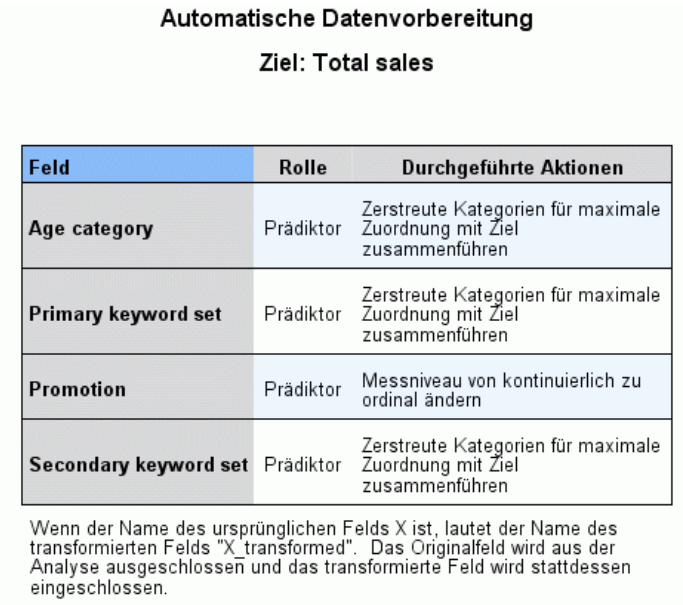
Tabelle. In der Tabelle werden einige Modelleinstellungen für ein hohes Niveau dargestellt, u. a.:

- der Name des Ziels,
- ob eine automatische Datenaufbereitung durchgeführt wurde, wie es in den Grundeinstellungen festgelegt wurde,
- die Modellauswahlmethode und das Auswahlkriterium, wie in den Einstellungen "Modellauswahl"festgelegt. Der Wert des Auswahlkriteriums für das endgültige Modell wird ebenfalls angezeigt und im Format "kleiner ist besser" dargestellt.

Diagramme. Das Diagramm zeigt die Genauigkeit des endgültigen Modells an, das in einem größeren und besseren Format dargestellt wird. Der Wert ist 100 × der eingestellten R^2 für das endgültige Modell.

Automatische Datenaufbereitung

Abbildung 15-9
Ansicht "Automatische Datenaufbereitung"



Automatische Datenvorbereitung

Ziel: Total sales

Feld	Rolle	Durchgeführte Aktionen
Age category	Prädiktor	Zerstreute Kategorien für maximale Zuordnung mit Ziel zusammenführen
Primary keyword set	Prädiktor	Zerstreute Kategorien für maximale Zuordnung mit Ziel zusammenführen
Promotion	Prädiktor	Messniveau von kontinuierlich zu ordinal ändern
Secondary keyword set	Prädiktor	Zerstreute Kategorien für maximale Zuordnung mit Ziel zusammenführen

Wenn der Name des ursprünglichen Felds X ist, lautet der Name des transformierten Felds "X_transformed". Das Originalfeld wird aus der Analyse ausgeschlossen und das transformierte Feld wird stattdessen eingeschlossen.

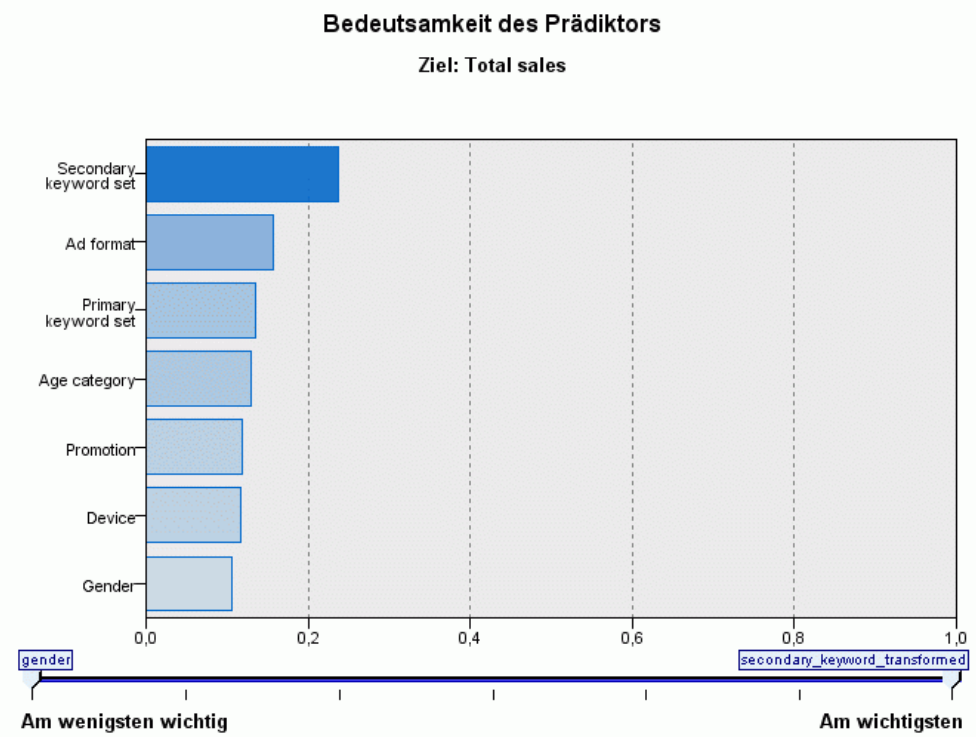
Diese Ansicht zeigt Informationen darüber an, welche Felder ausgeschlossen wurden und wie transformierte Felder im Schritt "automatische Datenaufbereitung" (ADP) abgeleitet wurden. Für jedes transformierte oder ausgeschlossene Feld listet die Tabelle den Feldnamen, die Rolle in der Analyse und die im ADP-Schritt vorgenommene Aktion auf. Die Felder werden in aufsteigender alphabetischer Reihenfolge der Feldnamen sortiert. Die möglichen für die einzelnen Felder vorgenommenen Aktionen umfassen Folgendes:

- Ableitung der Dauer: Monate berechnet die vergangene Zeit in Monaten, ausgehend von den Werten in einem Feld, das Datumsangaben bis zum aktuellen Systemdatum enthält.
- Ableitung der Dauer: Stunden berechnet die vergangene Zeit in Stunden, ausgehend von den Werten in einem Feld, das Uhrzeiten bis zur aktuellen Systemzeit enthält.
- Messniveau von stetig auf ordinal ändernwandelt stetige Felder mit weniger als fünf eindeutigen Werten in ordinale Felder um.
- Messniveau von ordinal auf stetig ändernwandelt ordinale Felder mit mehr als zehn eindeutigen Werten in stetige Felder um.
- Ausreißer kappen Werte stetiger Prädiktoren, die über einem Cutoff-Wert liegen (drei Standardabweichungen vom Mittelwert), werden auf den Cutoff-Wert gesetzt.
- Fehlende Werte ersetzen ersetzt fehlende Werte von nominalen Feldern durch den Modus, von ordinalen Feldern durch den Median und von stetigen Feldern durch den Mittelwert.

- Kategorien zusammenführen, um die Zuordnung zum Ziel zu maximieren ermittelt “ähnliche” Prädiktorkategorien auf der Grundlage der Beziehung zwischen der Eingabe und dem Ziel. Kategorien, die sich nicht signifikant unterscheiden (d. h. einen *p*-Wert aufweisen, der größer als 0,05 ist), werden zusammengeführt.
- Konstanten Prädiktor ausschließen/nach Ausreißer-Behandlung/nach der Zusammenführung von Kategorien entfernt Prädiktoren, die einen einzelnen Wert aufweisen, möglicherweise nachdem andere ADP-Aktionen ausgeführt wurden.

Bedeutsamkeit des Prädiktors

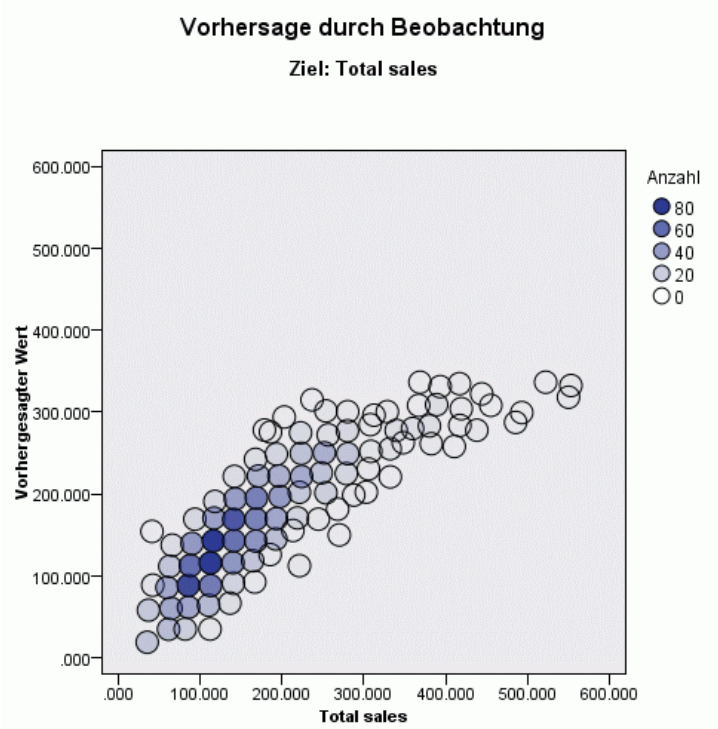
Abbildung 15-10
Ansicht “Bedeutsamkeit des Prädiktors”



In der Regel konzentriert man sich bei der Modellerstellung auf die Einflussvariablenfelder, die am wichtigsten sind, und vernachlässigt jene, die weniger wichtig sind. Dabei unterstützt Sie das Wichtigkeitsdiagramm für die Einflussvariablen, da es die relative Wichtigkeit der einzelnen Einflussvariablen für das Modell angibt. Da die Werte relativ sind, beträgt die Summe der Werte aller Einflussvariablen im Diagramm 1,0. Die Wichtigkeit der Einflussvariablen steht in keinem Bezug zur Genauigkeit des Modells. Sie bezieht sich lediglich auf die Wichtigkeit der einzelnen Einflussvariablen für eine Vorhersage und nicht auf die Genauigkeit der Vorhersage.

Vorhersage nach Beobachtung

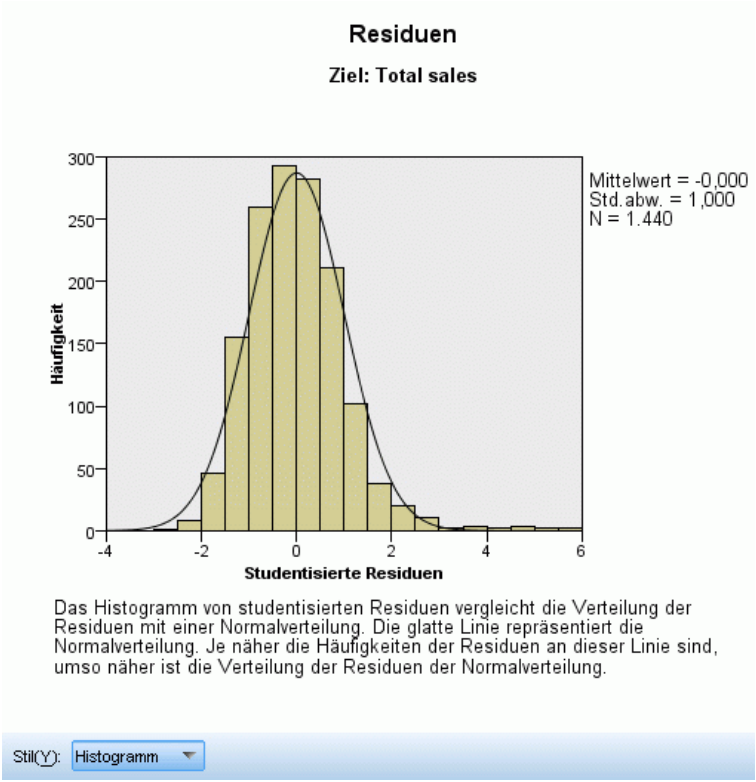
Abbildung 15-11
Ansicht "Vorhersage nach Beobachtung"



Diese Ansicht zeigt ein Bin-Streudiagramm der vorhergesagten Werte auf der vertikalen Achse durch die beobachteten Werte auf der horizontalen Achse. Idealerweise sollten die Werte entlang einer 45-Grad-Linie liegen. In dieser Ansicht können Sie erkennen, ob bestimmte Datensätze vom Modell besonders schlecht vorhergesagt werden.

Residuen

Abbildung 15-12
Ansicht "Residuen", Histogrammstil



Diese Ansicht zeigt ein Diagnosediagramm von Modellresiduen.

Diagrammstile. Für die Diagramme sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste Stil zugegriffen werden kann.

- **Histogramm.** Diese Ansicht zeigt ein Bin-Histogramm der studentisierten Residuen, das mit der normalen Verteilung überlagert ist. Lineare Modelle gehen davon aus, dass Residuen eine normale Verteilung aufweisen. Das Histogramm sollte sich also idealerweise einer nahezu glatten Linie annähern.
- **P-P-Diagramm.** Diese Ansicht zeigt ein Wahrscheinlichkeits-Wahrscheinlichkeits-Diagramm, bei dem die studentisierten Residuen mit einer normalen Verteilung verglichen werden. Wenn die Steigung der Diagrammpunkte weniger steil als die normale Linie ist, zeigen die Residuen eine größere Schwankung als eine normale Verteilung; ist die Steigung steiler, zeigen die Residuen weniger Schwankung als eine normale Verteilung. Wenn die Diagrammpunkte eine S-förmige Kurve aufweisen, ist die Verteilung der Residuen verzerrt.

Ausreißer

Abbildung 15-13
Ansicht "Ausreißer"

Ausreißer

Ziel: Total sales

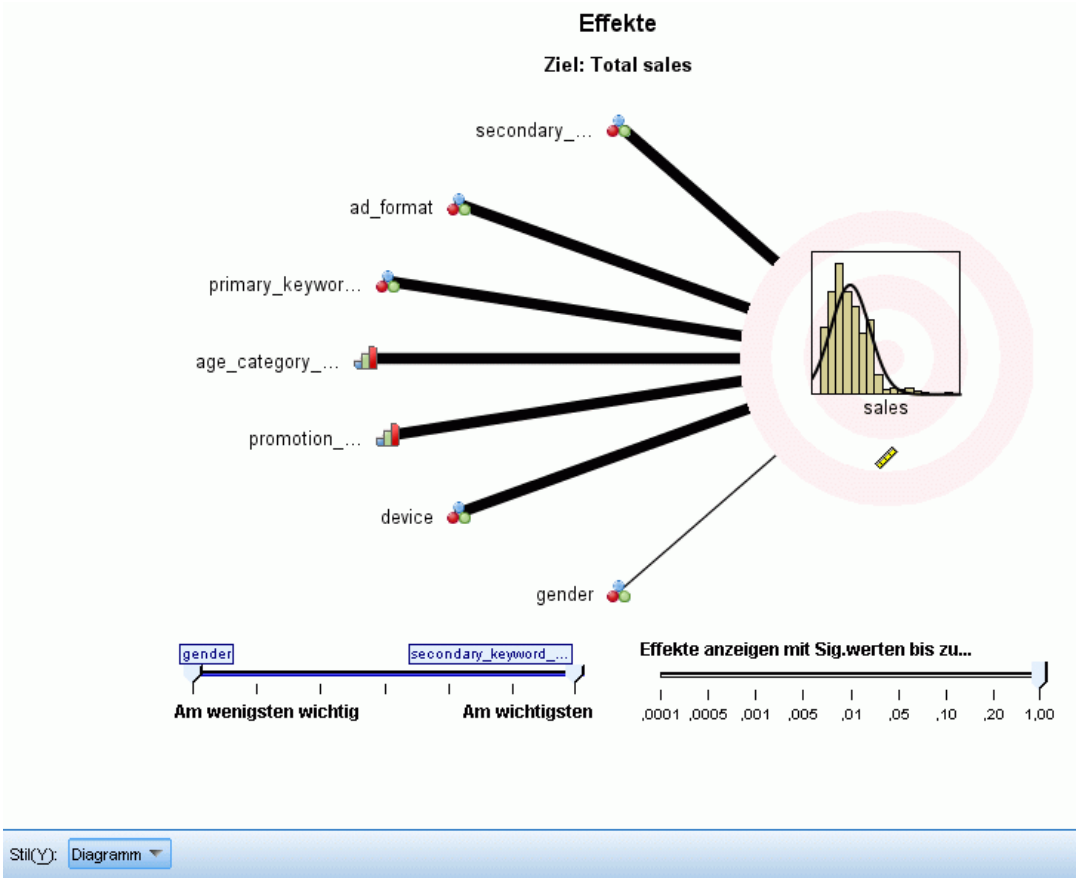
Total sales	Cook-Distanz
560.040	0,026
566.440	0,025
548.990	0,018
539.630	0,018
485.430	0,014
543.240	0,014

In dieser Tabelle sind Datensätze aufgelistet, die einen unverhältnismäßigen Einfluss auf das Modell ausüben. Außerdem werden die Datensatz-ID (sofern in der Registerkarte "Felder" angegeben), der Zielwert und die Cook-Distanz angezeigt. Die Cook-Distanz ist ein Maß dafür, wie stark sich die Residuen aller Datensätze ändern würden, wenn ein spezieller Datensatz von der Berechnung der Modellkoeffizienten ausgeschlossen würde. Ein großer Wert der Cook-Distanz zeigt an, dass der Ausschluss eines Datensatzes von der Berechnung die Koeffizienten substanziell verändert, und sollte daher als einflussreich betrachtet werden.

Einflussreiche Datensätze sollten sorgfältig untersucht werden, um zu entscheiden, ob ihnen bei der Schätzung des Modells weniger Gewicht gegeben werden kann, ob die extremen Werte auf einen akzeptablen Schwellenwert verringert werden können oder ob die einflussreichen Datensätze vollständig entfernt werden sollen.

Effekte

Abbildung 15-14
Ansicht "Effekte," Diagrammstil



Diese Ansicht zeigt die Größe der einzelnen Effekte im Modell.

Stile. Für die Diagramme sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste Stil zugegriffen werden kann.

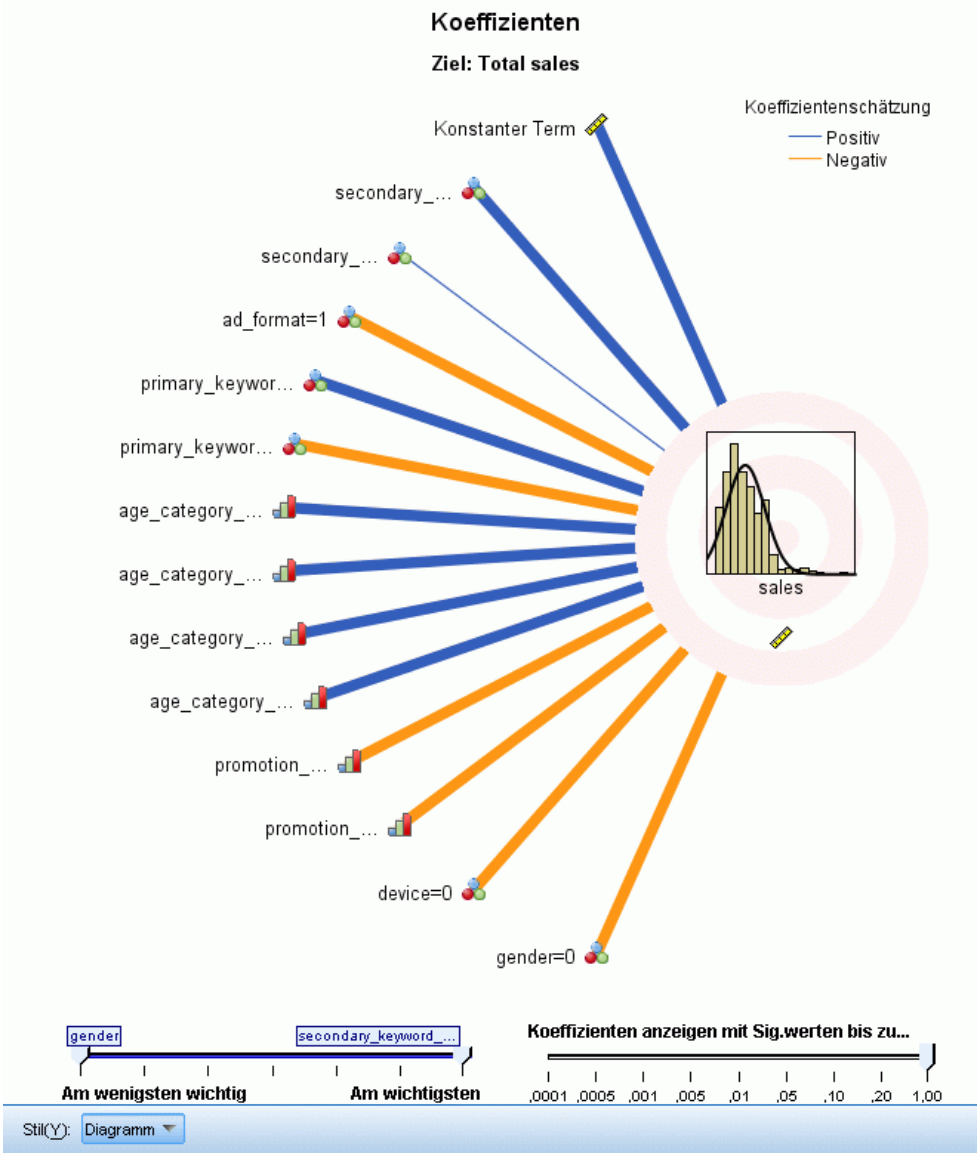
- **Diagramm.** In diesem Diagramm sind die Effekte von oben nach unten nach absteigender Bedeutsamkeit der Prädiktoren sortiert. Verbindungslinien im Diagramm sind basierend auf der Effektsignifikanz gewichtet, wobei eine größere Linienbreite signifikanteren Effekten entspricht (kleinere *p*-Werte). Wenn Sie den Mauszeiger über eine Verbindungslinie bewegen, wird eine QuickInfo mit dem *p*-Wert und der Bedeutung des Effekts angezeigt. Dies ist die Standardeinstellung.
- **Tabelle.** Diese Ansicht zeigt eine ANOVA-Tabelle für das Gesamtmodell und die einzelnen Modelleffekte. Die einzelnen Effekte sind von oben nach unten nach absteigender Bedeutsamkeit der Prädiktoren sortiert. Beachten Sie, dass die Tabelle standardmäßig minimiert ist, sodass nur die Ergebnisse des Gesamtmodells angezeigt werden. Klicken Sie in der Tabelle auf die Zelle für das korrigierte Modell, um die Ergebnisse für die einzelnen Modelleffekte anzuzeigen.

Bedeutsamkeit des Prädiktors. Für die Bedeutsamkeit des Prädiktors gibt es einen Schieberegler, mit dem eingestellt wird, welche Prädiktoren in der Ansicht gezeigt werden. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Prädiktoren konzentrieren. Standardmäßig werden die zehn besten Effekte angezeigt.

Signifikanz. Mit dem Signifikanz-Schieberegler kann noch weiter angegeben werden, welche Effekte in der Anzeige dargestellt werden. Diese Einstellungen gehen über die Eingaben, die auf der Bedeutsamkeit der Prädiktoren beruhen, hinaus. Effekte, deren Signifikanzwerte größer als der Wert des Schiebereglers sind, werden ausgeblendet. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Effekte konzentrieren. Standardmäßig ist der Wert 1,00 eingestellt, so dass keine Effekte basierend auf der Signifikanz herausgefiltert werden.

Koeffizienten

Abbildung 15-15
Ansicht "Koeffizienten", Diagrammstil



Diese Ansicht zeigt den Wert der einzelnen Koeffizienten im Modell. Hinweis: Faktoren (kategoriale Prädiktoren) sind innerhalb des Modells indikatorkodiert, sodass Faktoren, die **Effekte** enthalten, in der Regel mehrere zugehörige **Koeffizienten** aufweisen. Mit Ausnahme der Kategorie für den redundanten (Referenz-)Parameter erhält jede Kategorie einen solchen Koeffizienten.

Stile. Für die Diagramme sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste Stil zugegriffen werden kann.

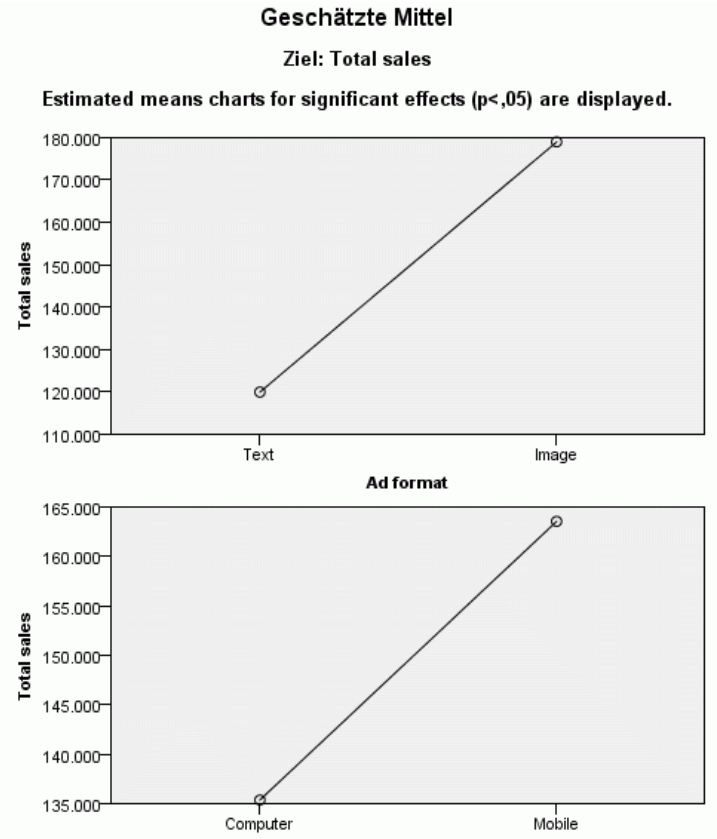
- **Diagramm.** In diesem Diagramm werden die konstanten Terme zuerst angezeigt, und dann die Effekte von oben nach unten nach absteigender Bedeutsamkeit der Prädiktoren sortiert. In Faktoren, die Effekte enthalten, werden die Koeffizienten in aufsteigender Reihenfolge der Datenwerte sortiert. Verbindungslinien im Diagramm sind basierend auf dem Vorzeichen des Koeffizienten farbig dargestellt (siehe Diagrammschlüssel) und auf der Grundlage der Koeffizientensignifikanz gewichtet, wobei eine größere Linienbreite signifikanteren Koeffizienten entspricht (kleinere *p*-Werte). Wenn Sie den Mauszeiger über eine Verbindungslinie bewegen, wird eine QuickInfo mit dem Wert des Koeffizienten, seinem *p*-Wert und der Bedeutung des Effekts angezeigt, mit dem der Parameter verbunden ist. Dies ist der Standardstil.
- **Tabelle.**Diese Tabelle zeigt die Werte, Signifikanztests und Konfidenzintervalle für die einzelnen Modellkoeffizienten. Nach dem konstanten Term sind die einzelnen Effekte von oben nach unten nach absteigender Bedeutsamkeit der Prädiktoren sortiert. In Faktoren, die Effekte enthalten, werden die Koeffizienten in aufsteigender Reihenfolge der Datenwerte sortiert. Beachten Sie, dass die Tabelle standardmäßig minimiert ist, sodass nur der Koeffizient, die Signifikanz und die Bedeutung der einzelnen Modellparameter angezeigt werden. Klicken Sie zum Anzeigen des Standardfehlers, der *t*-Statistik und des Konfidenzintervalls in der Tabelle auf die Zelle Koeffizient. Wenn Sie den Mauszeiger in der Tabelle über den Namen eines Modellparameters bewegen, wird eine QuickInfo mit dem Namen des Parameters, dem Effekt, mit dem der Parameter verbunden ist, und (für kategoriale Prädiktoren) den Wertelabels angezeigt, die mit dem Modellparameter verbunden sind. Dies kann besonders hilfreich sein, um die neuen Kategorien anzuzeigen, die erstellt werden, wenn bei der automatischen Datenaufbereitung ähnliche Kategorien eines kategorialen Prädiktors zusammengeführt werden.

Bedeutsamkeit des Prädiktors. Für die Bedeutsamkeit des Prädiktors gibt es einen Schieberegler, mit dem eingestellt wird, welche Prädiktoren in der Ansicht gezeigt werden. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Prädiktoren konzentrieren. Standardmäßig werden die zehn besten Effekte angezeigt.

Signifikanz. Mit dem Signifikanz-Schieberegler kann noch weiter angegeben werden, welche Koeffizienten in der Anzeige dargestellt werden. Diese Einstellungen gehen über die Eingaben, die auf der Bedeutsamkeit der Prädiktoren beruhen, hinaus. Koeffizienten, deren Signifikanzwerte größer als der Wert des Schiebereglers sind, werden ausgeblendet. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Koeffizienten konzentrieren. Standardmäßig ist der Wert 1,00 eingestellt, so dass keine Koeffizienten basierend auf der Signifikanz herausgefiltert werden.

Geschätzte Mittel

Abbildung 15-16
Ansicht "Geschätzte Mittel"



Diese Diagramme werden für signifikante Prädiktoren angezeigt. Das Diagramm zeigt den vom Modell geschätzten Zielwert auf der vertikalen Achse für jeden Prädiktorwert auf der horizontalen Achse, wobei alle anderen Prädiktoren konstant gehalten werden. Es gewährt eine nützliche Visualisierung der Effekte der einzelnen Prädiktorkoeffizienten auf dem Ziel.

Anmerkung: wenn keine Prädiktoren signifikant sind, werden keine geschätzten Mittel produziert.

Modellerstellungsübersicht

Abbildung 15-17
Ansicht "Modellerstellungsübersicht", Algorithmus "Schrittweise vorwärts"

Übersicht über Modellerstellung

Ziel: Total sales

		Schritt						
		1	2	3	4	5	6	7
Informationskriterium		11.949,413	11.597,758	11.347,000	11.118,878	10.965,287	10.816,338	10.803,021
Effekt	secondary_keyword_transformed	✓	✓	✓	✓	✓	✓	✓
	ad_format		✓	✓	✓	✓	✓	✓
	primary_keyword_transformed			✓	✓	✓	✓	✓
	age_category_transformed				✓	✓	✓	✓
	promotion_transformed					✓	✓	✓
	device						✓	✓
	gender							✓

Die Modellerstellungsmethode ist "Schrittweise vorwärts" mit dem "Informationskriterium".
Ein Häkchen bedeutet, dass sich der Effekt bei diesem Schritt im Modell befindet.

Wenn ein anderer Modellauswahlalgorithmus als Keiner in den Einstellungen "Modellauswahl" gewählt wird, werden einige Details zum Modellerstellungsprozess angegeben.

Schrittweise vorwärts Wenn der Auswahlalgorithmus "Schrittweise vorwärts" ist, zeigt die Tabelle die letzten zehn Schritte im schrittweisen Algorithmus an. Für jeden Schritt werden der Wert des Auswahlkriteriums und die Effekte im Modell an diesem Schritt angezeigt. Auf diese Weise bekommen Sie einen Eindruck davon, wie groß der Beitrag der einzelnen Schritte zum Modell ist. In jeder Spalte können Sie die Reihen so sortieren, dass Sie noch leichter erkennen können, welche Effekte sich bei einem bestimmten Schritt im Modell befinden.

Beste Untergruppen. Wenn der Auswahlalgorithmus "Beste Untergruppen" ist, zeigt die Tabelle die zehn besten Modelle an. Für jedes Modell werden der Wert des Auswahlkriteriums und die Effekte im Modell angezeigt. So erhalten Sie einen Eindruck der Stabilität der besten Modelle; wenn sie zu vielen ähnlichen Effekten mit wenigen Unterschieden neigen, können Sie sich auf das "Top"-Modell verlassen; wenn sie dagegen sehr unterschiedliche Effekte aufweisen, sind eventuell einige Effekte zu ähnlich und sollten kombiniert (oder entfernt) werden. In jeder Spalte können Sie die Reihen so sortieren, dass Sie noch leichter erkennen können, welche Effekte sich bei einem bestimmten Schritt im Modell befinden.

Kapitel
16

Lineare Regression

Mit “Lineare Regression” werden die Koeffizienten der linearen Gleichung unter Einbeziehung einer oder mehrerer unabhängiger Variablen geschätzt, die den Wert der abhängigen Variablen am besten vorhersagen. Sie können beispielsweise den Versuch unternehmen, die Jahresverkaufsbilanz eines Verkäufers (die abhängige Variable) nach unabhängigen Variablen wie Alter, Bildungsstand und Anzahl der Berufsjahre vorherzusagen.

Beispiel. Besteht ein Zusammenhang zwischen der Anzahl der in einer Saison gewonnenen Spiele eines Basketball-Teams und der pro Spiel erzielten mittleren Punktezahl des Teams? Einem Streudiagramm läßt sich entnehmen, dass zwischen diesen Variablen eine lineare Beziehung besteht. Die Anzahl gewonnener Spiele und die erzielte Punktezahl des Gegners stehen gleichfalls in linearer Beziehung zueinander. Diese Variablen enthalten eine negative Beziehung. Einer steigenden Anzahl gewonnener Spiele steht eine fallende mittlere Punktezahl des Gegners gegenüber. Mit der linearen Regression können Sie die Beziehung dieser Variablen modellieren. Mit einem geeigneten Modell lassen sich Spielgewinne von Teams vorhersagen.

Statistiken.Für jede Variable: Anzahl gültiger Fälle, Mittelwert und Standardabweichung. Für jedes Modell: Regressionskoeffizienten, Korrelationsmatrix, Teil- und partielle Korrelationen, multiples R, R^2, korrigiertes R^2, Änderung in R^2, Standardfehler der Schätzung, Tabelle der Varianzanalyse, vorhergesagte Werte und Residuen. Außerdem 95%-Konfidenzintervalle für jeden Regressionskoeffizienten, Varianz-Kovarianz-Matrix, Inflationsfaktor der Varianz, Toleranz, Durbin-Watson-Test, Distanzmaße (Mahalanobis, Cook und Hebelwerte), DfBeta, DfFit, Vorhersageintervalle und fallweise Diagnose. Diagramme: Streudiagramme, partielle Diagramme, Histogramme und Normalverteilungsdiagramme.

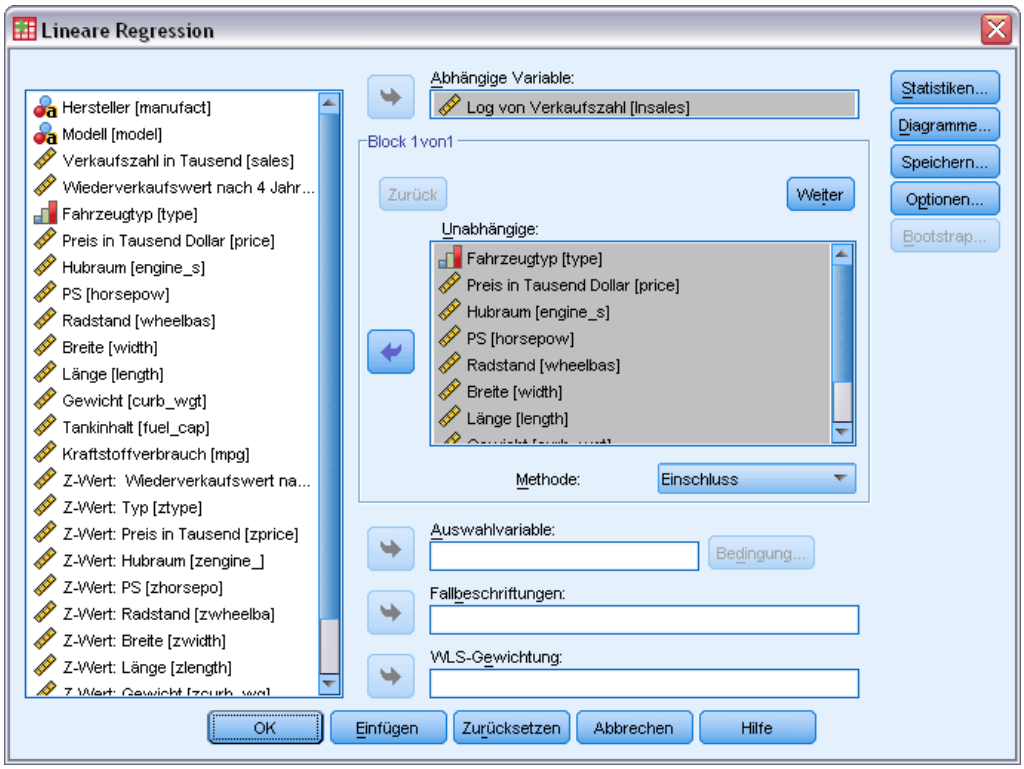
Daten. Die abhängigen und die unabhängigen Variablen müssen quantitativ sein. Kategoriale Variablen, wie beispielsweise Religion, Studienrichtung oder Wohnsitz, müssen in binäre (Dummy-)Variablen oder andere Typen von Kontrast-Variablen umkodiert werden.

Annahmen.Für jeden Wert der unabhängigen Variablen muss die abhängige Variable normalverteilt vorliegen. Die Varianz der Verteilung der abhängigen Variablen muss für alle Werte der unabhängigen Variablen konstant sein. Die Beziehung zwischen der abhängigen Variablen und allen unabhängigen Variablen sollte linear sein, und alle Beobachtungen sollten unabhängig sein.

So lassen Sie eine lineare Regressionsanalyse berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Regression > Linear...

Abbildung 16-1
Dialogfeld "Lineare Regression"



- ▶ Wählen Sie im Dialogfeld "Lineare Regression" eine numerische abhängige Variable aus.
- ▶ Wählen Sie eine oder mehrere numerische unabhängige Variablen aus.

Die folgenden Optionen sind verfügbar:

- Unabhängige Variablen können in Blöcken zusammengefaßt werden, und es können verschiedene Einschlussmethoden für unterschiedliche Untergruppen von Variablen angegeben werden.
- Auswahlvariablen zum Begrenzen der Analyse auf eine Untergruppe von Fällen mit einem bestimmten Wert oder bestimmten Werten für diese Variable können ausgewählt werden.
- Es können Variablen zur Fallunterscheidung ausgewählt werden, um Punkte in Diagrammen zu identifizieren.
- Wählen Sie eine numerische Variable für die WLS-Gewichtung aus, um eine Analyse der gewichteten kleinsten Quadrate durchzuführen.

WLS (Gewichtete kleinste Quadrate). Hiermit können Sie ein Modell gewichteter kleinster Quadrate berechnen. Die Datenpunkte werden mit dem reziproken Wert ihrer Varianzen gewichtet. Dies bedeutet, dass Beobachtungen mit großen Varianzen die Analyse weniger beeinflussen als Beobachtungen mit kleinen Varianzen. Wenn der Wert der Gewichtungsvariablen null, negativ oder fehlend ist, wird der Fall aus der Analyse ausgeschlossen.

Lineare Regression: Methode zur Auswahl von Variablen

Durch die Auswahl der Methode können Sie festlegen, wie unabhängige Variablen in die Analyse eingeschlossen werden. Anhand verschiedener Methoden können Sie eine Vielfalt von Regressionsmodellen mit demselben Satz von Variablen erstellen.

- **Einschluss (Regression).** Eine Prozedur für die Variablenauswahl, bei der alle Variablen eines Blocks in einem einzigen Schritt aufgenommen werden.
- **Schrittweise.** Bei jedem Schritt wird die noch nicht in der Gleichung enthaltene unabhängige Variable mit der kleinsten F-Wahrscheinlichkeit aufgenommen, sofern diese Wahrscheinlichkeit klein genug ist. Bereits in der Regressionsgleichung enthaltene Variablen werden entfernt, wenn ihre F-Wahrscheinlichkeit hinreichend groß wird. Das Verfahren endet, wenn keine Variablen mehr für Aufnahme oder Ausschluss infrage kommen.
- **Entfernen.** Ein Verfahren zur Variablenauswahl, bei dem alle Variablen eines Blocks in einem Schritt ausgeschlossen werden.
- **Rückwärtselimination.** Eine Methode zur Variablenauswahl, bei der alle Variablen in die Gleichung aufgenommen und anschließend sequenziell ausgeschlossen werden. Die Variable mit der kleinsten Teilkorrelation zur abhängigen Variablen wird als erste für den Ausschluss in Betracht gezogen. Wenn sie das Ausschlusskriterium erfüllt, wird sie entfernt. Nach dem Ausschluss der ersten Variablen wird die nächste Variable mit der kleinsten Teilkorrelation in Betracht gezogen. Das Verfahren wird beendet, wenn keine Variablen mehr zur Verfügung stehen, die die Ausschlusskriterien erfüllen.
- **Vorwärtsselektion.** Ein Verfahren zur schrittweisen Variablenauswahl, in dem die Variablen nacheinander in das Modell aufgenommen werden. Die erste Variable, die in Betracht gezogen wird, ist die mit der größten positiven bzw. negativen Korrelation mit der abhängigen Variablen. Diese Variable wird nur dann in die Gleichung aufgenommen, wenn sie das Aufnahmekriterium erfüllt. Wenn die erste Variable aufgenommen wurde, wird als Nächstes die unabhängige Variable mit der größten partiellen Korrelation betrachtet. Das Verfahren endet, wenn keine verbliebene Variable das Aufnahmekriterium erfüllt.

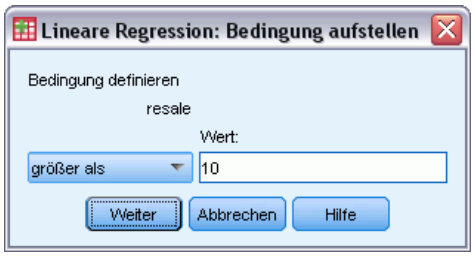
Die Signifikanzwerte in Ihrer Ausgabe basieren auf der Berechnung eines einzigen Modells. Deshalb sind diese generell ungültig, wenn eine schrittweise Methode (schrittweise, vorwärts oder rückwärts) verwendet wird.

Alle Variablen müssen das Toleranzkriterium erfüllen, um unabhängig von der angegebenen Einschlussmethode in die Gleichung einbezogen zu werden. In der Standardeinstellung beträgt der Toleranzwert 0,0001. Eine Variable wird auch dann nicht eingeschlossen, wenn dadurch die Toleranz einer Variablen im Modell unter das Toleranzkriterium abfallen würde.

Alle ausgewählten unabhängigen Variablen werden einem einzigen Regressionsmodell hinzugefügt. Sie können jedoch verschiedene Einschlussmethoden für unterschiedliche Untergruppen von Variablen angeben. Beispielsweise können Sie einen Block von Variablen durch schrittweises Auswählen und einen zweiten Block durch Vorwärtsselektion in das Regressionsmodell einschließen. Um einem Regressionsmodell einen zweiten Block von Variablen hinzuzufügen, klicken Sie auf Weiter.

Lineare Regression: Bedingung aufstellen

Abbildung 16-2
Dialogfeld "Lineare Regression: Bedingung aufstellen"



Die durch die Auswahlbedingung definierten Fälle werden in die Analyse eingeschlossen. Wenn Sie für die Variable beispielsweise gleich wählen und als Wert 5 eingeben, werden nur Fälle in die Analyse einbezogen, für die der Wert der gewählten Variablen gleich 5 ist. Ein String-Wert ist ebenfalls möglich.

Lineare Regression: Diagramme

Abbildung 16-3
Dialogfeld "Lineare Regression: Diagramme"

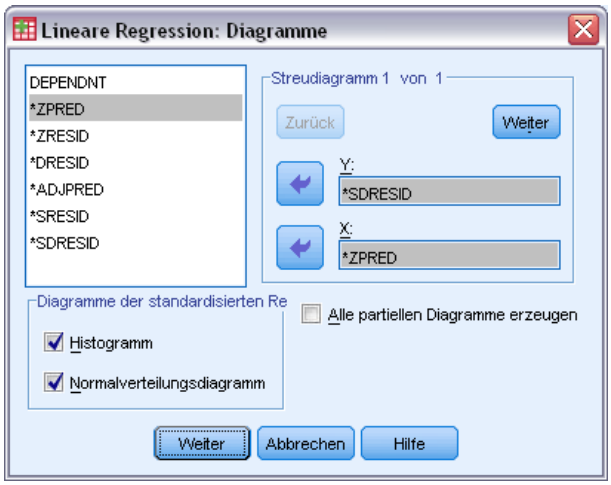


Diagramme können beim Validieren der Annahmen von Normalverteilung, Linearität und Varianz-Gleichheit hilfreich sein. Diagramme dienen auch zum Auffinden von Ausreißern, ungewöhnlichen Beobachtungen und Einflußfällen. Nachdem sie als neue Variablen gespeichert wurden, stehen im Daten-Editor vorhergesagte Werte, Residuen und andere diagnostische Hilfsmittel zum Erstellen von Diagrammen mit den unabhängigen Variablen zur Verfügung. Folgende Diagramme sind verfügbar:

Streudiagramme. Sie können je zwei der folgenden Elemente auftragen: die abhängige Variable, standardisierte vorhergesagte Werte, standardisierte Residuen, ausgeschlossene Residuen, korrigierte vorhergesagte Werte, studentisierte Residuen oder studentisierte ausgeschlossene Residuen. Tragen Sie die standardisierten Residuen über den standardisierten vorhergesagten Werten auf, um auf Linearität und Varianz-Gleichheit zu überprüfen.

Quellvariablenliste. Listet die abhängigen Variablen (DEPENDNT) und die folgenden vorhergesagten Variablen und Residuen-Variablen auf: standardisierte vorhergesagte Werte (*ZPRED), standardisierte Residuen (*ZRESID), ausgeschlossene Residuen (*DRESID), korrigierte vorhergesagte Werte (*ADJPRED), studentisierte Residuen (*SRESID) und studentisierte ausgeschlossene Residuen (*SDRESID).

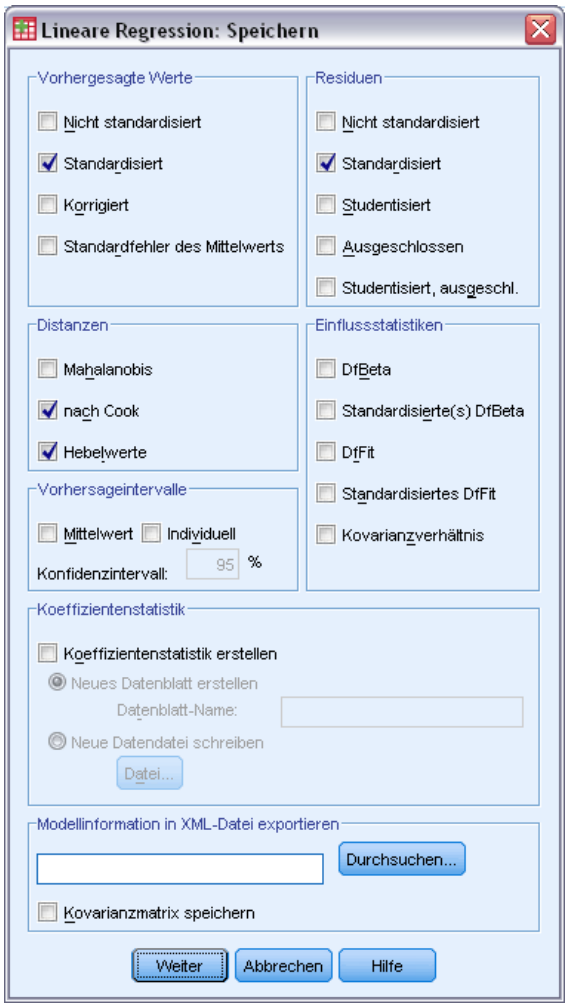
Alle partiellen Diagramme erzeugen. Erzeugt Streudiagramme der Residuen aller unabhängigen Variablen und der Residuen der abhängigen Variablen, wenn für den Rest der unabhängigen Variablen beide Variablen einer getrennten Regression unterzogen werden. Zum Erzeugen eines partiellen Diagramms müssen mindestens zwei unabhängige Variablen in der Gleichung enthalten sein.

Diagramme der standardisierten Residuen. Sie können Histogramme standardisierter Residuen und Normalverteilungsdiagramme anfordern, welche die Verteilung standardisierter Residuen mit einer Normalverteilung vergleichen.

Beim Anfordern von Diagrammen werden Auswertungsstatistiken für standardisierte vorhergesagte Werte und standardisierte Residuen (*ZPRED und *ZRESID) angezeigt.

Lineare Regression: Speichern von neuen Variablen

Abbildung 16-4
Dialogfeld "Lineare Regression: Speichern"



Vorhergesagte Werte, Residuen und andere für die Diagnose nützliche Statistiken können gespeichert werden. Mit jedem Auswahlvorgang werden Ihrer Datendatei eine oder mehrere neue Variablen hinzugefügt.

Vorhergesagte Werte. Dies sind die nach dem Regressionsmodell für jeden Fall vorhergesagten Werte.

- **Nicht standardisiert.** Der Wert, den das Modell für die abhängige Variable vorhersagt.
- **Standardisiert.** Eine Transformation jedes vorhergesagten Werts in dessen standardisierte Form. Das heißt, dass die Differenz zwischen dem vorhergesagten Wert und dem mittleren vorhergesagten Wert durch die Standardabweichung der vorhergesagten Werte geteilt wird. Standardisierte vorhergesagte Werte haben einen Mittelwert von 0 und eine Standardabweichung von 1.

- **Korrigiert.** Der vorhergesagte Wert für einen Fall, wenn dieser Fall von der Berechnung der Regressionskoeffizienten ausgeschlossen ist.
- **Standardfehler des Mittelwerts.** Standardfehler der vorhergesagten Werte. Ein Schätzwert der Standardabweichung des Durchschnittswertes der abhängigen Variablen für die Fälle, die denselben Werte für die unabhängigen Variablen haben.

Distanzen. Dies sind Maße zum Auffinden von Fällen mit ungewöhnlichen Wertekombinationen bei der unabhängigen Variablen und von Fällen, die einen großen Einfluß auf das Modell haben könnten.

- **Mahalanobis.** Dieses Maß gibt an, wie weit die Werte der unabhängigen Variablen eines Falles vom Mittelwert aller Fälle abweichen. Ein großer Mahalanobis-Abstand charakterisiert einen Fall, der bei einer oder mehreren unabhängigen Variablen Extremwerte besitzt.
- **nach Cook.** Ein Maß dafür, wie stark sich die Residuen aller Fälle ändern würden, wenn ein spezieller Fall von der Berechnung der Regressionskoeffizienten ausgeschlossen würde. Ein großer Wert der Cook-Distanz zeigt an, dass der Ausschluss eines Falles von der Berechnung der Regressionskoeffizienten die Koeffizienten substanziell verändert.
- **Hebelwerte.** Werte, die den Einfluss eines Punktes auf die Anpassung der Regression messen. Der zentrierte Wert für die Hebelwirkung bewegt sich zwischen 0 (kein Einfluss auf die Anpassung) und (N-1)/N.

Vorhersageintervalle. Die oberen und unteren Grenzen sowohl für Mittelwert als auch für einzelne Vorhersageintervalle.

- **Mittelwert.** Unter- und Obergrenze (zwei Variablen) für das Vorhersageintervall für den mittleren vorhergesagten Wert.
- **Individuell.** Unter- und Obergrenze (zwei Variablen) für das Vorhersageintervall der abhängigen Variablen für einen Einzelfall.
- **Konfidenzintervall.** Geben Sie einen Wert zwischen 1 und 99,99 ein, um das Konfidenzniveau für die beiden Vorhersageintervalle festzulegen. Wählen Sie "Mittelwert" oder "Individuell" aus, bevor Sie diesen Wert eingeben. Typische Werte für Konfidenzniveaus sind 90, 95 und 99.

Residuen. Der tatsächliche Wert der abhängigen Variablen minus des vorhergesagten Werts aus der Regressionsgleichung.

- **Nicht standardisiert.** Die Differenz zwischen einem beobachteten Wert und dem durch das Modell vorhergesagten Wert.
- **Standardisiert.** Der Quotient aus dem Residuum und einem Schätzer seiner Standardabweichung. Standardisierte Residuen, auch bekannt als Pearson-Residuen, haben einen Mittelwert von 0 und eine Standardabweichung von 1.
- **Studentisiert.** Ein Residuum, das durch seine geschätzte Standardabweichung geteilt wird, die je nach der Distanz zwischen den Werten der unabhängigen Variablen des Falles und dem Mittelwert der unabhängigen Variablen von Fall zu Fall variiert.

- **Ausgeschlossen.** Das Residuum für einen Fall, wenn dieser Fall nicht in die Berechnung der Regressionskoeffizienten eingegangen ist. Es ist die Differenz zwischen dem Wert der abhängigen Variablen und dem korrigierten Schätzwert.
- **Studentisiert, ausgeschl..** Der Quotient aus dem ausgeschlossenen Residuum eines Falles und seinem Standardfehler. Die Differenz zwischen einem studentisierten ausgeschlossenen Residuum und dem zugehörigen studentisierten Residuum gibt an, welchen Unterschied die Entfernung eines Falles für dessen eigene Vorhersage bewirkt.

Einflußstatistiken. Die Änderung in den Regressionskoeffizienten (DfBeta[s]) und vorhergesagten Werten (DfFit), die sich aus dem Ausschluss eines bestimmten Falls ergibt. Standardisierte DfBetas- und DfFit-Werte stehen zusammen mit dem Kovarianzverhältnis zur Verfügung.

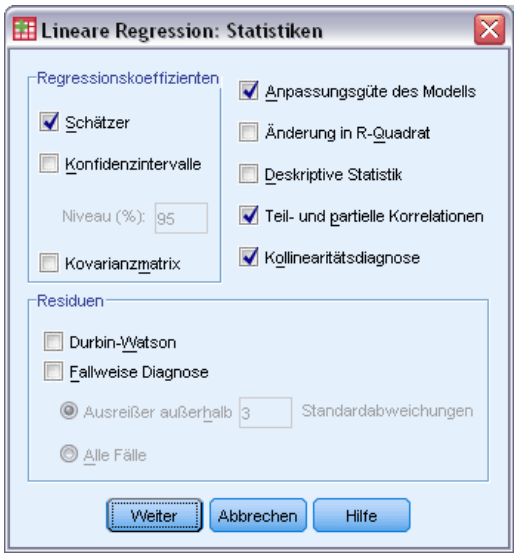
- **Differenz in Beta (DfBeta(s)).** Die Differenz im Beta-Wert entspricht der Änderung im Regressionskoeffizienten, die sich aus dem Ausschluss eines bestimmten Falls ergibt. Für jeden Term im Modell, einschließlich der Konstanten, wird ein Wert berechnet.
- **Standardisiertes DfBeta.** Die standardisierte Differenz im Beta-Wert. Die Änderung des Regressionskoeffizienten, die sich durch den Ausschluss eines bestimmten Falls ergibt. Es empfiehlt sich, Fälle mit absoluten Werten größer als 2 geteilt durch die Quadratwurzel von N zu überprüfen, wenn N die Anzahl der Fälle darstellt. Für jeden Term im Modell, einschließlich der Konstanten, wird ein Wert berechnet.
- **Differenz im vorhergesagten Wert (DfFit).** Die Änderung im vorhergesagten Wert, die sich aus dem Ausschluss eines bestimmten Falls ergibt.
- **Standardisiertes DfFit.** Die standardisierte Differenz im Anpassungswert. Die Änderung des vorhergesagten Werts, die sich durch den Ausschluss eines bestimmten Falls ergibt. Es empfiehlt sich, Fälle mit absoluten Werten größer als 2 geteilt durch die Quadratwurzel von p/N zu überprüfen, wobei p die Anzahl der unabhängigen Variablen im Modell und N die Anzahl der Fälle darstellt.
- **Kovarianzverhältnis.** Das Verhältnis der Determinante der Kovarianzmatrix bei Ausschluss eines bestimmten Falles von der Berechnung des Regressionskoeffizienten zur Determinante der Kovarianzmatrix bei Einschluss aller Fälle. Wenn der Quotient dicht bei 1 liegt, beeinflusst der ausgeschlossene Fall die Kovarianzmatrix nur unwesentlich.

Koeffizientenstatistik. Speichert den Regressionskoeffizienten in einem Daten-Set oder in einer Datendatei. Daten-Sets sind für die anschließende Verwendung in der gleichen Sitzung verfügbar, werden jedoch nicht als Dateien gespeichert, sofern Sie diese nicht ausdrücklich vor dem Beenden der Sitzung speichern. Die Namen von Daten-Sets müssen den Regeln zum Benennen von Variablen entsprechen.

Modellinformation in XML-Datei exportieren. Parameterschätzer und (wahlweise) ihre Kovarianzen werden in die angegebene Datei exportiert. Anhand dieser Modelldatei können Sie die Modellinformationen zu Bewertungszwecken auf andere Datendateien anwenden.

Lineare Regression: Statistiken

Abbildung 16-5
Dialogfeld "Statistiken"



Folgende Statistiken sind verfügbar:

Regressionskoeffizienten. Mit Schätzer zeigen Sie den Regressionskoeffizienten *B*, den Standardfehler von *B*, das Beta des standardisierten Koeffizienten, den *t*-Wert für *B* und das zweiseitige Signifikanzniveau von *t* an. Mit Konfidenzintervalle zeigen Sie Konfidenzintervalle mit dem angegebenen Konfidenzniveau für jeden Regressionskoeffizienten oder eine Kovarianzmatrix an. Mit Kovarianzmatrix wird eine Varianz-Kovarianz-Matrix von Regressionskoeffizienten mit Kovarianzen angezeigt, die nicht auf der Diagonalen liegen, und Varianzen, die auf der Diagonalen liegen. Außerdem wird eine Korrelationsmatrix angezeigt.

Anpassungsgüte des Modells. Die aufgenommenen und entfernten Variablen aus dem Modell werden aufgelistet, und die folgenden Statistiken der Anpassungsgüte werden angezeigt: multiples *R*, R^2 und korrigiertes R^2, Standardfehler der Differenz und eine Tabelle zur Varianzanalyse.

Änderung in R-Quadrat. Die Änderung in R^2, die aus dem Hinzufügen oder Entfernen einer unabhängigen Variablen resultiert. Wenn die durch eine Variable bewirkte Änderung in R^2 groß ist, bedeutet dies, dass diese Variable eine aussagekräftige Einflußvariable für die abhängige Variable ist.

Deskriptive Statistiken. Liefert die Anzahl gültiger Fälle, Mittelwert und Standardabweichung für jede Variable in der Analyse. Außerdem werden eine Korrelationsmatrix mit einem einseitigen Signifikanzniveau und die Anzahl der Fälle für jede Korrelation angezeigt.

Partielle Korrelation. Die Korrelation, die zwischen zwei Variablen verbleibt, nachdem die Korrelation entfernt wurde, die aus dem wechselseitigen Zusammenhang mit den anderen Variablen stammt. Die Korrelation zwischen der abhängigen Variablen und einer unabhängigen Variablen, wenn die linearen Effekte der anderen unabhängigen Variablen im Modell aus der unabhängigen Variablen entfernt wurden.

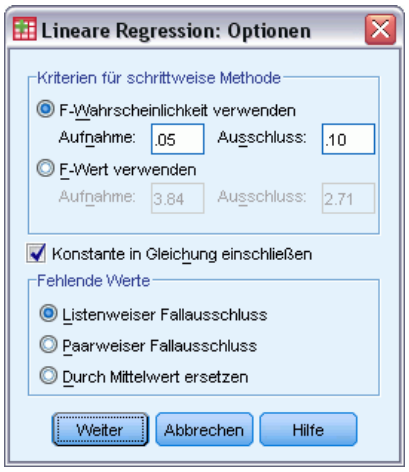
Teilkorrelation. Die Korrelation zwischen der abhängigen Variablen und einer unabhängigen Variablen, wenn die linearen Effekte der anderen unabhängigen Variablen im Modell aus der unabhängigen Variablen entfernt wurden. Die Korrelation entspricht der Änderung in R-Quadrat beim Addieren einer Variablen zu einer Gleichung. Zuweilen als semipartielle Korrelation bezeichnet.

Kollinearitätsdiagnose. Kollinearität (oder Multikollinearität) ist die unerwünschte Situation, in der eine unabhängige Variable eine lineare Funktion anderer unabhängiger Variablen ist. Eigenwerte der skalierten und unzentrierten Kreuzproduktmatrix, Bedingungsindexe und Proportionen der Varianzzerlegung werden zusammen mit Varianzfaktoren (VIF) und Toleranzen für einzelne Variablen angezeigt.

Residuen. Hiermit werden der Durbin-Watson-Test für Reihenkorrelationen der Residuen sowie die fallweise Diagnose für die Fälle angezeigt, die das Auswahlkriterium (Ausreißer über *n* Standardabweichungen) erfüllen.

Lineare Regression: Optionen

Abbildung 16-6
Dialogfeld "Lineare Regression: Optionen"



Die folgenden Optionen sind verfügbar:

Kriterien für schrittweise Methode. Diese Optionen eignen sich für den Fall, dass die Vorwärts-, Rückwärts- oder schrittweise Methode der Variablenauswahl angegeben wurde. Variablen im Modell können abhängig entweder von der Signifikanz (Wahrscheinlichkeit) des *F*-Werts oder vom *F*-Wert selbst eingeschlossen oder entfernt werden.

- **F-Wahrscheinlichkeit verwenden.** Eine Variable wird in das Modell aufgenommen, wenn das Signifikanzniveau ihres F-Werts kleiner ist als der Aufnahmewert. Sie wird ausgeschlossen, wenn das Signifikanzniveau größer ist als der Ausschlusswert. Der Aufnahmewert muss kleiner sein als der Ausschlusswert und beide Werte müssen positiv sein. Um mehr Variablen

in das Modell aufzunehmen, erhöhen Sie den Aufnahmewert. Um mehr Variablen aus dem Modell auszuschließen, senken Sie den Ausschlusswert.

- **F-Wert verwenden.** Eine Variable wird in ein Modell aufgenommen, wenn ihr F-Wert größer ist als der Aufnahmewert. Sie wird ausgeschlossen, wenn der F-Wert kleiner ist als der Ausschlusswert. Der Aufnahmewert muss größer sein als der Ausschlusswert und beide Werte müssen positiv sein. Um mehr Variablen in das Modell aufzunehmen, senken Sie den Aufnahmewert. Um mehr Variablen aus dem Modell auszuschließen, erhöhen Sie den Ausschlusswert.

Konstante in Gleichung einschließen. Als Voreinstellung enthält das Regressionsmodell einen konstanten Term. Wenn diese Option deaktiviert ist, wird die Regression durch den Ursprung gezwungen (selten verwendet). Manche Resultate einer durch den Ursprung verlaufenden Regression lassen sich nicht mit denen einer Regression vergleichen, die eine Konstante aufweist. Beispielsweise kann R^2 nicht in der üblichen Weise interpretiert werden.

Fehlende Werte. Sie können eine der folgenden Optionen auswählen:

- **Listenweiser Fallausschluss.**Nur Fälle mit gültigen Werten für alle Variablen werden in die Analyse einbezogen.
- **Paarweiser Fallausschluss.**Fälle mit vollständigen Daten für das korrelierte Variablenpaar werden zum Berechnen des Korrelationskoeffizienten verwendet, auf dem die Regressionsanalyse basiert. Freiheitsgrade basieren auf dem minimalen paarweisen *N*.
- **Durch Mittelwert ersetzen.**Alle Fälle werden für Berechnungen verwendet, wobei der Mittelwert der Variablen die fehlenden Beobachtungen ersetzt.

Zusätzliche Funktionen beim Befehl REGRESSION

Mit der Befehlssyntax können Sie auch Folgendes:

- Schreiben einer Korrelationsmatrix oder Einlesen einer Matrix anstelle der Rohdaten, um eine Regressionsanalyse zu erhalten (mit dem Unterbefehl `MATRIX`)
- Angeben von Toleranzniveaus (mit dem Unterbefehl `CRITERIA`)
- Berechnen mehrerer Modelle für dieselben oder unterschiedliche abhängige Variablen (mit den Unterbefehlen `METHOD` und `DEPENDENT`)
- Berechnen zusätzlicher Statistiken (mit den Unterbefehlen `DESCRIPTIVES` und `STATISTICS`)

Siehe *Befehlssyntaxreferenz* für die vollständigen Syntaxinformationen.

Kapitel
17

Ordinale Regression

Die ordinale Regression ermöglicht es, die Abhängigkeit einer polytomen ordinalen Antwortvariablen von einer Gruppe von Einflußvariablen zu modellieren. Bei diesen kann es sich um Faktoren oder Kovariaten handeln. Die Gestaltung der ordinalen Regression basiert auf der Methodologie von McCullagh (1980, 1998). In der Syntax wird diese Prozedur als `PLUM` bezeichnet.

Das Standardverfahren der linearen Regressionsanalyse beinhaltet die Minimierung der Summe von quadrierten Differenzen zwischen einer Antwortvariablen (abhängig) und einer gewichteten Kombination von Einflußvariablen (unabhängig). Die geschätzten Koeffizienten geben die Auswirkung einer Änderung in den Einflußvariablen auf die Antwortvariable wieder. Es wird angenommen, daß die Antwortvariable in dem Sinne numerisch ist, daß die Änderungen im Niveau der Antwortvariablen über die gesamte Spannweite der Antwortvariablen gleich sind. So beträgt die Differenz in der Körpergröße zwischen einer Person mit einer Größe von 150 cm und einer Person mit einer Größe von 140 cm beispielsweise 10 cm. Diese Angabe hat die gleiche Bedeutung wie die Differenz zwischen einer Person mit einer Größe von 210 cm und einer Person mit einer Größe von 200 cm. Bei ordinalen Variablen sind diese Beziehungen jedoch nicht notwendigerweise gegeben. Bei diesen Variablen kann die Auswahl und Anzahl von Antwortkategorien willkürlich ausfallen.

Beispiel. Die ordinale Regression kann verwendet werden, um die Reaktion von Patienten auf verschiedene Dosierungen eines Medikaments zu untersuchen. Die möglichen Reaktionen werden als *keine*, *mild*, *moderat* oder *stark* kategorisiert. Der Unterschied zwischen einer milden und einer moderaten Reaktion kann schwer oder gar nicht quantifiziert werden. Er gründet sich vielmehr auf reine Wahrnehmung. Der Unterschied zwischen einer milden und einer moderaten Reaktion kann darüber hinaus auch größer oder kleiner als der Unterschied zwischen einer moderaten und einer starken Reaktion ausfallen.

Statistiken und Diagramme. Beobachtete und erwartete Häufigkeiten und kumulative Häufigkeiten, Pearson-Residuen für Häufigkeiten und kumulative Häufigkeiten, beobachtete und erwartete Wahrscheinlichkeiten, beobachtete und erwartete kumulative Wahrscheinlichkeiten jeder Antwortkategorie nach Kovariaten-Struktur, asymptotische Korrelations- und Kovarianzmatrizen der Parameterschätzer, Pearson-Chi-Quadrat und Likelihood-Quotienten-Chi-Quadrat, Statistik der Anpassungsgüte, Iterationsprotokoll, Test der Annahme von parallelen Linien, Parameterschätzer, Standardfehler, Konfidenzintervalle sowie R^2 nach Cox und Snell, Nagelkerke und McFadden.

Daten. Es wird angenommen, daß die abhängige Variable ordinal ist. Sie kann eine numerische oder eine String-Variable sein. Die Reihenfolge richtet sich nach einer aufsteigenden Sortierung der Werte der abhängigen Variablen. Der niedrigste Wert entspricht der ersten Kategorie. Es wird angenommen, daß die Faktorvariablen kategorial sind. Die Kovariaten-Variablen müssen

numerisch sein. Beachten Sie, daß die Verwendung von mehr als einer stetigen Kovariate leicht zu einer sehr umfangreichen Tabelle mit Zellen-Wahrscheinlichkeiten führen kann.

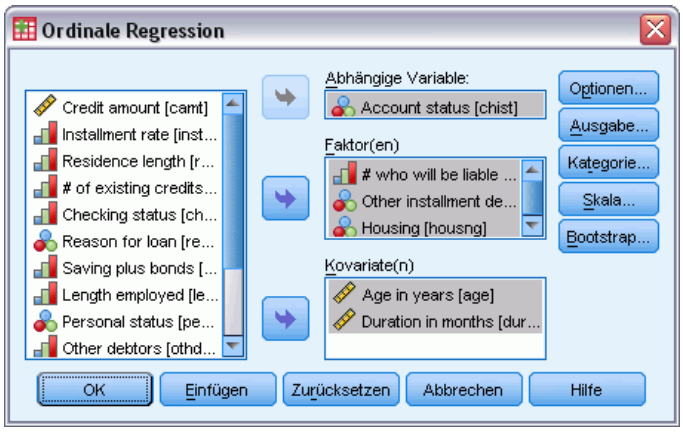
Annahmen. Es darf nur eine Responsevariable vorhanden sein, und diese muß angegeben werden. Zusätzlich wird angenommen, daß die Antworten bei jeder eindeutigen Wertstruktur in den unabhängigen Variablen unabhängige multinomiale Variablen darstellen.

Verwandte Prozeduren. Bei der nominalen logistischen Regression werden ähnliche Modelle für nominale abhängige Variablen verwendet.

Berechnen einer ordinalen Regression

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Regression > Ordinal...

Abbildung 17-1
Dialogfeld "Ordinale Regression"



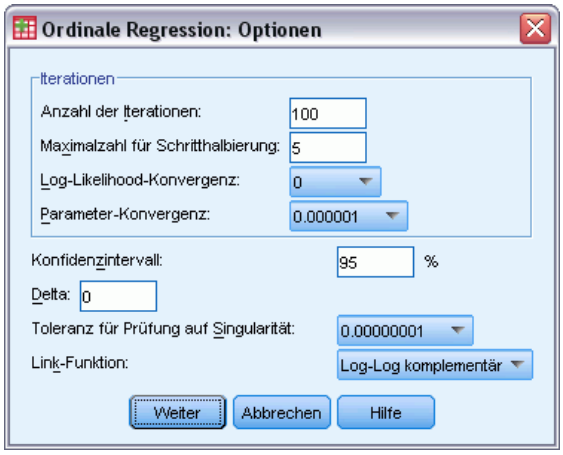
▶ Wählen Sie eine abhängige Variable aus.

▶ Klicken Sie auf OK.

Ordinale Regression: Optionen

Im Dialogfeld "Ordinale Regression: Optionen" können Sie die im iterativen Schätzprozeß verwendeten Parameter anpassen, ein Konfidenzniveau für die Parameterschätzer bestimmen und eine Verknüpfungsfunktion auswählen.

Abbildung 17-2
Dialogfeld "Ordinale Regression: Optionen"



Iterationen. Sie können den Iterationsprozeß anpassen.

- **Maximale Anzahl der Iterationen.**Geben Sie eine nichtnegative Ganzzahl an. Beim Wert 0 gibt die Prozedur die anfänglichen Schätzwerte zurück.
- **Maximalzahl für Schritt-Halbierung.**Geben Sie eine positive Ganzzahl ein.
- **Log-Likelihood-Konvergenz.**Der Prozeß wird beendet, wenn die absolute oder relative Änderung der Log-Likelihood kleiner als dieser Wert ist. Bei einem Wert von 0 wird dieses Kriterium nicht verwendet.
- **Parameter-Konvergenz.**Der Prozeß wird beendet, wenn die absolute oder relative Änderung in jedem der Parameterschätzer kleiner als dieser Wert ist. Bei einem Wert von 0 wird dieses Kriterium nicht verwendet.

Konfidenzintervall. Geben Sie einen Wert größer oder gleich 0 und kleiner als 100 ein.

DeltaDer Wert, der zu Zellen mit einer Häufigkeit von 0 addiert wird. Geben Sie eine nicht-negative Zahl kleiner als 1 an.

Toleranz für Prüfung auf Singularität. Wird zum Prüfen auf stark abhängige Einflußvariablen verwendet. Wählen Sie einen Wert aus der Liste der Optionen aus.

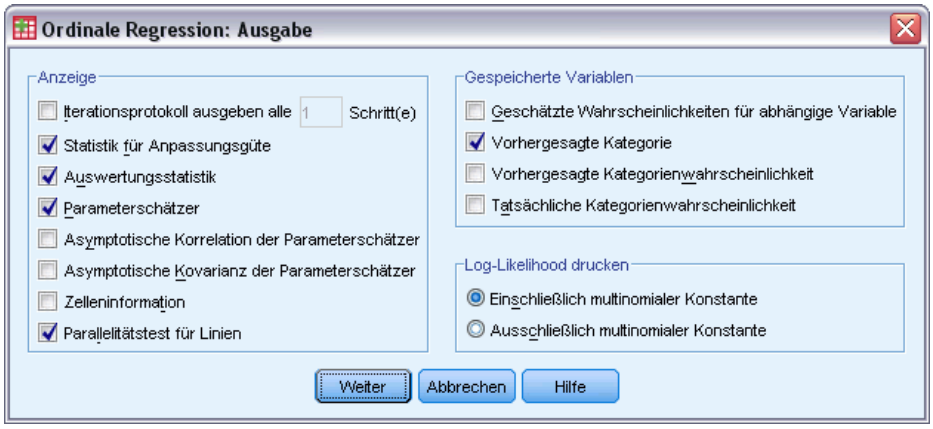
Verknüpfungsfunktion. Die Verknüpfungsfunktion ist eine Transformation der kumulativen Wahrscheinlichkeiten, die eine Schätzung des Modells ermöglicht. Es stehen fünf Verknüpfungsfunktionen zur Verfügung, die in der folgenden Tabelle zusammengefasst sind.

Funktion (Script window, New Procedure)	Form	Typische Anwendung
Logit	$\log(\xi / (1-\xi))$	Gleichmäßig verteilte Kategorien
Log-Log komplementär	$\log(-\log(1-\xi))$	Höhere Kategorien wahrscheinlicher
Log-Log negativ	$-\log(-\log(\xi))$	Niedrigere Kategorien wahrscheinlicher
Probit	$\Phi^{-1}(\xi)$	Latente Variable ist normalverteilt
Cauchit (Inverse von Cauchy)	$\tan(\pi(\xi-0,5))$	Latente Variable weist viele Extremwerte auf

Ordinale Regression: Ausgabe

Im Dialogfeld "Ordinale Regression: Ausgabe" können Sie festlegen, welche Tabellen im Viewer angezeigt werden und ob Variablen in der Arbeitsdatei gespeichert werden.

Abbildung 17-3
Dialogfeld "Ordinale Regression: Ausgabe"



Anzeigen. Es werden die folgenden Tabellen erstellt:

- **Iterationsprotokoll ausgeben.**Die Log-Likelihood und die Parameterschätzer werden mit der hier angegebenen Häufigkeit ausgegeben. Die erste und letzte Iteration wird immer ausgegeben.
- **Statistik für Anpassungsgüte.**Gibt die Chi-Quadrat-Statistik nach Pearson und die Likelihood-Quotienten-Chi-Quadrat-Statistik aus. Diese werden anhand der in der Variablenliste angegebenen Klassifikation berechnet.
- **Auswertungsstatistik.**R^2-Statistik nach Cox und Snell, Nagelkerke und McFadden.
- **Parameterschätzer.**Parameterschätzer, Standardfehler und Konfidenzintervalle.
- **Asymptotische Korrelation der Parameterschätzer.** Matrix der Parameterschätzer-Korrelationen.
- **Asymptotische Kovarianz der Parameterschätzer.** Matrix der Parameterschätzer-Kovarianzen.
- **Zelleninformationen.** Beobachtete und erwartete Häufigkeiten und kumulative Häufigkeiten, Pearson-Residuen für Häufigkeiten und kumulative Häufigkeiten, beobachtete und erwartete Wahrscheinlichkeiten sowie beobachtete und erwartete kumulative Wahrscheinlichkeiten jeder Antwortkategorie nach Kovariaten-Struktur. Bedenken Sie, daß diese Option bei Modellen mit vielen Kovariaten-Strukturen (beispielsweise bei Modellen mit stetigen Kovariaten) zu einer sehr umfassenden, unübersichtlichen Tabelle führen kann.
- **Parallelitätstest für Linien.**Test der Hypothese, daß die Kategorieparameter über alle Niveaus der abhängigen Variablen gleich sind. Dies ist nur bei reinen Kategoriemodellen verfügbar.

Gespeicherte Variablen. Es werden die folgenden Variablen in der Arbeitsdatei gespeichert:

- **Geschätzte Antwortwahrscheinlichkeiten.**Aus dem Modell geschätzte Wahrscheinlichkeiten, daß eine Faktor-/Kovariaten-Struktur in eine Antwortkategorie klassifiziert wird. Es gibt so viele Wahrscheinlichkeiten wie die Anzahl der Antwortkategorien.

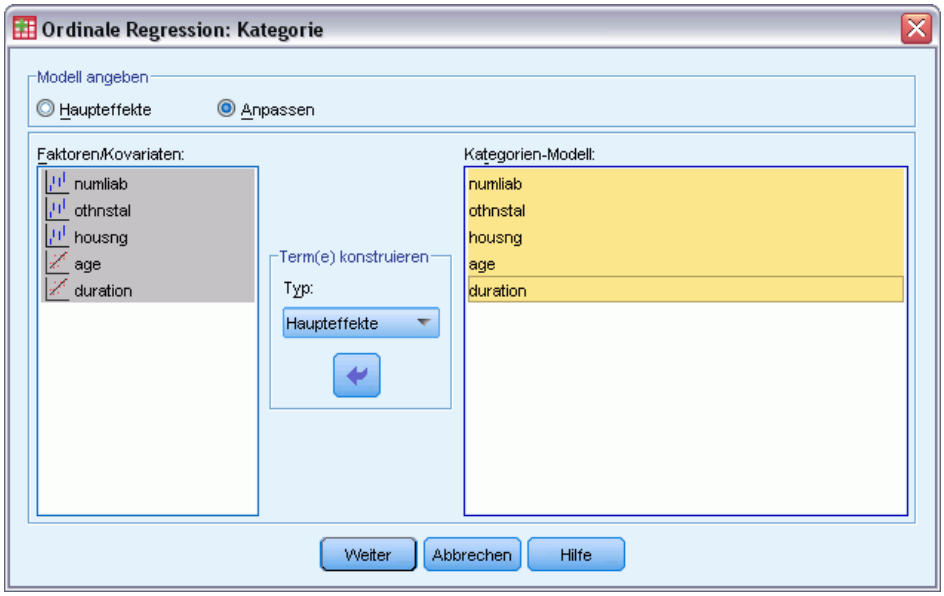
- **Vorhergesagte Kategorie.**Die Antwortkategorie mit der größten geschätzten Wahrscheinlichkeit für eine Faktor-/Kovariaten-Struktur.
- **Vorhergesagte Kategorienwahrscheinlichkeit.**Geschätzte Wahrscheinlichkeit, daß eine Faktor-/Kovariaten-Struktur in die vorhergesagte Kategorie klassifiziert wird. Diese Wahrscheinlichkeit entspricht außerdem der größten geschätzten Wahrscheinlichkeit der Faktor-/Kovariaten-Struktur.
- **Tatsächliche Kategorienwahrscheinlichkeit.**Geschätzte Wahrscheinlichkeit, daß eine Faktor-/Kovariaten-Struktur in die tatsächliche Kategorie klassifiziert wird.

Log-Likelihood drucken. Hiermit wird die Ausgabe der Log-Likelihood festgelegt. Mit Einschließlich multinomialer Konstante wird der vollständige Wert der Likelihood ausgegeben. Wenn Sie die Ergebnisse mit anderen Produkten vergleichen möchten, bei denen keine Konstante vorhanden ist, können Sie diese ausschließen.

Ordinale Regression: Kategorie

Im Dialogfeld "Ordinale Regression: Kategorie" können Sie das Modell für die Analyse kategorisieren.

Abbildung 17-4
Dialogfeld "Ordinale Regression: Kategorie"



Modell bestimmen. Ein Modell mit Haupteffekten enthält die Haupteffekte der Faktoren und Kovariaten, aber keine Wechselwirkungseffekte. Sie können ein benutzerdefiniertes Modell erstellen, um Teilgruppen von Wechselwirkungen zwischen Faktoren oder Kovariaten zu bestimmen.

Faktoren/Kovariaten. Die Faktoren und Kovariaten werden aufgelistet.

Modell kategorisieren. Das Modell ist abhängig von den gewählten Haupt- und Wechselwirkungseffekten.

Terme konstruieren

Für die ausgewählten Faktoren und Kovariaten:

Wechselwirkung. Hiermit wird der Wechselwirkungsterm mit der höchsten Ordnung von allen ausgewählten Variablen erzeugt. Dies ist die Standardeinstellung.

Haupteffekte. Legt einen Haupteffekt-Term für jede ausgewählte Variable an.

Alle 2-Weg. Hiermit werden alle möglichen 2-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 3-Weg. Hiermit werden alle möglichen 3-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

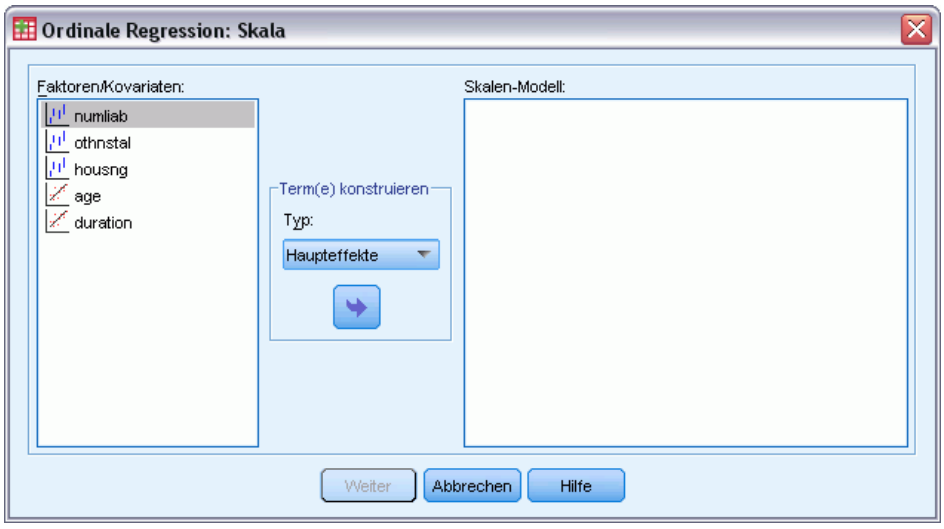
Alle 4-Weg. Hiermit werden alle möglichen 4-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 5-Weg. Hiermit werden alle möglichen 5-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Ordinale Regression: Skala

Im Dialogfeld "Ordinale Regression: Skala" können Sie das Modell für die Analyse skalieren.

Abbildung 17-5
Dialogfeld "Ordinale Regression: Skala"



Faktoren/Kovariaten. Die Faktoren und Kovariaten werden aufgelistet.

Modell skalieren. Das Modell ist abhängig von den gewählten Haupt- und Wechselwirkungseffekten.

Terme konstruieren

Für die ausgewählten Faktoren und Kovariaten:

Wechselwirkung. Hiermit wird der Wechselwirkungsterm mit der höchsten Ordnung von allen ausgewählten Variablen erzeugt. Dies ist die Standardeinstellung.

Haupteffekte. Legt einen Haupteffekt-Term für jede ausgewählte Variable an.

Alle 2-Weg. Hiermit werden alle möglichen 2-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 3-Weg. Hiermit werden alle möglichen 3-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 4-Weg. Hiermit werden alle möglichen 4-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 5-Weg. Hiermit werden alle möglichen 5-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Zusätzliche Funktionen beim Befehl PLUM

Sie können die ordinale Regression an Ihre Bedürfnisse anpassen, wenn Sie ihre Auswahl in ein Syntax-Fenster einfügen und die resultierende Befehlssyntax für den Befehl `PLUM` bearbeiten. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Angepaßte Hypothesentests können durch Festlegen von Nullhypothesen als lineare Parameterkombinationen erstellt werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
18

Kurvenanpassung

Mit der Prozedur "Kurvenanpassung" werden Regressionsstatistiken zur Kurvenanpassung und zugehörige Diagramme für 11 verschiedene Regressionsmodelle zur Kurvenanpassung erstellt. Für jede abhängige Variable wird ein separates Modell erstellt. Außerdem können Sie vorhergesagte Werte, Residuen und Vorhersageintervalle als neue Variablen speichern.

Beispiel. Ein Internet-Dienstanbieter verfolgt den Prozentsatz des mit Viren infizierten E-Mail-Verkehrs über die Netzwerke im Lauf der Zeit. Ein Streudiagramm zeigt, dass eine nichtlineare Beziehung vorliegt. Sie können ein quadratisches oder kubisches Modell an die Daten anpassen und die Gültigkeit der Annahmen sowie die Güte der Anpassung des Modells prüfen.

Statistiken. Für jedes Modell: Regressionskoeffizienten, multiples R, R^2, korrigiertes R^2, Standardfehler des Schätzers, Tabelle für die Varianzanalyse, vorhergesagte Werte, Residuen und Vorhersageintervalle. Modelle: linear, logarithmisch, invers, quadratisch, kubisch, Potenz, zusammengesetzt, S-Kurve, logistisch, Wachstum und exponentiell.

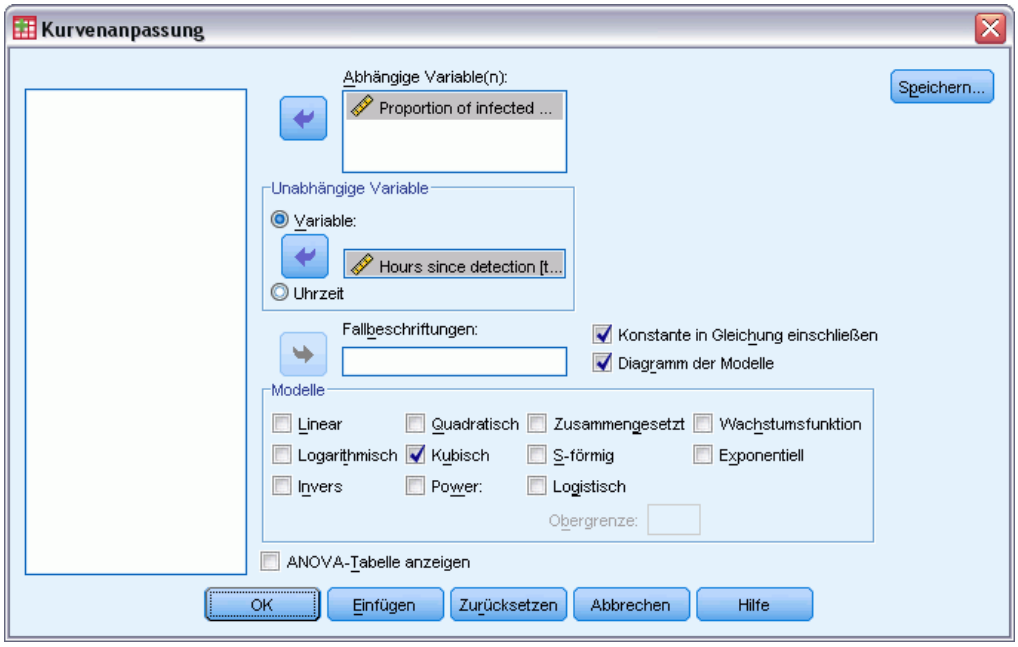
Daten. Die abhängigen und die unabhängigen Variablen müssen quantitativ sein. Wenn Sie aus der Arbeitsdatei Zeit als unabhängige Variable ausgewählt haben (statt eine Variable auszuwählen), erzeugt die Prozedur "Kurvenanpassung" eine Zeitvariable mit gleichen Zeitabständen zwischen den Fällen. Wenn Zeit ausgewählt wurde, sollte die abhängige Variable eine Zeitreihenmessung sein. Zur Zeitreihenanalyse ist eine Datendateistruktur erforderlich, in der jeder Fall (jede Zeile) einen Satz von Beobachtungen zu unterschiedlichen Zeiten bei gleichen Zeitabständen zwischen den Fällen darstellt.

Annahmen. Stellen Sie Ihre Daten grafisch dar, um den Zusammenhang zwischen den unabhängigen und den abhängigen Variablen (linear, exponentiell usw.) erkennen zu können. Die Residuen eines guten Modells müssen willkürlich und normalverteilt sein. Bei einem linearen Modell müssen folgende Annahmen erfüllt werden: Für jeden Wert der unabhängigen Variablen muss die abhängige Variable normalverteilt vorliegen. Die Varianz der Verteilung der abhängigen Variablen muss für alle Werte der unabhängigen Variablen konstant sein. Die abhängige Variable und die unabhängige Variable müssen linear zusammenhängen, und alle Beobachtungen müssen unabhängig sein.

So führen Sie eine Kurvenanpassung durch:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Regression > Kurvenanpassung…

Abbildung 18-1
Dialogfeld "Kurvenanpassung"



- Wählen Sie eine oder mehrere abhängige Variablen aus. Für jede abhängige Variable wird ein separates Modell erstellt.
- Wählen Sie eine unabhängige Variable aus (wählen Sie entweder eine Variable aus der Arbeitsdatei oder wählen Sie Zeit aus).
- Die folgenden Optionen sind verfügbar:
  - Eine Variable zum Beschriften der Fälle in Streudiagrammen auswählen. Sie können für jeden Punkt im Streudiagramm das Symbol zum Identifizieren von Punkten verwenden, um den Wert der Variablen für die "Fallbeschriftung" anzeigen zu lassen.
  - Klicken Sie auf Speichern, um vorhergesagte Werte, Residuen und Vorhersageintervalle als neue Variablen zu speichern.

Außerdem sind folgende Optionen verfügbar:

- **Konstante in Gleichung einschließen.** Mit dieser Option wird ein konstanter Term in der Regressionsgleichung geschätzt. In der Standardeinstellung ist die Konstante eingeschlossen.
- **Diagramm der Modelle.** Mit dieser Option werden für alle ausgewählten Modelle die Werte der abhängigen Variablen über der unabhängigen Variablen grafisch dargestellt. Für jede abhängige Variable wird ein eigenes Diagramm erzeugt.
- **ANOVA-Tabelle anzeigen.** Mit dieser Option wird für jedes ausgewählte Modell eine Zusammenfassung für die Varianzanalyse angezeigt.

Modelle für die Kurvenanpassung

Sie können ein oder mehrere Regressionsmodelle für die Kurvenanpassung auswählen. Stellen Sie Ihre Daten grafisch dar, um zu ermitteln, welches Modell Sie verwenden sollten. Wenn Ihre Variablen in einem linearen Zusammenhang zu stehen scheinen, verwenden Sie ein einfaches lineares Regressionsmodell. Wenn Ihre Variablen in keinem linearen Zusammenhang stehen, transformieren Sie diese. Wenn eine Transformation keine Abhilfe schafft, benötigen Sie möglicherweise ein komplizierteres Modell. Betrachten Sie ein Streudiagramm Ihrer Daten. Wenn das Diagramm einer Ihnen bekannten mathematischen Funktion ähnelt, passen Sie Ihre Daten an diesen Modelltyp an. Wenn Ihre Daten zum Beispiel einer Exponentialfunktion ähneln, verwenden Sie ein exponentielles Modell.

Linear (GLM Ptable). Ein Modell mit der Gleichung Y = b0 + (b1 * t). Die Werte der Zeitreihe werden als lineare Funktion der Zeit aufgefasst.

Logarithmisch. Ein Modell mit der Gleichung Y = b0 + (b1 * ln(t)).

Invers. Ein Modell mit der Gleichung Y = b0 + (b1 / t).

Quadratisch (GLM Ptable). Ein Modell mit folgender Gleichung: Y = b0 + (b1 * t) + (b2 * t**2). Das quadratische Modell kann zum Modellieren von Zeitreihen verwendet werden, die "abheben" oder gedämpft verlaufen.

Kubisch (GLM Ptable). Ein Modell mit folgender Gleichung:Y = b0 + (b1 * t) + (b2 * t**2) + (b3 * t**3).

Exponent. Ein Modell mit folgender Gleichung: Y = b0 * (t**b1) oder ln(Y) = ln(b0) + (b1 * ln(t)).

Zusammengesetzt. Dieses Modell basiert auf folgender Gleichung: Y = b0 * (b1**t) oder ln(Y) = ln(b0) + (ln(b1) * t).

S-Kurve. Ein Modell, dessen Gleichung lautet: Y = e**(b0 + (b1/t)) oder ln(Y) = b0 + (b1/t).

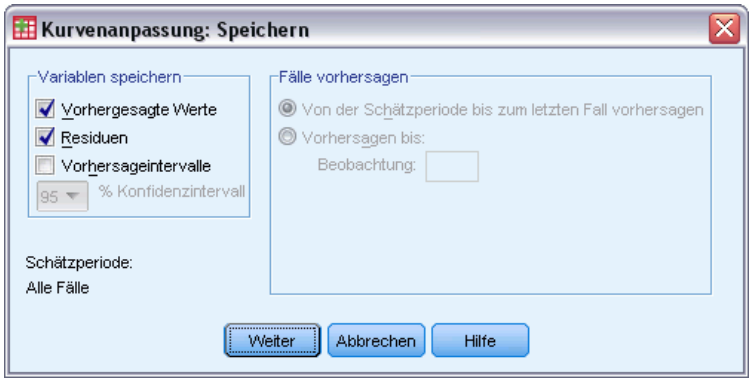
Logistisch (PP Plots Test Dist.). Die Gleichung für dieses Modell lautet Y = 1 / (1/u + (b0 * (b1**t))) oder ln(1/y-1/u) = ln (b0) + (ln(b1) * t), wobei u die obere Schranke ist. Nach der Auswahl von "Logistisch" muss der Wert der oberen Schranke angegeben werden, der in der Regressionsgleichung verwendet werden soll. Der Wert muss eine positive Zahl sein, die größer ist als der größte Wert der abhängigen Variablen.

Wachstum. Ein Modell, dessen Gleichung lautet: Y = e**(b0 + (b1 * t)) oder ln (Y) = b0 + (b1 * t).

Exponentialverteilung. Ein Modell mit folgender Gleichung: Y = b0 * (e**(b1 * t)) oder ln (Y) = ln (b0) + (b1 * t).

Kurvenanpassung: Speichern

Abbildung 18-2
Dialogfeld "Kurvenanpassung: Speichern"



Variablen speichern. Für jedes ausgewählte Modell können Sie vorhergesagte Werte, Residuen (beobachteter Wert der abhängigen Variablen minus vorhergesagter Wert des Modells) und Vorhersageintervalle (Ober- und Untergrenzen) speichern. Die neuen Variablennamen werden mit den beschreibenden Labels in einer Tabelle im Ausgabefenster angezeigt.

Fälle vorhersagen. Wenn Sie in der Arbeitsdatei statt einer Variablen Zeit als unabhängige Variable ausgewählt haben, können Sie nach dem Ende der Zeitreihe eine Vorhersageperiode angeben. Sie können eine der folgenden Möglichkeiten wählen:

- **Von der Schätzperiode bis zum letzten Fall vorhersagen.** Hiermit werden auf der Grundlage der Fälle in der Schätzperiode Werte für alle Fälle in der Datei vorhergesagt. Die unten im Dialogfeld angezeigte Schätzperiode wird im Menü "Daten", Option "Fälle auswählen", Dialogfeld "Fälle auswählen:Bereich" festgelegt. Wenn keine Schätzperiode definiert wurde, werden alle Fälle zur Vorhersage der Werte verwendet.
- **Vorhersagen bis.** Hiermit werden auf der Grundlage der Fälle in der Schätzperiode Werte bis zum angegebenen Datum, zur angegebenen Uhrzeit oder zur angegebenen Beobachtungsnummer vorhergesagt. Mit dieser Funktion können Werte nach dem letzten Fall in der Zeitreihe vorhergesagt werden. Die gegenwärtig definierten Datumsvariablen bestimmen, welche Textfelder zur Verfügung stehen, um das Ende der Vorhersageperiode anzugeben. Wenn keine Datumsvariablen definiert sind, können Sie die letzte Beobachtungs- bzw. Fallnummer angeben.

Datumsvariablen erstellen Sie im Menü "Daten" mit der Option "Datum definieren".

Kapitel
19

Regression mit partiellen kleinsten Quadraten

Die Prozedur "Regression mit partiellen kleinsten Quadraten" schätzt Regressionsmodelle mit partiellen kleinsten Quadraten (Partial Least Squares, PLS, auch als "Projektion auf latente Struktur" (Projection to Latent Structure) bezeichnet). PLS ist ein Vorhersageverfahren, das eine Alternative zum Regressionsmodell der gewöhnlichen kleinsten Quadrate (Ordinary Least Squares,OLS), zur kanonischen Korrelation bzw. zur strukturierten Gleichungsmodellierung darstellt und besonders nützlich ist, wenn die Einflussvariablen eine hohe Korrelation aufweisen oder wenn die Anzahl der Einflussvariablen die Anzahl der Fälle übersteigt.

PLS kombiniert Merkmale der Hauptkomponentenanalyse mit Merkmalen der mehrfachen Regression. Zunächst wird ein Set latenter Faktoren extrahiert, die einen möglichst großen Anteil der Kovarianz zwischen den unabhängigen und den abhängigen Variablen erklären. Anschließend werden in einem Regressionsschritt die Werte der abhängigen Variablen mithilfe der Zerlegung der unabhängigen Variablen vorhergesagt.

Verfügbarkeit. PLS ist ein Erweiterungsbefehl, für den das IBM® SPSS® Statistics - Integration Plug-In for Python auf dem System installiert sein muss, auf dem PLS. Das PLS-Erweiterungsmodul muss separat installiert werden. Es kann auf folgender Webseite heruntergeladen werden: *http://www.spss.com/devcentral*.

Hinweis: Das PLS-Erweiterungsmodul ist von Python-Software abhängig. SPSS Inc. ist nicht der Inhaber bzw. Lizenzgeber der Python-Software. Alle Benutzer von Python müssen den Bedingungen der Python-Lizenzvereinbarung zustimmen, die sich auf der Python-Website befinden. SPSS Inc. macht keinerlei Aussagen über die Qualität des Python-Programms. SPSS Inc. schließt jegliche Haftung im Zusammenhang mit Ihrer Nutzung des Python-Programms aus.

Tabellen. Der Anteil der (durch den latenten Faktor) erklärten Varianz, die Gewichtungen latenter Faktoren, die Ladungen latenter Faktoren, die Bedeutung der unabhängigen Variablen in der Projektion (VIP) und die Schätzer für Regressionsparameter (nach abhängiger Variablen) werden jeweils standardmäßig angegeben.

Diagramme. Die Bedeutung der Variablen in der Projektion (Variable Importance in Projection, VIP), Faktor-Scores, Faktorgewichtungen für die ersten drei latenten Faktoren und die Distanz zum Modell werden jeweils über die Registerkarte Optionen erstellt.

Messniveau. Die abhängigen und unabhängigen Variablen (Einflussvariablen) können metrisch, nominal oder ordinal sein. Bei der Prozedur wird davon ausgegangen, dass allen Variablen das richtige Messniveau zugewiesen wurde. Sie können das Messniveau für eine Variable jedoch vorübergehend ändern. Klicken Sie hierzu mit der rechten Maustaste auf die Variable in der Liste der Quellvariablen und wählen Sie das gewünschte Messniveau im Kontextmenü aus. Kategoriale (nominale bzw. ordinale) Variablen werden von der Prozedur als äquivalent behandelt.

Kodierung für kategoriale Variablen. Die Prozedur kodiert vorübergehend für die Dauer des Verfahrens kategoriale abhängige Variablen mithilfe der "Eins-aus-*c*"-Kodierung neu. Wenn es *c* Kategorien für eine Variable gibt, wird die Variable als *c* Vektoren gespeichert. Dabei wird die erste Kategorie als (1,0,...,0) angegeben, die zweite Kategorie als (0,1,0,...,0), ... und die letzte Kategorie als (0,0,...,0,1). Kategoriale abhängige Variablen werden mithilfe von Dummy-Codierung dargestellt, d. h. es wird einfach der Indikator weggelassen, der der Referenzkategorie entspricht.

Häufigkeitsgewichtungen. Gewichtungswerte werden vor der Verwendung auf die nächste ganze Zahl gerundet. Fälle mit fehlenden Gewichten oder Gewichten unter 0,5 werden in der Analyse nicht verwendet.

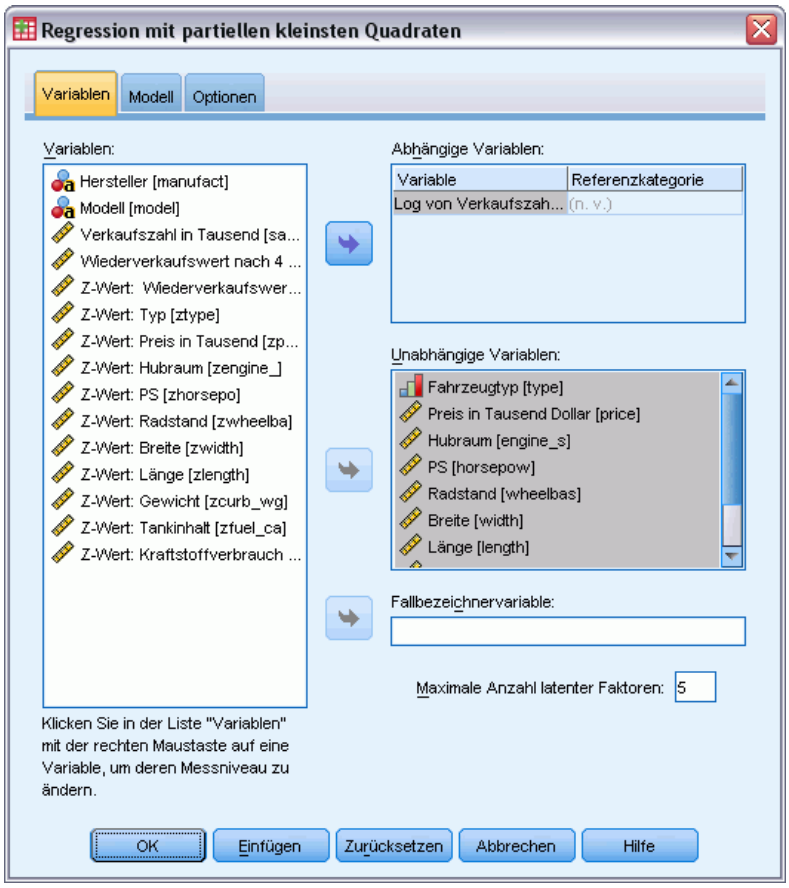
Fehlende Werte. Benutzer- und systemdefiniert fehlende Werte werden als ungültig behandelt.

Neuskalierung. Alle Modellvariablen werden zentriert und standardisiert, einschließlich der Indikatorvariablen die für kategoriale Variablen stehen.

So lassen Sie eine Regression mit partiellen kleinsten Quadraten berechnen:

Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Regression > Partielle kleinste Quadrate...

Abbildung 19-1
Regression mit partiellen kleinsten Quadraten – Registerkarte "Variablen"



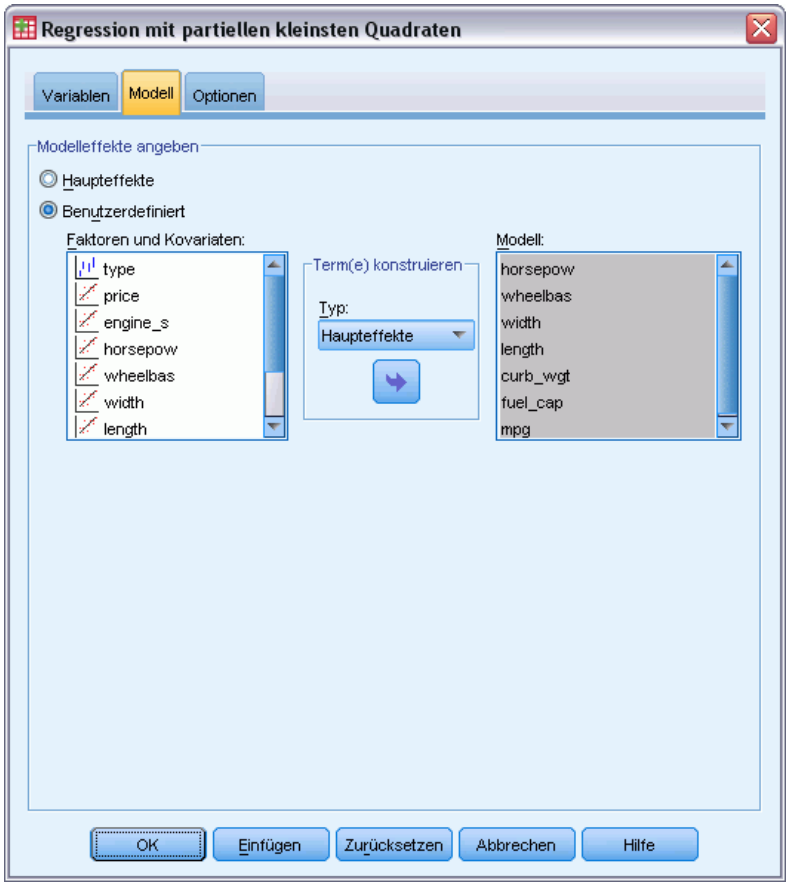
- ▶ Wählen Sie mindestens eine abhängige Variable aus.
- ▶ Wählen Sie mindestens eine unabhängige Variable aus.

Die folgenden Optionen sind verfügbar:

- Angabe einer Referenzkategorie für kategoriale (nominale bzw. ordinale) abhängige Variablen.
- Angabe einer Variablen, die als eindeutige Kennung für die fallweise Ausgabe und für die gespeicherten Daten-Sets verwendet werden soll.
- Angabe einer Obergrenze für die Anzahl der zu extrahierenden latenten Faktoren.

Modell

Abbildung 19-2
Regression mit partiellen kleinsten Quadraten – Registerkarte "Modell"



Modell-Effekte angeben. Ein Modell mit Haupteffekten enthält die Haupteffekte aller Faktoren und Kovariaten. Wählen Sie Benutzerdefiniert, um Interaktionen anzugeben. Sie müssen alle in das Modell zu übernehmenden Terme angeben.

Faktoren und Kovariaten. Die Faktoren und Kovariaten werden aufgelistet.

Modell. Das Modell ist von der Art Ihrer Daten abhängig. Nach der Auswahl von Anpassen können Sie die Haupteffekte und Wechselwirkungen auswählen, die für Ihre Analyse von Interesse sind.

Terme konstruieren

Für die ausgewählten Faktoren und Kovariaten:

Wechselwirkung. Hiermit wird der Wechselwirkungsterm mit der höchsten Ordnung von allen ausgewählten Variablen erzeugt. Dies ist die Standardeinstellung.

Haupteffekte. Legt einen Haupteffekt-Term für jede ausgewählte Variable an.

Alle 2-Weg. Hiermit werden alle möglichen 2-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 3-Weg. Hiermit werden alle möglichen 3-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 4-Weg. Hiermit werden alle möglichen 4-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Alle 5-Weg. Hiermit werden alle möglichen 5-Weg-Wechselwirkungen der ausgewählten Variablen erzeugt.

Optionen

Abbildung 19-3
Regression mit partiellen kleinsten Quadraten – Registerkarte "Optionen"

Auf der Registerkarte "Optionen" kann der Benutzer Modellschätzer für einzelne Fälle, latente Faktoren und Einflussvariablen speichern und grafisch darstellen lassen.

Geben Sie für jeden Datentyp den Namen eines Daten-Sets an. Die Namen der Daten-Sets müssen eindeutig sein. Wenn Sie den Namen eines bestehenden Daten-Sets angeben, werden dessen Inhalte ersetzt; ansonsten wird ein neues Daten-Set erstellt.

- **Schätzer für einzelne Fälle speichern.**Speichert die folgenden fallweisen Modellschätzer: vorhergesagte Werte, Residuen, Distanz zum Modell mit latenten Faktoren und Scores für latente Faktoren. Außerdem werden die Scores für latente Faktoren grafisch dargestellt.

- **Schätzer für latente Faktoren speichern.** Speichert die Ladungen und Gewichtungen latenter Faktoren. Außerdem werden die Gewichtungen für latente Faktoren grafisch dargestellt.
- **Schätzer für unabhängige Variablen speichern.** Speichert Schätzer für Regressionsparameter und die Bedeutung der unabhängigen Variablen in der Projektion (VIP). Außerdem werden die VIP-Werte für die einzelnen latente Faktoren grafisch dargestellt.

Kapitel
20

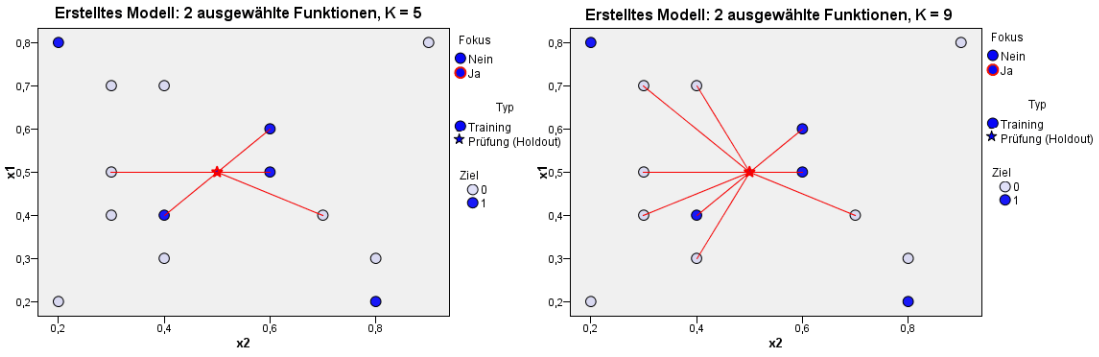
Analyse Nächstgelegener Nachbar

Die Analyse Nächstgelegener Nachbar ist eine Methode für die Klassifizierung von Fällen nach ihrer Ähnlichkeit mit anderen Fällen. Für Machine Learning wurde sie als Methode für die Mustererkennung in Daten ohne exakte Entsprechung mit gespeicherten Mustern oder Fällen entwickelt. Ähnliche Fälle liegen nah beieinander und Fälle mit geringer Ähnlichkeit sind weit voneinander entfernt. Daher kann der Abstand zwischen zwei Fällen als Maß für ihre Unähnlichkeit herangezogen werden.

Fälle, die nah beieinander liegen, werden als "Nachbarn" bezeichnet. Wenn ein neuer Fall (Holdout) vorgelegt wird, wird sein Abstand zu den einzelnen Fällen im Modell berechnet. Die Klassifizierungen der ähnlichsten Fälle – der nächstgelegenen Nachbarn – werden ermittelt und der neue Fall wird in die Kategorie eingeordnet, die die größte Anzahl an nächstgelegenen Nachbarn aufweist.

Sie können die Anzahl an nächstgelegenen Nachbarn angegeben, die untersucht werden sollen. Dieser Wert wird als *k* bezeichnet. Die Bilder zeigen, wie ein neuer Fall mit zwei unterschiedlichen *k*-Werten klassifiziert würde. Bei $k = 5$ wird der neue Fall in die Kategorie *1* eingeordnet, weil ein Großteil der nächstgelegenen Nachbarn der Kategorie *1* angehört. Bei $k = 9$ wird der neue Fall hingegen in die Kategorie *0* eingeordnet, weil ein Großteil der nächstgelegenen Nachbarn der Kategorie *0* angehört.

Abbildung 20-1
Die Effekte der Änderung von k bei der Klassifizierung



Die Analyse Nächstgelegene Nachbarn kann auch für die Berechnung von Werten für ein stetiges Ziel verwendet werden. Hierbei wird der Durchschnitts- oder Median-Zielwert der nächstgelegenen Nachbarn verwendet, um den vorhergesagten Wert für den neuen Fall zu beziehen.

Ziel und Merkmale. Folgende Ziele und Merkmale sind möglich:

- **Nominal.** Eine Variable kann als nominal behandelt werden, wenn ihre Kategorien sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit.
- **Ordinal.** Eine Variable kann als ordinal behandelt werden, wenn ihre Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Ordinale Variablen treten beispielsweise bei Einstellungsmessungen (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf.
- **Metrisch.** Eine Variable kann als metrisch (stetig) behandelt werden, wenn ihre Werte geordnete Kategorien mit einer sinnvollen Metrik darstellen, sodass man sinnvolle Aussagen über die Abstände zwischen den Werten machen kann. Metrische Variablen sind beispielsweise Alter (in Jahren) oder Einkommen (in Geldeinheiten).

  Nominale und ordinale Variablen werden in der Nächste-Nachbarn-Analyse gleich behandelt. Bei der Prozedur wird davon ausgegangen, dass allen Variablen das richtige Messniveau zugewiesen wurde. Sie können das Messniveau für eine Variable jedoch vorübergehend ändern. Klicken Sie hierzu mit der rechten Maustaste auf die Variable in der Liste der Quellvariablen und wählen Sie das gewünschte Messniveau im Kontextmenü aus.

Messniveau und Datentyp sind durch ein Symbol neben der jeweiligen Variablen in der Variablenliste gekennzeichnet:

Messniveau	**Datentyp**			
	Numerisch	Zeichenfolge	Datum	Zeit
Metrisch (stetig)		entfällt		
Ordinal				
Nominal				

Kodierung für kategoriale Variablen. Die Prozedur kodiert vorübergehend für die Dauer des Verfahrens kategoriale Einflussvariablen und abhängige Variablen mithilfe der "Eins-aus-*c*"-Kodierung neu. Wenn es *c* Kategorien für eine Variable gibt, wird die Variable als c Vektoren gespeichert. Dabei wird die erste Kategorie als (1,0,...,0) angegeben, die zweite Kategorie als (0,1,0,...,0), ... und die letzte Kategorie als (0,0,...,0,1).

Dieses Codierungsschema steigert die Dimensionalität des Funktionsbereichs. Die Gesamtanzahl an Dimensionen ist die Anzahl an metrischen Einflussvariablen plus die Anzahl an Kategorien in allen kategorialen Einflussvariablen. Daher kann dieses Codierungsschema zu einer Verlangsamung des Trainings führen. Wenn das Training der nächstgelegenen Nachbarn sehr langsam vorangeht, können Sie versuchen, die Anzahl der Kategorien der kategorialen Einflussvariablen zu verringern, indem Sie ähnliche Kategorien zusammenfassen oder Fälle ausschließen, die extrem seltene Kategorien aufweisen, bevor Sie die Prozedur ausführen.

Jegliche "Eins-aus-*c*"-Codierung beruht auf den Trainingsdaten, selbst wenn eine Holdout-Stichprobe definiert wurde (siehe Partitionen). Wenn die Holdout-Stichprobe daher Fälle mit Einflussvariablen-Kategorien enthält, die in den Trainingsdaten nicht enthalten sind, werden diese Fälle nicht beim Scoring verwendet. Wenn die Holdout-Stichprobe Fälle mit Kategorien abhängiger Variablen enthält, die in den Trainingsdaten nicht enthalten sind, werden diese Fälle beim Scoring verwendet.

Neuskalierung. Metrische Funktionen werden standardmäßig normalisiert. Jegliche Neuskalierung beruht auf den Trainingsdaten, selbst wenn eine Holdout-Stichprobe definiert wurde (siehe Partitionen auf S. 143). Wenn Sie eine Variable zur Festlegung von Partitionen angeben, müssen diese Funktionen in der Trainings- und Holdout-Stichprobe ähnliche Verteilungen aufweisen. Verwenden Sie beispielsweise die Prozedur Explorative Datenanalyse, um die Verteilungen in den verschiedenen Partitionen zu untersuchen.

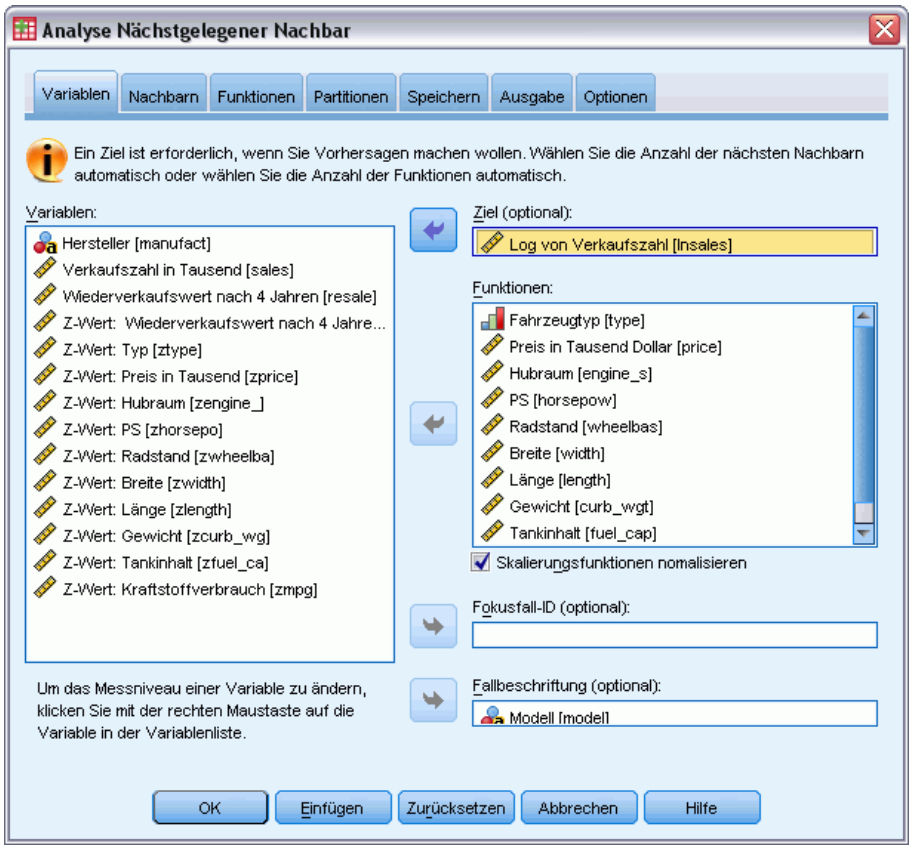
Häufigkeitsgewichtungen. Häufigkeitsgewichtungen werden von dieser Prozedur ignoriert.

Reproduzieren der Ergebnisse. Die Prozedur verwendet Zufallszahlengenerierung während der Zufallszuweisung von Partitionen und Kreuzvalidierungs-Teilstichproben. Wenn Sie Ihre Ergebnisse exakt reproduzieren möchten, müssen Sie nicht nur dieselben Einstellungen für die Prozedur, sondern auch einen Startwert für den Mersenne-Twister festlegen (siehe Partitionen auf S. 143) oder Variablen für die Definition von Partitionen und Kreuzvalidierungs-Teilstichproben verwenden.

So definieren Sie die Analyse der nächstgelegenen Nachbarn:

Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Klassifizieren > Nächstgelegener Nachbar...

Abbildung 20-2
Registerkarte "Variablen" im Dialogfeld "Analyse nächstgelegener Nachbar"



- Geben Sie ein oder zwei Funktionen an, die als unabhängige Variablen oder Einflussvariablen betrachtet werden können, falls ein Ziel vorhanden ist.

Ziel (optional). Wenn kein Ziel (abhängige Variable oder Antwort) angegeben ist, findet die Prozedur nur die *k* nächstgelegenen Nachbarn – es wird keine Klassifizierung oder Prognose vorgenommen.

Metrische Funktionen normalisieren. Normalisierungsfunktionen weisen denselben Wertebereich auf. Das kann die Leistung des Schätzungsalgorithmus verbessern. Es wird eine korrigierte Normalisierung, [2*(x−min)/(max−min)]−1, angewendet. Korrigierte, normalisierte Werte liegen im Bereich zwischen −1 und 1.

Fokusfall-ID (optional). Mit dieser Option können Sie Fälle von besonderem Interesse markieren. Zum Beispiel möchte ein Forscher bestimmen, welche Testergebnisse aus einem Schulbezirk – der Fokusfall – vergleichbar sind mit denen aus ähnlichen Schulbezirken. Er verwendet die Analyse nächstgelegener Nachbar, um die Schulbezirke zu finden, die sich hinsichtlich einer festgelegten Menge an Merkmalen am ähnlichsten sind. Anschließend vergleicht er die Testergebnisse des untersuchten Schulbezirks mit jenen der nächstgelegenen Nachbarn.

Fokusfälle können auch in klinischen Studien für die Auswahl von Vergleichsfällen verwendet werden, die den klinischen Fällen ähnlich sind. Die Fokusfälle werden in der Tabelle der *k* nächstgelegenen Nachbarn und Abstände, im Funktionsbereichsdiagramm, im Peer-Diagramm und in der Quadrantenkarte dargestellt. Informationen zu Fokusfällen werden in den Dateien gespeichert, die auf der Registerkarte "Ausgabe" angegeben sind.

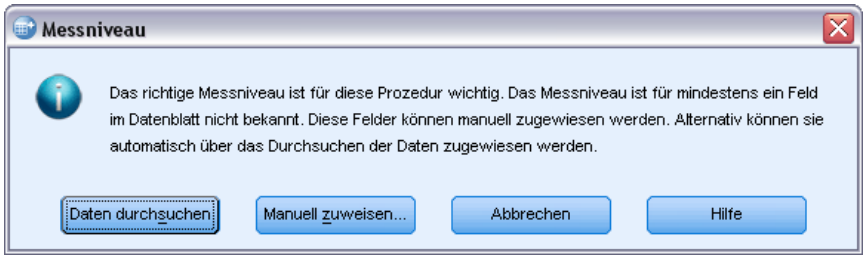
Fälle mit einem positiven Wert für die angegebene Variable werden als Fokusfälle behandelt. Variablen ohne positive Werte können nicht angegeben werden.

Fallbeschriftung (optional). Fälle werden im Funktionsbereichsdiagramm, im Peer-Diagramm und in der Quadrantenkarte mit diesen Werten beschriftet.

Felder mit unbekanntem Messniveau

Die Messniveau-Warnmeldung wird angezeigt, wenn das Messniveau für mindestens eine Variable (ein Feld) im Datenblatt unbekannt ist. Da sich das Messniveau auf die Berechnung der Ergebnisse für diese Prozedur auswirkt, müssen alle Variablen ein definiertes Messniveau aufweisen.

Abbildung 20-3
Messniveau-Warnmeldung

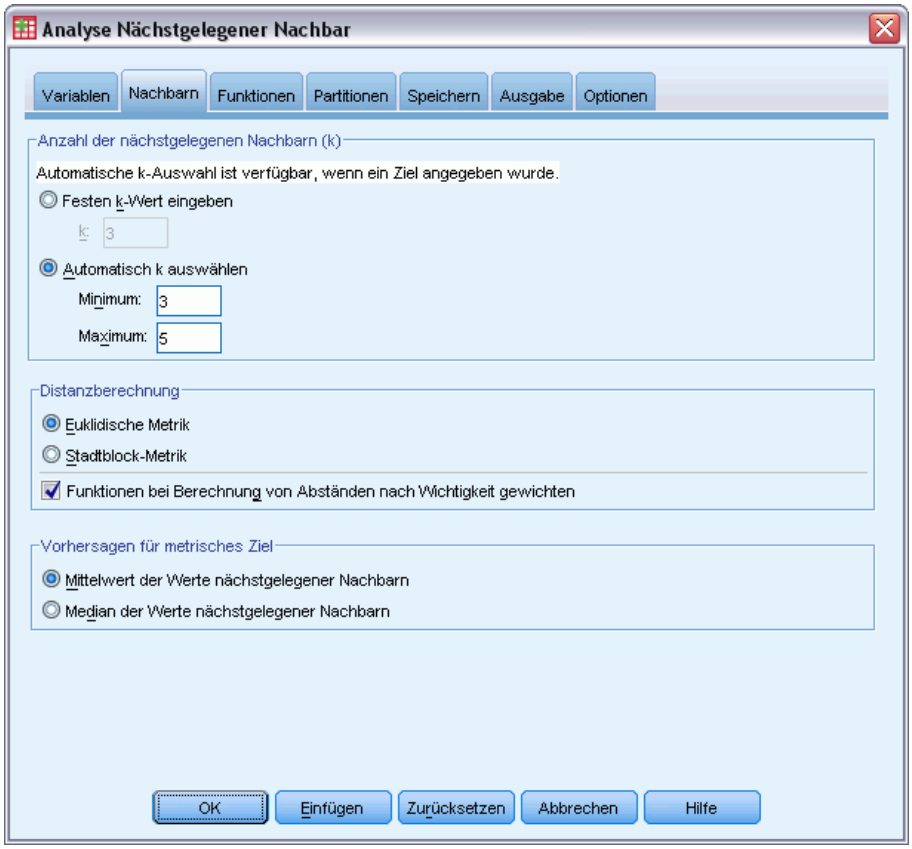


- **Daten durchsuchen.** Liest die Daten im aktiven Datenblatt (Arbeitsdatei) und weist allen Feldern, deren Messniveau zurzeit nicht bekannt ist, das Standardmessniveau zu. Bei großen Datenblättern kann dieser Vorgang einige Zeit in Anspruch nehmen.
- **Manuell zuweisen.** Öffnet ein Dialogfeld, in dem alle Felder mit unbekanntem Messniveau aufgeführt werden. Mit diesem Dialogfeld können Sie diesen Feldern ein Messniveau zuweisen. Außerdem können Sie in der Variablenansicht des Daten-Editors ein Messniveau zuweisen.

Da das Messniveau für diese Prozedur bedeutsam ist, können Sie erst dann auf das Dialogfeld zur Ausführung dieser Prozedur zugreifen, wenn für alle Felder ein Messniveau definiert wurde.

Nachbarn

Abbildung 20-4
Registerkarte "Nachbarn" im Dialogfeld "Analyse nächstgelegener Nachbar"



Anzahl der nächstgelegenen Nachbarn (k). Geben Sie die Anzahl der nächstgelegenen Nachbarn an. Beachten Sie dabei, dass eine höhere Anzahl an Nachbarn nicht unbedingt ein präziseres Modell hervorbringt.

Wenn ein Ziel auf der Registerkarte "Variablen" angegeben wurde, können Sie alternativ einen Wertebereich angeben und die Prozedur die "beste" Anzahl an Nachbarn in diesem Bereich ermitteln lassen. Wie die Anzahl an nächstgelegenen Nachbarn bestimmt wird, hängt davon ab, ob auf der Registerkarte "Funktionen" die Funktionsauswahl angegeben wurde.

- Wenn die Funktionsauswahl aktiviert wurde, wird für jeden Wert von *k* im angegebenen Bereich eine Funktionsauswahl durchgeführt und *k* und die zugehörige Funktionsgruppe mit der niedrigsten Fehlerrate (oder dem geringsten Quadratsummen-Fehler, falls das Ziel metrisch ist) werden ausgewählt.
- Wenn die Funktionsauswahl nicht aktiviert ist, wird eine *V*-fache Kreuzvalidierung angewendet, um die "beste" Anzahl an Nachbarn zu ermitteln. Informationen zur Zuweisung von Aufteilungen finden Sie unter der Registerkarte "Partition".

Distanzberechnung. Mit diesem Wert wird das Längenmaßsystem für die Messung der Ähnlichkeit von Fällen festgelegt.

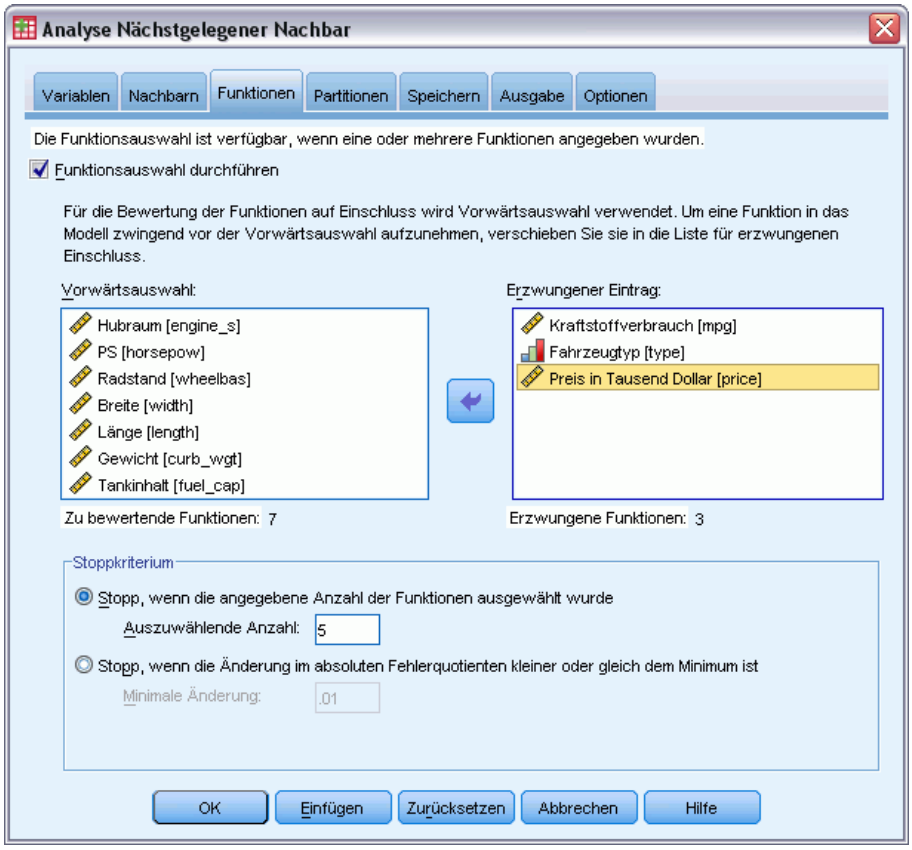
- **Euklidisch.** Der Abstand zwischen zwei Fällen, x und y, ergibt sich aus der Quadratwurzel der Summe, über alle Dimensionen, der quadrierten Differenzen zwischen den Werten für die Fälle.
- **Stadtblock.** Die Distanz zwischen zwei Fällen ergibt sich aus der Summe, über alle Dimensionen, der absoluten Differenzen zwischen den Werten der Fälle. Dies wird auch als Manhattan-Distanz bezeichnet.

Wenn auf der Registerkarte "Variablen" ein Ziel angegeben wurde, können Sie die Funktionen bei der Berechnung der Distanzen auch mit der normalisierten Wichtigkeit gewichten. Die Wichtigkeit der Funktionen für eine Einflussvariable ergibt sich aus dem Verhältnis der Fehlerrate oder dem Quadratsummenfehler des Modells, wobei die Einflussvariable bis zum Quadratsummenfehler für das gesamte Modell vom Modell entfernt wird. Die normalisierte Wichtigkeit wird durch die Neugewichtung der Werte der Funktionswichtigkeit berechnet, so dass deren Summe 1 ergibt.

Vorhersagen für das metrische Ziel. Wenn auf der Registerkarte "Variablen" ein metrisches Ziel angegeben ist, legt dieser Wert fest, ob der Vorhersagewert basierend auf dem Mittelwert oder dem Median der nächstgelegenen Nachbarn berechnet wird.

Funktionen

Abbildung 20-5
Registerkarte "Funktionen" im Dialogfeld "Analyse nächstgelegener Nachbar"



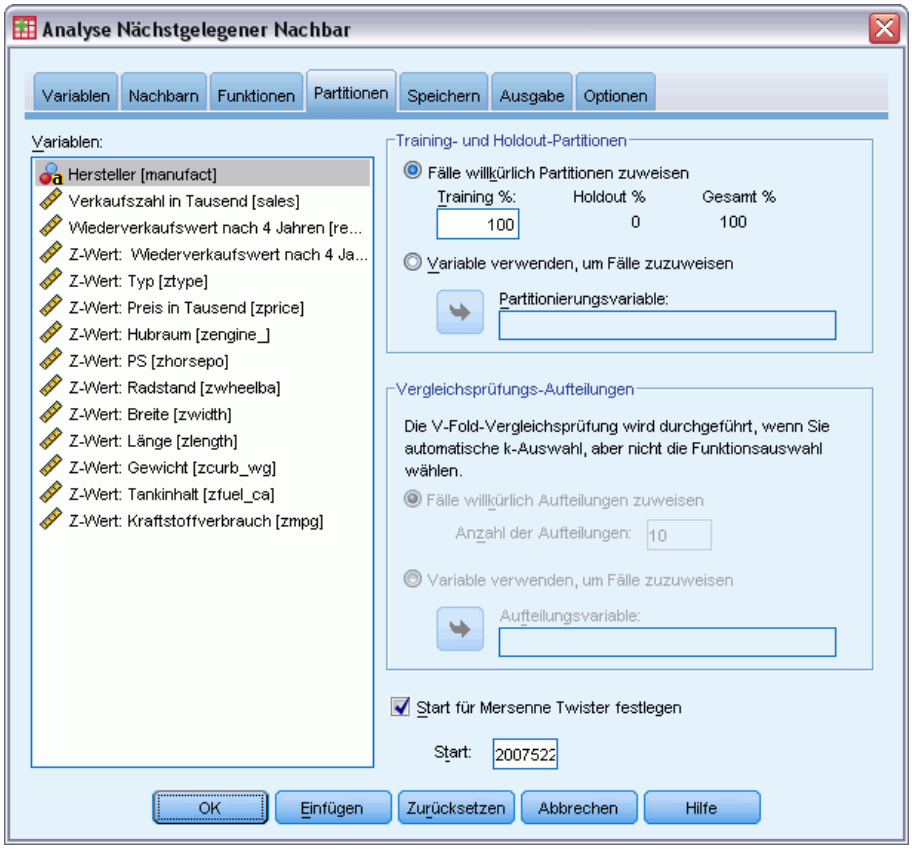
Auf der Registerkarte “Funktionen” können Sie Optionen für die Funktionsauswahl angeben, wenn auf der Registerkarte “Variablen” ein Ziel angegeben ist. Standardmäßig werden bei der Funktionsauswahl alle Funktionen berücksichtigt, Sie können optional aber auch eine Untergruppe von Funktionen auswählen, die in das Modell aufgenommen werden sollen.

Stoppkriterien. Bei jedem Schritt wird die Funktion, deren Integration in das Modell den geringsten Fehler hervorruft (für kategoriale Ziele als Fehlerrate und für metrische Ziele als Quadratsummenfehler berechnet), für die Integration in das Modell in Betracht gezogen. Die Vorwärtsselektion wird fortgesetzt, bis die angegebene Bedingung erfüllt wird.

- **Feste Anzahl an Funktionen.** Der Algorithmus fügt neben den erzwungenen Funktionen eine feste Anzahl an Funktionen in das Modell ein. Geben Sie eine positive Ganzzahl ein. Eine geringere Anzahl an Werten führt zu einem sparsameren Modell. Dabei läuft man allerdings Gefahr, wichtige Funktionen zu vernachlässigen. Bei einer höheren Anzahl an Werten werden alle wichtigen Funktionen erfasst, dafür läuft man aber Gefahr, Funktionen einzufügen, die den Modellfehler erhöhen.
- **Minimale Änderung im absoluten Fehlerquotienten.** Der Algorithmus wird beendet, wenn die Änderung im absoluten Fehlerquotienten vermuten lässt, dass das Modell durch Hinzufügen weiterer Funktionen nicht mehr weiter optimiert werden kann. Geben Sie eine positive Zahl an. Bei einem geringeren Wert für die minimale Änderung werden in der Regel mehr Funktionen aufgenommen. Dabei können allerdings auch Funktionen aufgenommen werden, die das Modell nicht wesentlich verbessern. Bein einem höheren Wert für die minimale Änderungen werden mehr Funktionen ausgeschlossen, was dazu führen kann, dass Funktionen ausgeschlossen werden, die wichtig für das Modell wären. Der “optimale” Wert für die minimale Änderung hängt von den jeweiligen Daten und dem Anwendungsbereich ab. Informationen dazu, wie Sie beurteilen, welche Funktionen am wichtigsten sind, finden Sie im Protokoll über die Funktionsauswahlfehler in der Ausgabe. Für weitere Informationen siehe Thema Funktionsauswahl-Fehlerprotokoll auf S. 156.

Partitionen

Abbildung 20-6
Registerkarte "Partitionen" im Dialogfeld "Analyse nächstgelegener Nachbar"



Auf der Registerkarte "Partitionen" können Sie das Daten-Set in Trainings- und Holdout-Sets unterteilen und gegebenenfalls Vergleichsprüfungs-Aufteilungen Fälle zuweisen.

Training- und Holdout-Partitionen. Diese Gruppe gibt die Methode zur Partitionierung der Arbeitsdatei in eine Trainings- und eine Holdout-Stichprobe an. Die **Trainingsstichprobe** umfasst die Datensätze, die zum Trainieren des Modells der nächstgelegenen Nachbarn verwendet wurden; ein gewisser Prozentsatz der Fälle im Daten-Set muss der Trainingsstichprobe zugewiesen werden, um ein Modell zu erhalten. Die **Holdout-Stichprobe** ist ein unabhängiger Satz von Datensätzen, der zur Bewertung des endgültigen Modells verwendet wird; der Fehler für die Houldout-Stichprobe bietet eine "ehrliche" Schätzung der Vorhersagekraft des Modells, da die Prüffälle (die Fälle in der Holdout-Stichprobe) nicht zur Erstellung des Modells verwendet wurden.

- **Fälle willkürlich Partitionen zuweisen.** Legen Sie den Prozentsatz der Fälle fest, die der Trainingsstichprobe zugewiesen werden sollen. Die übrigen Fälle werden der Holdout-Stichprobe zugewiesen.
- **Variable zum Zuweisen von Fällen verwenden.** Geben Sie eine numerische Variable an, die jeden Fall in der Arbeitsdatei der Trainings- bzw. Holdout-Stichprobe zuweist. Fälle mit einem positiven Wert für die Variable werden der Trainingsstichprobe zugewiesen, Fälle mit dem Wert 0 und einem negativen Wert der Holdout-Stichprobe. Fälle mit einem systemdefiniert

fehlenden Wert werden aus der Analyse ausgeschlossen. Alle benutzerdefiniert fehlenden Werte für die Partitionsvariable werden immer als gültig behandelt.

Vergleichsprüfungs-Aufteilungen Um die "beste" Anzahl an Nachbarn zu ermitteln wird eine *V*-fache Vergleichsprüfung durchgeführt. Bei Funktionsauswahl ist sie aus Leistungsgründen nicht verfügbar.

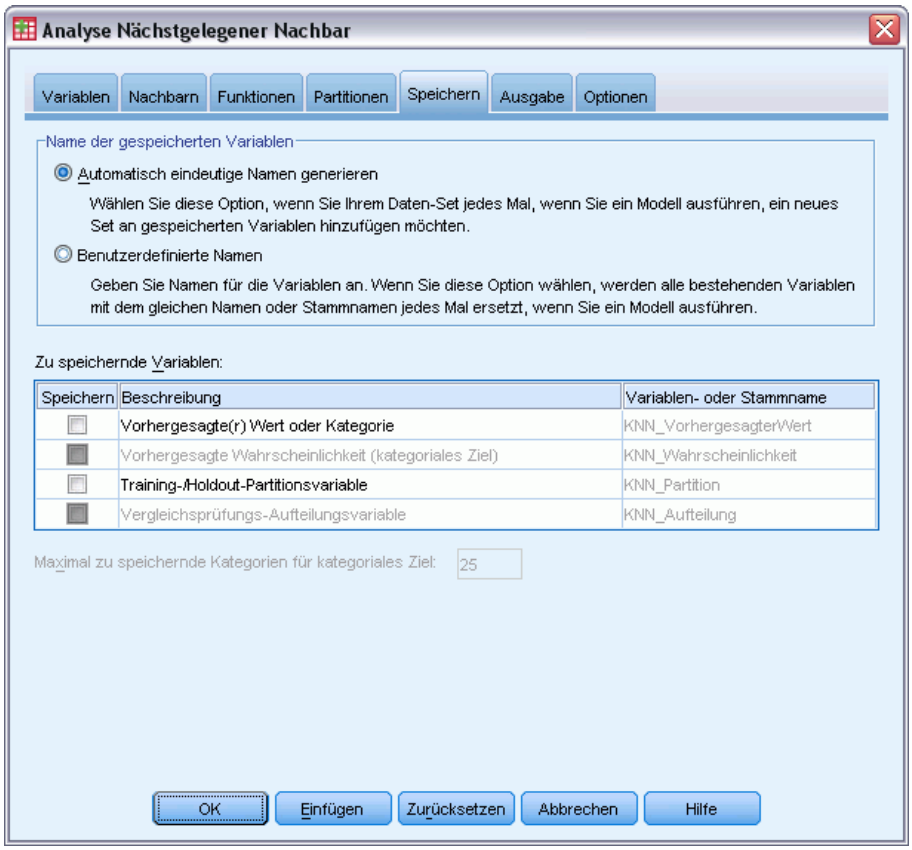
Bei der Vergleichsprüfung wird die Stichprobe in mehrere Teilstichproben oder Aufteilungen gegliedert. Anschließend werden Nächste-Nachbarn-Modelle erzeugt; dabei werden nacheinander die Daten der einzelnen Stichproben ausgeschlossen. Das erste Modell beruht auf allen Fällen mit Ausnahme der Fälle in der ersten Stichprobenaufteilung, das zweite Modell auf allen Fällen mit Ausnahme der Fälle in der zweiten Stichprobenaufteilung usw. Bei jedem Modell wird jeweils der Fehler geschätzt. Hierzu wird das Modell auf die Teilstichprobe angewendet, die beim Erstellen des Modells ausgeschlossen war. Die "beste" Anzahl an nächstgelegenen Nachbarn ist die Anzahl, die die wenigsten Fehler für alle Aufteilungen erzeugt.

- **Aufteilungen willkürlich Fälle zuweisen.** Geben Sie die Anzahl an Aufteilungen an, die für die Vergleichsprüfung herangezogen werden sollen. Die Prozedur weist Fälle willkürlich Aufteilungen zu und nummeriert sie von 1 bis *V*, die Anzahl an Aufteilungen.
- **Variable zum Zuweisen von Fällen verwenden.** Geben Sie eine numerische Variable an, die jeden Fall in der Arbeitsdatei einer Aufteilung zuweist. Die Variable muss numerisch sein und Werte von 1 bis *V* annehmen. Wenn Werte in diesem Bereich und bei aufgeteilten Dateien in Aufteilungen fehlen, ruft das Fehler hervor.

Startwert für Mersenne-Twister festlegen. Wenn Sie einen Startwert festlegen, können Sie Analysen reproduzieren. Die Verwendung dieses Steuerelements gleicht der Festlegung eines Mersenne-Twisters als aktivem Generator und eines festen Startpunkts für das Dialogfeld "Zufallszahlengeneratoren", mit dem wichtigen Unterschied, dass die Festlegung des Startpunkts in diesem Dialogfeld den aktuellen Status des Zufallszahlengenerators beibehält und diesen Status nach Abschluss der Analyse wiederherstellt.

Speichern

Abbildung 20-7
Registerkarte "Speichern" im Dialogfeld "Analyse nächstgelegener Nachbar"



Namen der gespeicherten Variablen. Durch eine automatische Generierung von Namen wird sichergestellt, dass Ihre Arbeit nicht verloren geht. Mit benutzerdefinierten Namen können Sie Ergebnisse aus früheren Durchgängen verwerfen/ersetzen, ohne zuerst die gespeicherten Variablen im Daten-Editor löschen zu müssen.

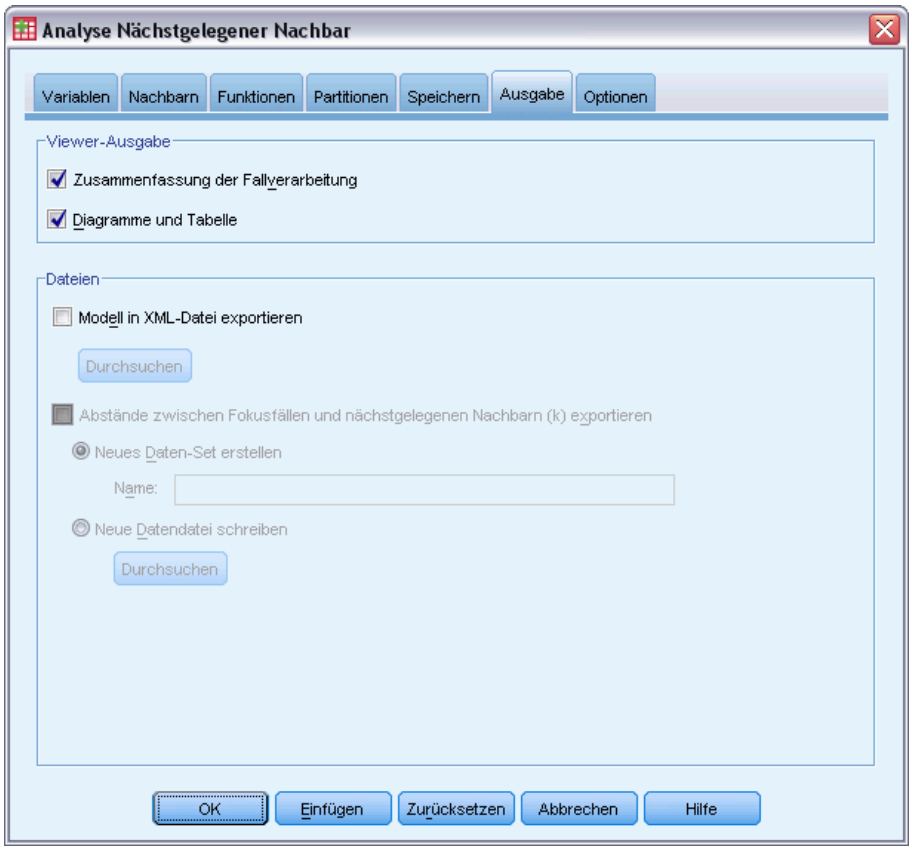
Zu speichernde Variablen

- **Vorhergesagte(r) Wert oder Kategorie.** Damit wird bei metrischen Zielen der vorhergesagte Wert und bei kategorialen Zielen die vorhergesagte Kategorie gespeichert.
- **Vorhergesagte Wahrscheinlichkeit.** Damit werden bei kategorialen Zielen die vorhergesagten Wahrscheinlichkeiten gespeichert. Für die ersten *n* Kategorien wird eine separate Variable gespeichert. Dabei wird *n* im Steuerelement Maximale Anzahl der zu speichernden Kategorien für kategoriale Ziele angegeben.

- **Trainings-/Holdout-Partitionsvariablen.** Wenn Fälle den Trainings- und Holdout-Stichproben auf der Registerkarte "Partitionen" willkürlich zugewiesen werden, wird mit dieser Einstellung der Wert der Partition (Training oder Holdout) gespeichert, der der Fall zugewiesen wurde.
- **Vergleichsprüfungs-Aufteilungsvariable.** Wenn Fälle auf der Registerkarte "Partitionen" Vergleichsprüfungs-Aufteilungen willkürlich zugewiesen werden, wird mit dieser Einstellung der Wert der Aufteilung gespeichert, der dieser Fall zugewiesen wurde.

Ausgabe

Abbildung 20-8
Registerkarte "Ausgabe" im Dialogfeld "Analyse nächstgelegener Nachbar"



Viewer-Ausgabe

- **Zusammenfassung der Fallverarbeitung.** Zeigt die Tabelle mit der Zusammenfassung der Fallverarbeitung an, die die Anzahl der in der Analyse ein- und ausgeschlossenen Fälle zusammenfasst (insgesamt und nach Trainings- und Holdout-Stichprobe geordnet).
- **Diagramme und Tabellen.** Enthält modellbezogene Ausgaben einschließlich Tabellen und Diagrammen. Die Tabellen in der Modellansicht enthalten *k* nächstgelegene Nachbarn und die Abstände für Fokusfälle, eine Klassifizierung der kategorialen Antwortvariablen und eine Zusammenfassung der Fehler. Die grafische Ausgabe in der Modellansicht enthält ein Auswahlfehler-Protokoll, ein Wichtigkeitsdiagramm für die Funktionen, ein

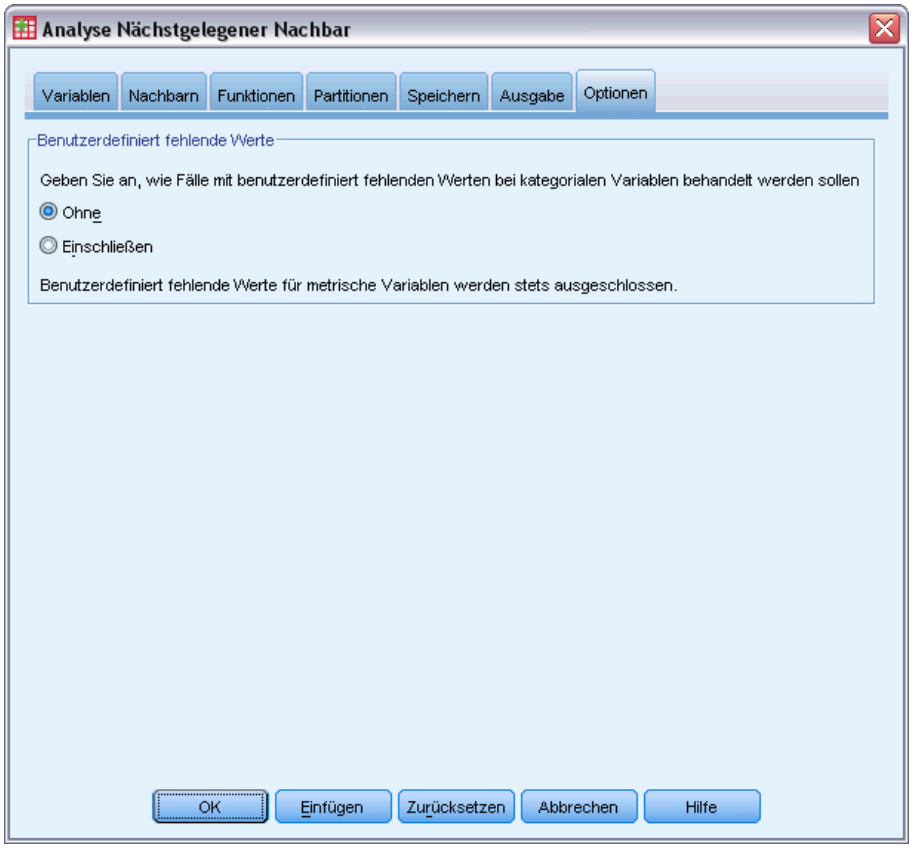
Funktionsbereichsdiagramm, ein Peer-Diagramm und eine Quadrantenkarte. Für weitere Informationen siehe Thema Modellansicht auf S. 148.

Dateien

- **Modell als XML exportieren.** Anhand dieser Modelldatei können Sie die Modellinformationen zu Bewertungszwecken auf andere Datendateien anwenden. Diese Option ist nicht verfügbar, wenn aufgeteilte Dateien definiert wurden.
- **Abstände zwischen Fokusfällen und k nächstgelegenen Nachbarn exportieren.** Für jeden Fokusfall wird eine separate Variable für jeden der *k* nächstgelegenen Nachbarn (aus der Trainingsstichprobe) und die entsprechenden *k* nächstgelegenen Abstände erzeugt.

Optionen

Abbildung 20-9
Registerkarte "Optionen" im Dialogfeld "Analyse nächstgelegener Nachbar"

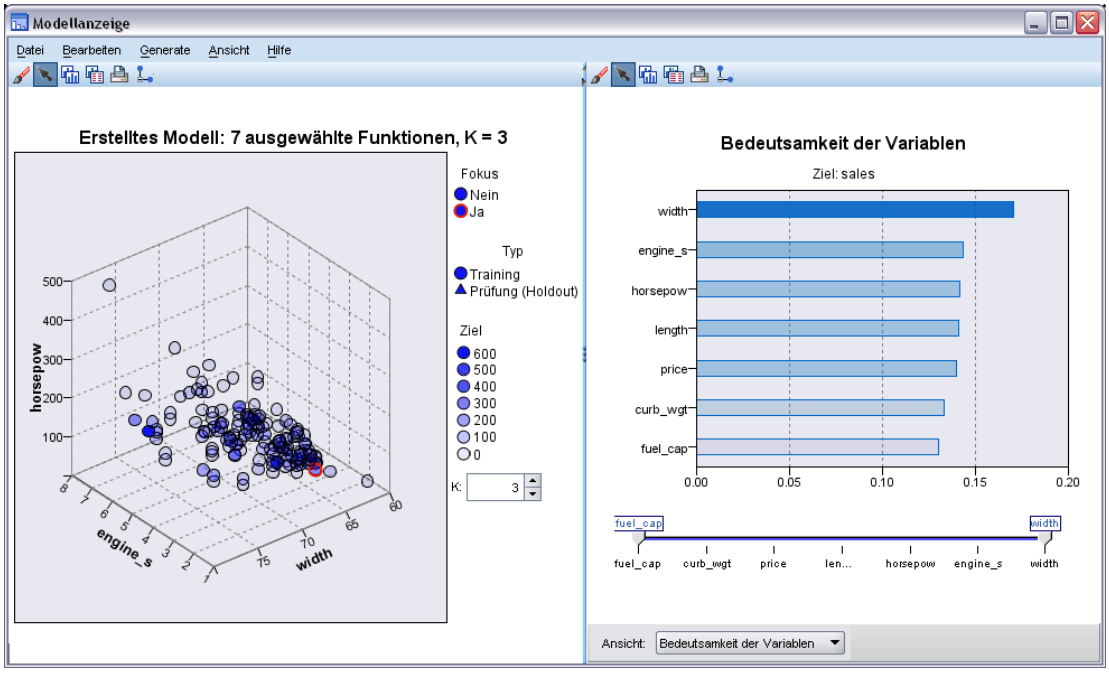


Benutzerdefinierte fehlende Werte. Kategoriale Variablen müssen gültige Werte für einen Fall aufweisen, um in die Analyse aufgenommen zu werden. Mit diesen Steuerungen legen Sie fest, ob benutzerdefiniert fehlende Werte bei den kategorialen Variablen als gültige Werte behandelt werden sollen.

Systemdefinierte fehlende Werte und fehlende Werte für methrische Variablen werden immer als ungültige Werte behandelt.

Modellansicht

Abbildung 20-10
Modellansicht für die Analyse nächstgelegener Nachbar



Wenn Sie auf der Registerkarte “Ausgabe” die Option Diagramme und Tabellen wählen, erstellt die Prozedur ein Nächster-Nachbar-Modell-Objekt im Viewer. Wenn Sie dieses Objekt durch einen Doppelklick aktivieren, erhalten Sie eine interaktive Ansicht des Modells. Das Fenster der Modellansicht setzt sich aus zwei Bereichen zusammen:

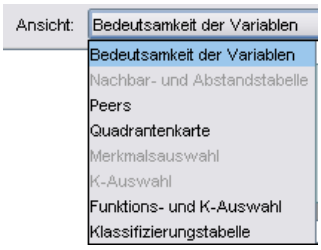
- Im ersten Bereich wird eine Übersicht des Modells, die sogenannte Hauptansicht, angezeigt.
- Im zweiten Bereich wird eine der beiden folgenden Ansichten angezeigt:

  Die Hilfsmodellansicht enthält mehr Informationen zum Modell, ist dafür aber weniger stark auf das Modell an sich konzentriert.

  Die verknüpfte Ansicht zeigt Details zu einer bestimmten Funktion des Modells an, wenn der Benutzer einen Teil der Hauptansicht ansteuert.

Standardmäßig wird im ersten Bereich der Funktionsbereich und im zweiten Bereich das Wichtigkeitsdiagramm der Variablen angezeigt. Wenn das Wichtigkeitsdiagramm der Variablen nicht verfügbar ist, d. h. wenn auf der Registerkarte “Funktionen” nicht die Option Funktionen nach Wichtigkeit gewichten ausgewählt wurde, wird im ersten Bereich die Dropdown-Liste “Ansicht” angezeigt.

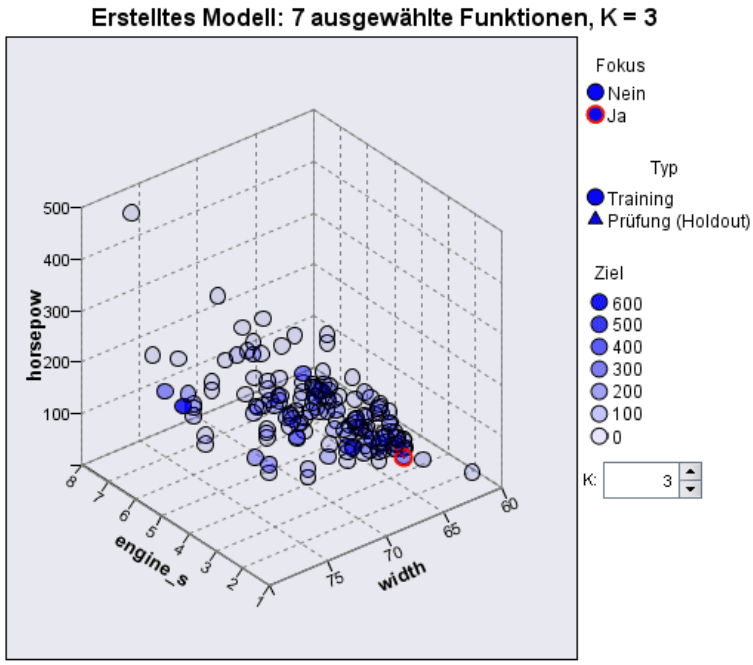
Abbildung 20-11
Dropdown-Liste "Modellansicht" im Dialogfeld "Analyse nächstgelegener Nachbar"



Wenn für eine Ansicht keine Informationen zur Verfügung stehen, ist der zugehörige Text in der Dropdown-Liste "Ansicht" deaktiviert.

Funktionsbereich

Abbildung 20-12
Funktionsbereich



Das Funktionsbereichsdiagramm ist ein interaktives Diagramm für den Funktionsbereich (bzw. -unterbereich, bei mehr als drei Funktionen). Jede Achse stellt eine Funktion im Modell dar und die Position der Punkte in der Tabelle gibt die Werte dieser Funktionen für Fälle in den Trainings- und Holdout-Partitionen an.

Erläuterungen Neben den Funktionswerten liefern die Punkte im Diagramm weitere Informationen.

- Die Form gibt die Partition an, zu der ein Punkt gehört (Training oder Holdout).

- Die Farbe/Schattierung eines Punkts gibt den Wert des Ziels für diesen Fall an. Dabei entsprechen eindeutige Farbwerte den Kategorien eines kategorialen Ziels und Schattierungen dem Wertebereich eines stetigen Ziels. Für Trainings-Partitionen ist der angegebene Wert der festgestellte Wert. Für Holdout-Partitionen handelt es sich um den vorhergesagten Wert. Wenn kein Ziel angegeben ist, wird diese Erläuterung nicht angezeigt.
- Kräftigere Umrisse weisen auf Fokusfälle hin. Fokusfälle werden im Zusammenhang mit ihren *k* nächstgelegenen Nachbarn angezeigt.

Steuerelemente und Interaktivität. Sie können den Funktionsbereich mit einer Reihe an Steuerelementen im Diagramm untersuchen.

- Sie können festlegen, welche Untermenge an Funktionen im Diagramm angezeigt werden soll, und ändern, welche Funktionen in den Dimensionen dargestellt werden.
- Fokusfälle sind Punkte, die im Funktionsbereichsdiagramm ausgewählt wurden. Wenn Sie eine Fokusfallvariable angegeben haben, werden zuerst die Punkte ausgewählt, die die Fokusfälle darstellen. Es kann jedoch jeder Punkt vorübergehend ein Fokusfall werden, wenn Sie ihn auswählen. Die gängigen Steuerelemente für Punkte sind verfügbar: Wenn Sie auf einen Punkt klicken, wird dieser Punkt ausgewählt und die Auswahl aller anderen Punkte aufgehoben. Wenn Sie die Strg-Taste drücken und auf einen Punkt klicken, wird er der Menge an gewählten Punkten hinzugefügt. Verknüpfte Ansichten wie das Peers-Diagramm werden automatisch mit den Fällen aktualisiert, die im Funktionsbereich ausgewählt werden.
- Sie können die Anzahl an für Fokusfälle anzuzeigenden nächstgelegenen Nachbarn (*k*) ändern.
- Wenn Sie die Maus über einen Punkt im Diagramm bewegen, wird eine QuickInfo mit dem Wert der Fallbeschriftung oder, wenn keine Fallbeschriftungen definiert sind, der Fallnummer und dem festgestellten und vorhergesagten Zielwert angezeigt.
- Sie können den Funktionsbereich über die Schaltfläche “Zurücksetzen” wieder in seinen Originalzustand versetzen.

Hinzufügen und Entfernen von Feldern/Variablen

Sie können dem Funktionsbereich neue Felder/Variablen hinzufügen oder akuell angezeigte Felder/Variablen entfernen.

Variablenpalette

Abbildung 20-13
Variablenpalette

Die Variablenpalette muss angezeigt werden, bevor Sie Variablen hinzufügen und entfernen können. Um die Variablenpalette anzuzeigen, muss sich die Modellanzeige im Bearbeitungsmodus befinden und im Funktionsbereich muss ein Fall ausgewählt sein.

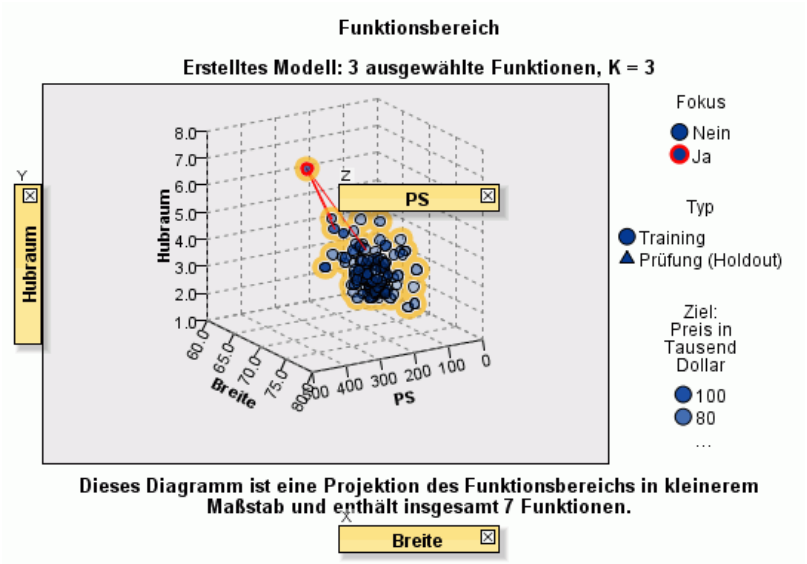
- Um die Modellanzeige in den Barbeitungsmodus zu versetzen, wählen Sie die folgenden Menübefehle aus:
  Ansicht > Bearbeitungsmodus
- Klicken Sie im Bearbeitungsmodus auf einen beliebigen Fall im Funktionsbereich.
- Zum Anzeigen der Variablenpalette wählen Sie die folgenden Menübefehle aus:
  Ansicht > Paletten > Variablen

  In der Variablenpalette sind alle Variablen im Funktionsbereich aufgeführt. Das Symbol neben dem Variablennamen zeigt das Messniveau der Variablen an.
- Um das Messniveau einer Variablen vorübergehend zu ändern, klicken Sie in der Variablenpalette mit der rechten Maustaste auf die Variable und wählen eine Option.

Variablenzonen

Variablen werden im Funktionsbereich zu “Zonen” hinzugefügt. Um die Zonen anzuzeigen, ziehen Sie eine Variable aus der Variablenpalette oder wählen Zonen anzeigen.

Abbildung 20-14
Variablenzonen



Der Funktionsbereiche hat Zonen für die *x*-, die *y*- und die *z*-Achse.

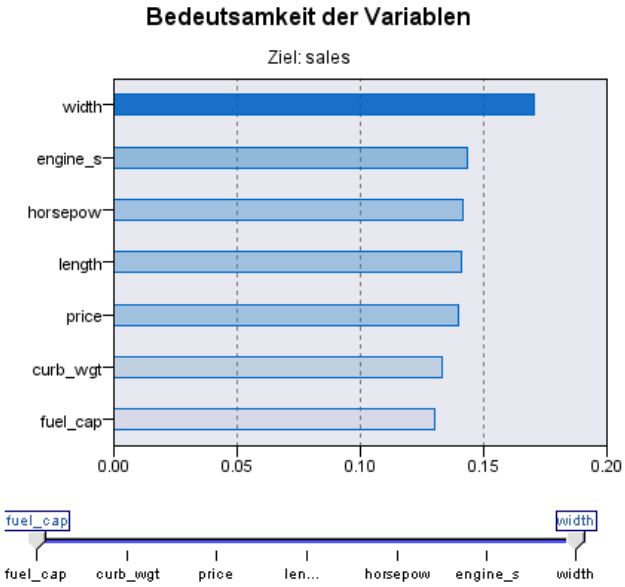
Variablen in Zonen verschieben

Allgemeine Regeln und Tipps zum Verschieben von Variablen in Zonen:

- Um eine Variabele in eine Zone zu verschieben, klicken Sie auf die Variable und ziehen Sie sie aus der Variablenpalette in die Zone. Wenn Sie Zonen anzeigen wählen, können Sie auch mit der rechten Maustaste auf eine Zone klicken und eine Variable auswählen, die Sie dieser Zone hinzufügen möchten.
- Wenn Sie eine Variable aus der Variablenpalette in eine Zone ziehen, in der sich bereits eine andere Variable befindet, wird die alte Variable durch die neue ersetzt.
- Wenn Sie eine Variable aus einer Zone in eine andere ziehen, in der sich bereits eine andere Variable befindet, werden die beiden Variablen vertauscht.
- Wenn Sie in einer Zone auf "X" klicken, wird die Variable aus dieser Zone entfernt.
- Falls sich in der Visualisierung mehrere Grafikelemente befinden, kann jedes Grafikelement über eigene Variablenzonen verfügen. Wählen Sie zuerst das gewünschte Grafikelement aus.

Variablenwichtigkeit

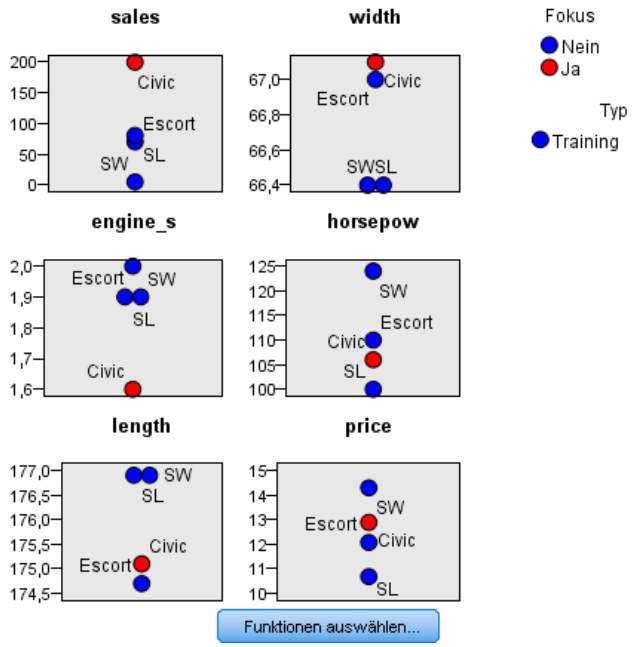
Abbildung 20-15
Variablenwichtigkeit



In der Regel konzentriert man sich bei der Modellerstellung auf die Variablen, die am wichtigsten sind, und vernachlässigt jene, die weniger wichtig sind. Dabei unterstützt Sie das Wichtigkeitsdiagramm der Variablen, da es die relative Wichtigkeit der einzelnen Variablen für das Modell angibt. Da die Werte relativ sind, beträgt die Summe der Werte aller Variablen im Diagramm 1,0. Die Variablenwichtigkeit bezieht sich nicht auf die Genauigkeit des Modells. Sie bezieht sich lediglich auf die Wichtigkeit der einzelnen Variablen für eine Vorhersage und nicht auf die Genauigkeit der Vorhersage.

Gruppen

Abbildung 20-16
Peers-Diagramm



Dieses Diagramm enthält die Fokusfälle und ihre *k* nächstgelegenen Nachbarn für jede Funktion im Ziel. Es ist verfügbar, wenn ein Fokusfall im Funktionsbereich ausgewählt ist.

Verknüpfungsverhalten. Das Peers-Diagramm ist auf zwei Arten mit dem Funktionsbereich verknüpft.

- Im Peers-Diagramm werden die im Funktionsbereich gewählten Fokusfälle sowie ihre *k* nächstgelegenen Nachbarn angezeigt.
- Der Wert *k* wird im Funktionsbereich gewählt und im Peers-Diagramm herangezogen.

Abstände zwischen nächstgelegenen Nachbarn

Abbildung 20-17
Abstände zwischen nächstgelegenen Nachbarn

Fokusfall	Nächstgelegene Nachbarn			Nächstgelegene Abstände		
	1	2	3	1	2	3
Civic	SL	Escort	SW	0,053	0,059	0,064

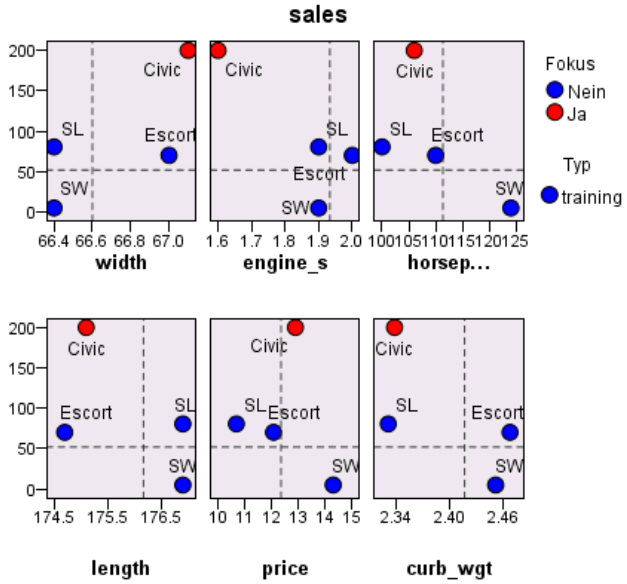
Diese Tabelle zeigt nur die *k* nächstgelegenen Nachbarn und Abstände für Fokusfälle an. Sie ist verfügbar, wenn eine Fokusfall-ID auf der Registerkarte “Variable” angegeben ist, und zeigt nur Fokusfälle an, die mit dieser Variablen angegeben werden.

Jede Zeile der:

- Spalte Fokusfall enthält den Wert der Fallbeschriftungsvariablen für den Fokusfall. Wenn keine Fallbeschriftungen angegeben wurden, enthält diese Spalte die Fallnummer des Fokusfalls.
- Die *i*. Spalte unter der Gruppe der nächstgelegenen Nachbarn enthält den Wert der Fallbeschriftungsvariablen für den *i*. nächsten Nachbarn des Fokusfalls. Wenn keine Fallbeschriftungen definiert wurden, enthält diese Spalte die Fallnummer des *i*. nächstgelegenen Nachbarn des Fokusfalls.
- Die *i*. Spalte unter der Gruppe der kürzesten Abstände enthält den Abstand des *i*. nächstgelegenen Nachbarn zum Fokusfall.

Quadrantenkarte

Abbildung 20-18
Quadrantenkarte

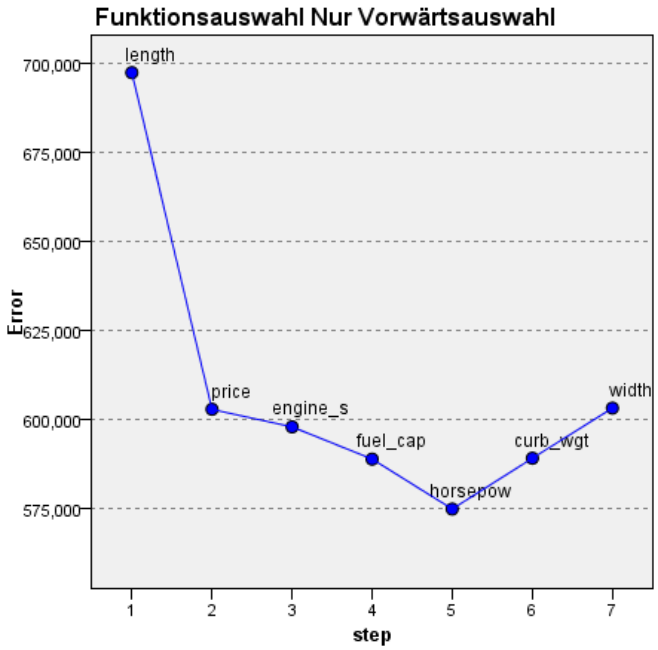


Dieses Diagramm zeigt die Fokusfälle und ihre *k* nächstgelegenen Nachbarn als Streudiagramm (oder Punktdiagramm, je nach Messniveau des Ziels) mit dem Ziel auf der *y*-Achse und einer metrischen Funktion auf der *x*-Achse nach Funktionen in einzelne Felder unterteilt an. Es ist verfügbar, wenn ein Ziel vorhanden und ein Fokusfall im Funktionsbereich ausgewählt ist.

- Für stetige Variablen werden bei den Mittelwerten der Variablen in der Trainingspartition Referenzlinien gezogen.

Funktionsauswahl-Fehlerprotokoll

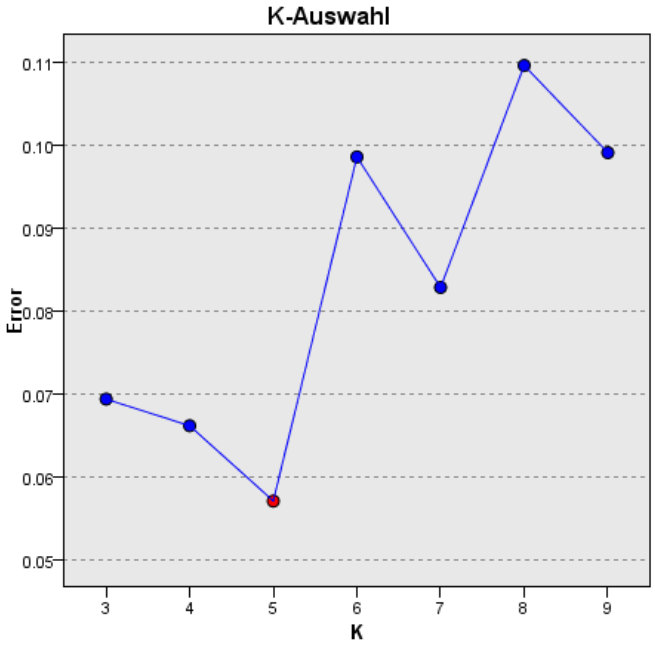
Abbildung 20-19
Funktionsauswahl



Punkte im Diagramm zeigen den Fehler (je nach Messniveau des Ziels entweder die Fehlerrate oder den Quadratsummenfehler) auf der *y*-Achse für das Modell mit der Funktion auf der *x*-Achse an (plus allen Funktionen weiter links auf der *x*-Achse). Dieses Diagramm ist verfügbar, wenn ein Ziel und eine Funktionsauswahl aktiviert sind.

k-Auswahl-Fehlerprotokoll

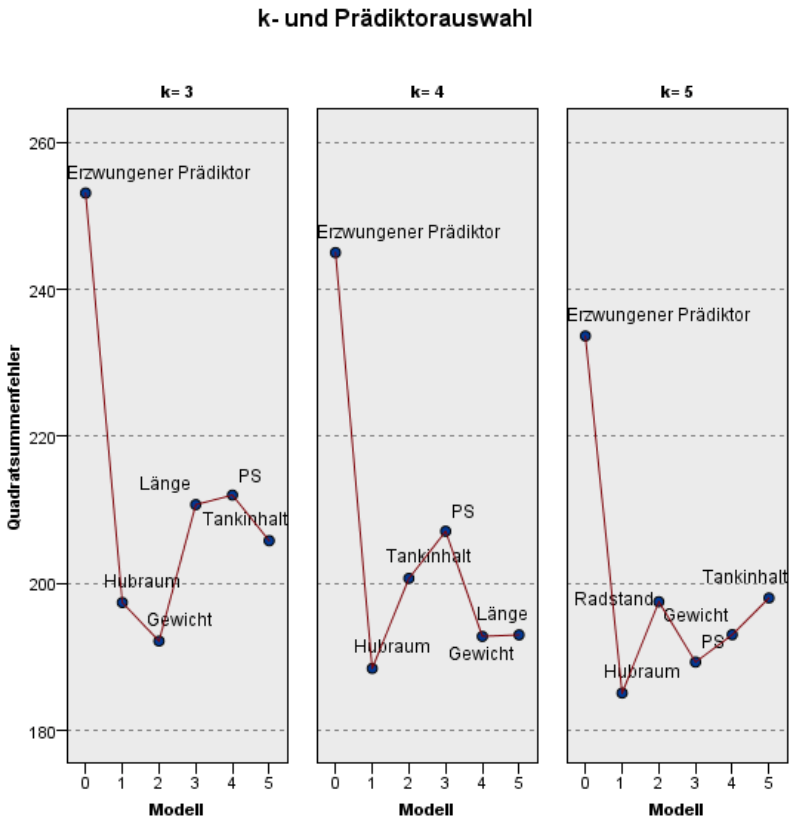
Abbildung 20-20
k-Auswahl



Punkte im Diagramm zeigen den Fehler (je nach Messniveau des Ziels entweder die Fehlerrate oder den Quadratsummenfehler) auf der *y*-Achse für das Modell mit der Anzahl an nächstgelegenen Nachbarn (*k*) auf der *x*-Achse an. Dieses Diagramm ist verfügbar, wenn ein Ziel und eine *k*-Auswahl aktiviert sind.

k- und Funktions-Auswahlfehler-Protokoll

Abbildung 20-21
k- und Funktions-Auswahl



Hierbei handelt es sich um Funktionsauswahldiagramme (siehe Funktionsauswahl-Fehlerprotokoll auf S. 156), durch *k* in Felder unterteilt. Dieses Diagramm ist verfügbar, wenn ein Ziel und die *k*- und Funktions-Auswahl aktiviert sind.

Klassifikationsmatrix

Abbildung 20-22
Klassifikationsmatrix

Partition		Vorhergesagt		
		0	**1**	**Prozent korrekt**
Training	**0**	111	1	99.11%
	1	7	33	82.50%
	Prozent (insgesamt)	77.64%	22.37%	94.74%

Diese Tabelle enthält die Kreuzklassifikation der festgestellten Werte im Vergleich zu den vorhergesagten Werten des Ziels nach Partitionen. Sie ist verfügbar, wenn ein kategoriales Ziel vorhanden ist.

- Die Zeile (Fehlend) in der Holdout-Partition enthält Holdout-Fälle mit fehlenden Werten im Ziel. Diese Fälle tragen zur Prüfstichprobe bei: Gesamtprozentwerte, aber nicht die Werte für "Prozent korrekt".

Fehlerzusammenfassung

Abbildung 20-23
Fehlerzusammenfassung

Partition	Sum-of-Squares Error
Training	622043

Diese Tabelle ist verfügbar, wenn eine Zielvariable vorhanden ist. Sie enthält die Fehler für das Modell, Quadratsummenfehler für stetige Ziele und die Fehlerrate (100 % − Gesamtprozent korrekt) für kategoriale Ziele.

Kapitel
21

Diskriminanzanalyse

Die Diskriminanzanalyse erstellt ein Vorhersagemodell für Gruppenzugehörigkeiten. Dieses Modell besteht aus einer Diskriminanzfunktion (oder bei mehr als zwei Gruppen ein Set von Diskriminanzfunktionen) auf der Grundlage derjenigen linearen Kombinationen der Einflußvariablen, welche die beste Diskriminanz zwischen den Gruppen ergeben. Die Funktionen werden aus einer Stichprobe der Fälle erzeugt, bei denen die Gruppenzugehörigkeit bekannt ist. Diese Funktionen können dann auf neue Fälle mit Messungen für die Einflußvariablen, aber unbekannter Gruppenzugehörigkeit angewandt werden.

Hinweis: Die Gruppenvariable kann mehr als zwei Werte besitzen. Die Codes für die Gruppenvariable müssen allerdings ganzzahlige Werte sein, und Sie müssen hierfür die minimalen und maximalen Werte festlegen. Fälle mit Werten außerhalb dieser Grenzen werden von der Analyse ausgeschlossen.

Beispiel. Im Durchschnitt verbrauchen Personen in kühlen Ländern mehr Kalorien pro Tag als Bewohner der Tropen, und ein größerer Anteil der Personen in den kühlen Ländern sind Stadtbewohner. Ein Forscher möchte diese Informationen in einer Funktion zusammenfassen, um zu bestimmen, wie gut eine bestimmte Person diesen beiden Ländergruppen zugeordnet werden kann. Der Forscher nimmt an, dass auch die Bevölkerungsgröße und Wirtschaftsinformationen relevant sein könnten. Mit der Diskriminanzanalyse können Sie die Koeffizienten der linearen Diskriminanzfunktion schätzen, die im Prinzip genauso wie die rechte Seite einer Regressionsgleichung bei mehrfacher Regression aufgebaut ist. Unter Verwendung der Koeffizienten *a*, *b*, *c* und *d* lautet die Funktion also:

D = a * Klima + b * Städtisch + c * Bevölkerung + d * Bruttosozialprodukt der Region je Einwohner.

Wenn diese Variablen für die Unterscheidung zwischen den beiden Klimazonen relevant sind, müssen sich die Werte von *D* für tropische und kühlere Länder unterscheiden. Falls Sie eine schrittweise Methode für die Variablenauswahl verwenden, stellen Sie unter Umständen fest, dass nicht alle vier Variablen in die Funktion aufgenommen werden müssen.

Statistiken.Für jede Variable: Mittelwerte, Standardabweichungen, univariate ANOVA. Für jede Analyse: Box-*M*, Korrelationsmatrix innerhalb der Gruppen, Kovarianzmatrix innerhalb der Gruppen, Kovarianzmatrix der einzelnen Gruppen, gesamte Kovarianzmatrix. Für jede kanonische Diskriminanzfunktion: Eigenwert, Prozentwert der Varianz, kanonische Korrelation, Wilks-Lambda, Chi-Quadrat. Für jeden Schritt: a-priori-Wahrscheinlichkeit, Funktionskoeffizienten nach Fisher, nicht standardisierte Funktionskoeffizienten, Wilks-Lambda für jede kanonische Funktion.

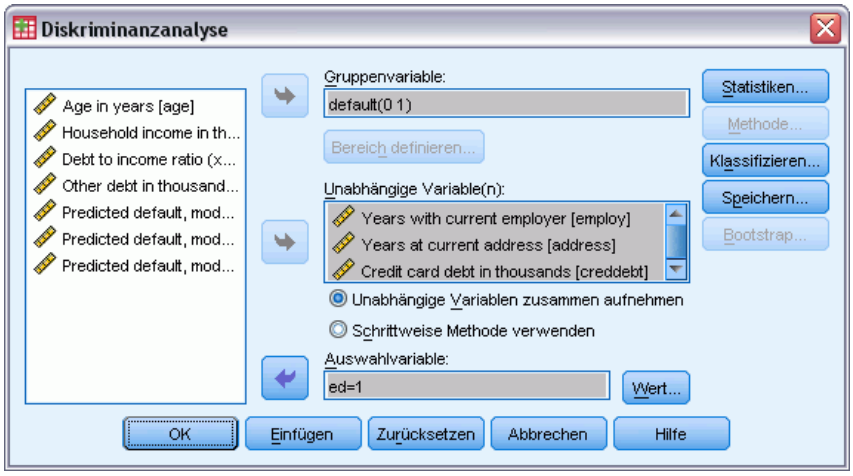
Daten. Die Gruppenvariable muss über eine begrenzte Anzahl unterschiedener Kategorien verfügen, die als ganzzahlige Werte kodiert werden. Unabhängige nominale Variablen müssen in Dummy- oder Kontrastvariablen umkodiert werden.

Annahmen. Die Fälle müssen unabhängig sein. Einflußvariablen müssen in multivariater Normalverteilung vorliegen, und die Varianz-Kovarianz-Matrizen innerhalb der Gruppen müssen zwischen den Gruppen gleich groß sein. Die Gruppenzugehörigkeit muss sich wechselseitig ausschließen (das heißt, kein Fall gehört zu mehr als einer Gruppe) und umfassend sein (das heißt, alle Fälle gehören zu einer Gruppe). Diese Prozedur ist am effektivsten, wenn die Gruppenzugehörigkeit eine rein kategoriale Variable ist. Wenn die Gruppenzugehörigkeit hingegen auf den Werten einer stetigen Variablen basiert (zum Beispiel bei einem Vergleich von IQ-Werten), sollten Sie die lineare Regression in Betracht ziehen, um von den reichhaltigeren Informationen zu profitieren, die in der stetigen Variablen selbst enthalten sind.

So lassen Sie eine Diskriminanzanalyse berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Klassifizieren > Diskriminanzanalyse

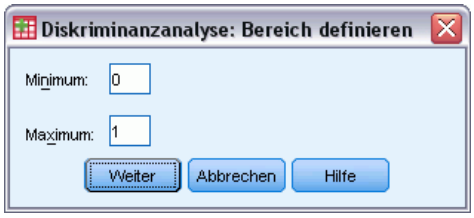
Abbildung 21-1
Dialogfeld "Diskriminanzanalyse"



- Wählen Sie eine Gruppenvariable mit ganzzahligen Werten aus und klicken Sie auf Bereich definieren, um die gewünschten Kategorien festzulegen.
- Wählen Sie die unabhängigen Variablen oder Einflußvariablen aus. (Wenn die Gruppenvariable nichtganzzahlig ist, können Sie eine Variable mit dieser Eigenschaft im Menü "Transformieren" mit dem Befehl "Automatisch umkodieren" erstellen.)
- Wählen Sie die gewünschte Methode für die Eingabe der unabhängigen Variablen aus.
  - **Unabhängige Variablen zusammen aufnehmen.** Nimmt alle unabhängigen Variablen, welche die Toleranzkriterien erfülllen, gleichzeitig auf.
  - **Schrittweise Methode verwenden.**Verwendet ein schrittweises Verfahren zur Steuerung von Variablenaufnahme und Variablenausschluss.
- Wahlweise können Sie die Fälle auch mithilfe einer Auswahlvariablen auswählen.

Diskriminanzanalyse: Bereich definieren

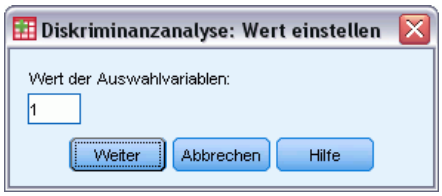
Abbildung 21-2
Dialogfeld "Diskriminanzanalyse: Bereich definieren"



Geben Sie den kleinsten (Minimum) und den größten (Maximum) Wert der Gruppenvariablen für die Analyse an. Fälle mit Werten außerhalb dieses Bereichs werden in der Diskriminanzanalyse nicht verwendet, aber ausgehend von den Ergebnissen der Analyse in eine der vorhandenen Gruppen eingeordnet. Die Minimum- und Maximumwerte müssen ganzzahlig sein.

Diskriminanzanalyse: Fälle auswählen

Abbildung 21-3
Dialogfeld "Diskriminanzanalyse: Wert einstellen"



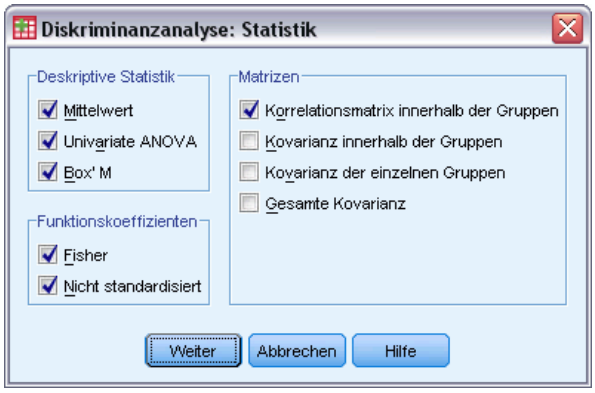
So wählen Sie die Fälle für die Analyse aus:

- Wählen Sie im Dialogfeld "Diskriminanzanalyse" eine Auswahlvariable aus.
- Klicken Sie auf Wert, um eine Ganzzahl als Auswahlvariable einzugeben.

Bei der Ableitung der Diskriminanzfunktionen werden nur die Fälle verwendet, deren Auswahlvariablen den angegebenen Wert aufweisen. Statistiken und Klassifikationsergebnisse werden sowohl für die ausgewählten als auch für die nicht ausgewählten Fälle erzeugt. Mit diesem Prozess liegt ein Mechanismus vor, mit dem neue Fälle anhand von bereits vorhandenen Daten klassifiziert werden können oder mit dem Sie Ihre Daten in Teilmengen von Lern- und Testfällen einteilen können, um so eine Gültigkeitsprüfung des erzeugten Modells durchzuführen.

Diskriminanzanalyse: Statistik

Abbildung 21-4
Dialogfeld "Diskriminanzanalyse: Statistik"



Deskriptive Statistiken. Verfügbare Optionen sind Mittelwerte (einschließlich Standardabweichungen), univariate ANOVA und der Box-*M*-Test.

- **Mittelwerte.** Zeigt Gesamt- und Gruppenmittelwerte sowie Standardabweichungen für die unabhängigen Variablen an.
- **Univariate ANOVA.** Führt für jede unabhängige Variable eine einfaktorielle Varianzanalyse durch, d. h. einen Test auf Gleichheit der Gruppenmittelwerte.
- **Box-M.** Ein Test auf Gleichheit der Kovarianzmatrizen der Gruppen. Bei hinreichend großen Stichproben bedeutet ein nichtsignifikanter p-Wert, dass die Anhaltspunkte für unterschiedliche Matrizen nicht ausreichend sind. Der Test ist empfindlich gegenüber Abweichungen von der multivariaten Normalverteilung.

Funktionskoeffizienten. Verfügbare Optionen sind Klassifikationskoeffizienten nach Fisher und nicht standardisierte Koeffizienten.

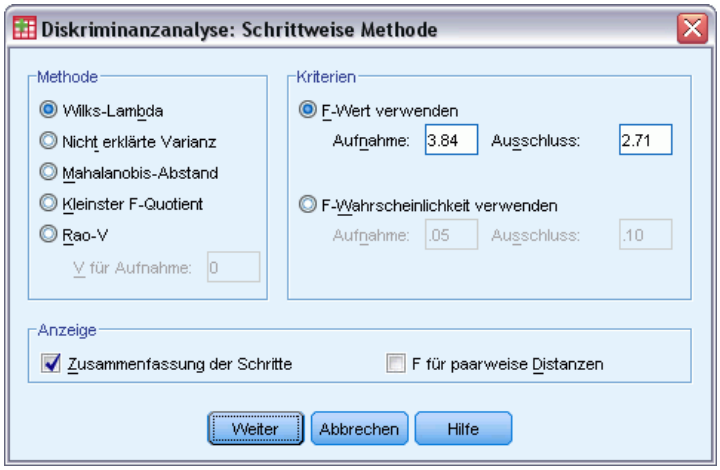
- **Fisher.** Zeigt die Koeffizienten der Klassifizierungsfunktion nach Fisher an, die direkt für die Klassifizierung verwendet werden können. Es wird ein eigenes Set von Koeffizienten der Klassifizierungsfunktion für jede Gruppe ermittelt. Ein Fall wird der Gruppe zugewiesen, für die er den größten Diskriminanzwert (Klassifizierungsfunktionswert) aufweist.
- **Nichtstandardisiert.** Zeigt die nichtstandardisierten Koeffizienten der Diskriminanzfunktion an.

Matrizen. Als Koeffizientenmatrizen für unabhängige Variablen stehen die Korrelationsmatrix innerhalb der Gruppen, die Kovarianzmatrix innerhalb der Gruppen, die Kovarianzmatrix der einzelnen Gruppen und die gesamte Kovarianzmatrix zur Verfügung.

- **Korrelationsmatrix innerhalb der Gruppen.** Zeigt eine gemeinsame Korrelationsmatrix innerhalb der Gruppen an, die als Mittel der separaten Kovarianzmatrizen für alle Gruppen vor der Berechnung der Korrelationen bestimmt wird.
- **Kovarianz innerhalb der Gruppen.** Zeigt eine gemeinsame Kovarianzmatrix innerhalb der Gruppen an, die sich von der Gesamt-Kovarianzmatrix unterscheiden kann. Die Matrix wird als Mittel der einzelnen Kovarianzmatrizen für alle Gruppen berechnet.
- **Kovarianz der einzelnen Gruppen.** Zeigt separate Kovarianzmatrizen für jede Gruppe an.
- **Gesamte Kovarianz.** Zeigt die Kovarianzmatrix für alle Fälle an, so als wären sie aus einer einzigen Stichprobe.

Diskriminanzanalyse: Schrittweise Methode

Abbildung 21-5
Dialogfeld "Diskriminanzanalyse: Schrittweise Methode"



Methode. Wählen Sie die Statistiken aus, die für die Aufnahme oder den Ausschluß neuer Variablen dienen sollen. Die Optionen Wilks-Lambda, nicht erklärte Varianz, Mahalanobis-Abstand, kleinster *F*-Quotient und Rao-*V* stehen zur Verfügung. Mit Rao-*V* können Sie den Mindestanstieg von *V* für eine einzugebende Variable angeben.

- **Wilks-Lambda.** Eine Auswahlmethode für Variablen bei der schrittweisen Diskriminanzanalyse. Die Aufnahme von Variablen in die Gleichung erfolgt anhand der jeweiligen Verringerung von Wilks-Lambda. Bei jedem Schritt wird diejenige Variable aufgenommen, die den Gesamtwert von Wilks-Lambda am meisten vermindert.
- **Nicht erklärte Varianz.** Bei jedem Schritt wird die Variable aufgenommen, welche die Summe der nicht erklärten Variation zwischen den Gruppen minimiert.
- **Mahalanobis-Abstand.** Dieses Maß gibt an, wie weit die Werte der unabhängigen Variablen eines Falles vom Mittelwert aller Fälle abweichen. Ein großer Mahalanobis-Abstand charakterisiert einen Fall, der bei einer oder mehreren unabhängigen Variablen Extremwerte besitzt.
- **Kleinster F-Quotient.** Eine Methode für die Variablenauswahl in einer schrittweisen Analyse. Sie beruht auf der Maximierung eines F-Quotienten, der aus dem Mahalanobis-Abstand zwischen den Gruppen errechnet wird.
- **Rao-V.** Ein Maß für die Unterschiede zwischen Gruppenmittelwerten. Auch Lawley-Hotelling-Spur genannt. Bei jedem Schritt wird die Variable aufgenommen, die den Anstieg des Rao-V maximiert. Wenn Sie diese Option ausgewählt haben, geben Sie den Minimalwert ein, den eine Variable für die Aufnahme in die Analyse aufweisen muss.

Kriterien. Verfügbar sind F-Wert verwenden und F-Wahrscheinlichkeit verwenden. Geben Sie Werte für die Aufnahme und den Ausschluß der Variablen an.

- **F-Wert verwenden.** Eine Variable wird in ein Modell aufgenommen, wenn ihr F-Wert größer ist als der Aufnahmewert. Sie wird ausgeschlossen, wenn der F-Wert kleiner ist als der Ausschlusswert. Der Aufnahmewert muss größer sein als der Ausschlusswert und beide Werte müssen positiv sein. Um mehr Variablen in das Modell aufzunehmen, senken Sie

den Aufnahmewert. Um mehr Variablen aus dem Modell auszuschließen, erhöhen Sie den Ausschlusswert.

- **F-Wahrscheinlichkeit verwenden.** Eine Variable wird in das Modell aufgenommen, wenn das Signifikanzniveau ihres F-Werts kleiner ist als der Aufnahmewert. Sie wird ausgeschlossen, wenn das Signifikanzniveau größer ist als der Ausschlusswert. Der Aufnahmewert muss kleiner sein als der Ausschlusswert und beide Werte müssen positiv sein. Um mehr Variablen in das Modell aufzunehmen, erhöhen Sie den Aufnahmewert. Um mehr Variablen aus dem Modell auszuschließen, senken Sie den Ausschlusswert.

Anzeigen. Mit Zusammenfassung der Schritte können Sie nach jedem Schritt die Statistiken für alle Variablen anzeigen lassen. Bei Auswahl von F für paarweise Distanzen wird für jedes Gruppenpaar eine Matrix des paarweisen *F*-Quotienten angezeigt.

Diskirminanzanalyse: Klassifizieren

Abbildung 21-6
Diskriminanzanalyse – Dialogfeld "Klassifizieren"

A-priori-Wahrscheinlichkeit. Diese Option legt fest, ob die Klassifikationskoeffizientenare A-priori-Wissen der Gruppenzugehörigkeit angepasst werden.

- **Alle Gruppen gleich.** Gleiche A-priori-Wahrscheinlichkeit wird für alle Gruppen angenommen; dies wirkt sich nicht auf die Koeffizienten aus.
- **Aus der Gruppengröße berechnen.** Die beobachteten Gruppengrößen in Ihrer Stichprobe bestimmen die A-priori-Wahrscheinlichkeiten der Gruppenzugehörigkeit. Wenn zum Beispiel 50 % der Beobachtungen der Analyse in die erste, 25 % in die zweite und 25 % in die dritte Gruppe fallen, werden die Klassifikationskoeffizienten angepasst, um die Wahrscheinlichkeit der Zugehörigkeit in der ersten Gruppe relativ zu den beiden anderen zu erhöhen.

Anzeigen. Die verfügbaren Anzeigeoptionen lauten: "Fallweise Ergebnisse", "Zusammenfassende Tabelle" und "Klassifikation mit Fallauslassung".

- **Fallweise Ergebnisse.** Für jeden Fall werden Codes für die tatsächliche Gruppe, die vorhergesagte Gruppe, A-posteriori-Wahrscheinlichkeiten und Diskriminanzwerte angezeigt.

- **Zusammenfassende Tabelle.** Die Anzahl der Fälle, die auf Grundlage der Diskriminanzanalyse jeder der Gruppen richtig oder falsch zugeordnet werden. Zuweilen auch als Klassifikationsmatrix bezeichnet.
- **Klassifikation mit Fallauslassung.** Jeder Fall der Analyse wird durch Funktionen aus allen anderen Fällen unter Auslassung dieses Falls klassifiziert. Diese Klassifikation wird auch als "U-Methode" bezeichnet.

Fehlende Werte durch Mittelwert ersetzen. Wenn Sie diese Option wählen, werden fehlende Werte durch den Mittelwert der jeweiligen unabhängigen Variablen ersetzt, allerdings nur während der Klassifikation der Gruppen.

Kovarianzmatrix verwenden. Sie können wählen, ob zur Klassifikation der Fälle die Kovarianzmatrix innerhalb der Gruppen oder die gruppenspezifische Kovarianzmatrix verwendet werden soll.

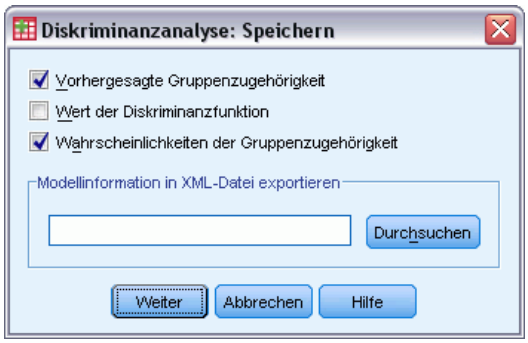
- **Innerhalb der Gruppen.** Zur Klassifizierung von Fällen wird die gemeinsame Kovarianzmatrix innerhalb der Gruppen verwendet.
- **Gruppenspezifisch.** Für die Klassifizierung werden gruppenspezifische Kovarianzmatrizen verwendet. Da die Klassifizierung auf Diskriminanzfunktionen und nicht auf ursprünglichen Variablen basiert, entspricht diese Option nicht immer der Verwendung einer quadratischen Diskriminanzfunktion.

Diagramme. Die verfügbaren Diagrammoptionen sind “Kombinierte Gruppen”, “Gruppenspezifisch” und “Territorien”.

- **Kombinierte Gruppen.** Erzeugt ein alle Gruppen umfassendes Streudiagramm der Werte für die ersten beiden Diskriminanzfunktionen. Wenn nur eine Funktion vorliegt, wird stattdessen ein Histogramm angezeigt.
- **Gruppenspezifisch.** Erzeugt gruppenspezifische Streudiagramme der Werte für die ersten beiden Diskriminanzfunktionen. Wenn nur eine Funktion vorliegt, werden stattdessen Histogramme angezeigt.
- **Territorien.** Ein Diagramm der Grenzen, mit denen Fälle auf der Grundlage von Funktionswerten in Gruppen klassifiziert werden. Die Zahlen entsprechen den Gruppen, in die die Fälle klassifiziert wurden. Der Mittelwert jeder Gruppe wird durch einen darin liegenden Stern (*) angezeigt. Dieses Diagramm wird nicht angezeigt, wenn nur eine einzige Diskriminanzfunktion vorliegt.

Diskriminanzanalyse: Speichern

Abbildung 21-7
Dialogfeld "Diskriminanzanalyse: Speichern"



Sie können der aktiven Datendatei neue Variablen hinzufügen. Die verfügbaren Optionen sind "Vorhergesagte Gruppenzugehörigkeit" (eine einzelne Variable), "Wert der Diskriminanzfunktion" (eine Variable für jede Diskriminanzfunktion in der Lösung) und "Wahrscheinlichkeiten der Gruppenzugehörigkeit" unter Berücksichtigung der Werte der Diskriminanzfunktion (eine Variable pro Gruppe).

Des weiteren können Sie Modellinformationen in die angegebene Datei exportieren. Anhand dieser Modelldatei können Sie die Modellinformationen zu Bewertungszwecken auf andere Datendateien anwenden.

Zusätzliche Funktionen beim Befehl DISCRIMINANT

Mit der Befehlssyntax können Sie auch Folgendes:

- Durchführen von mehreren Diskriminanzanalysen (mit einem Befehl) und Festlegen der Reihenfolge, in der die Variablen eingegeben werden (mit dem Unterbefehl `ANALYSIS`).
- Eingeben von a-priori-Wahrscheinlichkeiten für den Klassifikation (mit dem Unterbefehl `PRIORS`).
- Anzeigen von rotierten Mustern und Strukturmatrizen (mit dem Unterbefehl `ROTATE`).
- Begrenzen der Anzahl von extrahierten Diskriminanzfunktionen (mit dem Unterbefehl `FUNCTIONS`).
- Beschränken der Klassifikation auf die Fälle, die für die Analyse ausgewählt (oder nicht ausgewählt) wurden (mit dem Unterbefehl `SELECT`).
- Einlesen und Analysieren der Korrelationsmatrix (mit dem Unterbefehl `MATRIX`).
- Schreiben einer Korrelationsmatrix für die spätere Analyse (mit dem Unterbefehl `MATRIX`).

Siehe *Befehlssyntaxreferenz* für die vollständigen Syntaxinformationen.

Kapitel
22

Faktorenanalyse

Mit der Faktorenanalyse wird versucht, die zugrunde liegenden Variablen oder **Faktoren** zu bestimmen, welche die Korrelationsmuster innerhalb eines Satzes beobachteter Variablen erklären. Die Faktorenanalyse wird häufig zur Datenreduktion verwendet, indem wenige Faktoren identifiziert werden, welche den größten Teil der in einer großen Anzahl manifester Variablen aufgetretenen Varianz erklären. Die Faktorenanalyse kann auch zum Erzeugen von Hypothesen über kausale Mechanismen oder zum Sichten von Variablen für die anschließende Analyse verwendet werden (zum Beispiel, um vor einer linearen Regressionsanalyse Kollinearität zu erkennen).

Die Prozedur "Faktorenanalyse" bietet ein hohes Maß an Flexibilität:

- Es stehen sieben Methoden der Faktorextraktion zur Verfügung.
- Es sind fünf Rotationsmethoden verfügbar, einschließlich der direkten Oblimin-Methode und Promax-Methode für nicht orthogonale Rotationen.
- Für die Berechnung von Faktorwerten stehen drei Methoden zur Verfügung. Die Werte können für weitere Analysen als Variablen gespeichert werden.

Beispiel. Welche Einstellungen der befragten Personen liegen den gegebenen Antworten bei einer politischen Untersuchung zugrunde? Bei der Untersuchung der Korrelationen zwischen den Themen der Umfrage zeigen sich signifikante Überschneidungen zwischen verschiedenen Untergruppen von Themen. Fragen zu Steuern korrelieren gewöhnlich miteinander, ebenso wie Fragen zum Thema Bundeswehr und so weiter. Mit der Faktorenanalyse können Sie die Anzahl der zugrunde liegenden Faktoren untersuchen und in vielen Fällen die konzeptionelle Bedeutung der Faktoren bestimmen. Zusätzlich können Sie für jeden Fall Faktorwerte berechnen lassen, die sich dann für weiterführende Analysen verwenden lassen. Zum Beispiel könnten Sie ein logistisches Regressionsmodell erstellen, um das Wahlverhalten auf der Grundlage von Faktorwerten vorherzusagen.

Statistiken. Für jede Variable: Anzahl gültiger Fälle, Mittelwert und Standardabweichung. Für jede Faktorenanalyse: Korrelationsmatrix der Variablen mit Signifikanzniveaus, Determinante, Inverse; reproduzierte Korrelationsmatrix mit Anti-Image; Anfangslösung (Kommunalitäten, Eigenwerte und Prozentsatz der erklärten Varianz); Kaiser-Meyer-Olkin-Maß für die Angemessenheit der Stichproben und Bartlett-Test auf Sphärizität; nicht rotierte Lösung mit Faktorladungen, Kommunalität und Eigenwerten; sowie rotierte Lösung mit rotierter Mustermatrix und Transformationsmatrix. Für schiefe Rotationen: rotierte Muster- und Strukturmatrizen; Koeffizientenmatrix der Faktorwerte und Kovarianzmatrix des Faktors. Diagramme: Screeplot von Eigenwerten und Diagramm der Ladungen der ersten zwei oder drei Faktoren.

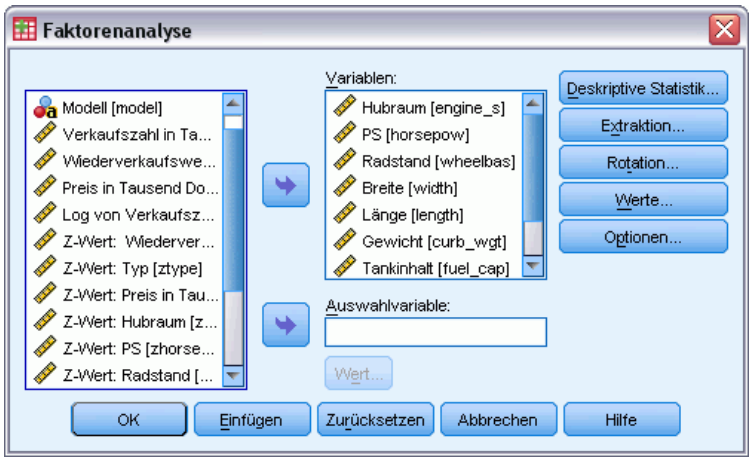
Daten. Die Variablen müssen auf dem **Intervall-** oder **Verhältnis-**Niveau quantitativ sein. Kategoriale Daten (wie beispielsweise Religion oder Geburtsland) sind für die Faktorenanalyse nicht geeignet. Daten, für welche die Korrelationskoeffizienten nach Pearson sinnvoll berechnet werden können, eignen sich gewöhnlich für eine Faktorenanalyse.

Annahmen. Die Daten sollten für jedes Variablenpaar in einer bivariaten Normalverteilung vorliegen. Beobachtungen müssen unabhängig sein. Im Modell der Faktorenanalyse ist festgelegt, dass Variablen durch gemeinsame Faktoren (die vom Modell geschätzten Faktoren) und eindeutige Faktoren (die sich nicht zwischen den beobachteten Variablen überschneiden) bestimmt sind. Die errechneten Schätzwerte basieren auf der Annahme, dass alle eindeutigen Faktoren weder miteinander noch mit den gemeinsamen Faktoren korrelieren.

So lassen Sie eine Faktorenanalyse berechnen:

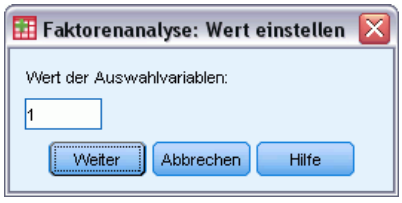
- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Dimensionsreduzierung > Faktorenanalyse...
- Wählen Sie die Variablen für die Faktorenanalyse aus.

Abbildung 22-1
Dialogfeld "Faktorenanalyse"



Faktorenanalyse: Fälle auswählen

Abbildung 22-2
Dialogfeld "Faktorenanalyse: Wert einstellen"



So wählen Sie die Fälle für die Analyse aus:

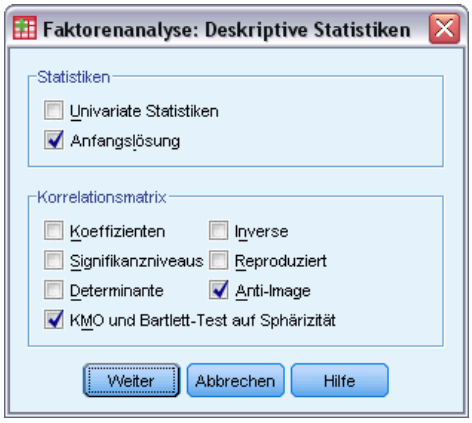
- Wählen Sie eine Auswahlvariable aus.

▶ Klicken Sie auf Wert, um eine Ganzzahl als Auswahlvariable einzugeben.

Nur Fälle mit diesem Wert für die Auswahlvariable werden für die Faktorenanalyse verwendet.

Faktorenanalyse: Deskriptive Statistiken

Abbildung 22-3
Dialogfeld "Faktorenanalyse: Deskriptive Statistiken"



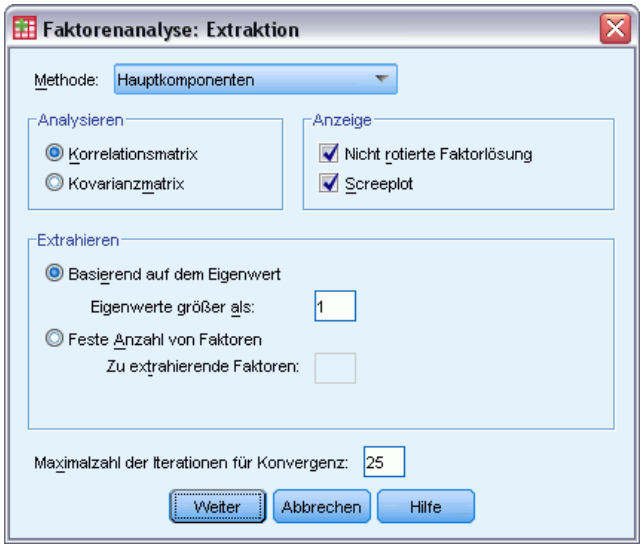
Statistik. Univariate Statistiken enthalten den Mittelwert, die Standardabweichung und die Anzahl gültiger Fälle für jede Variable. Die Anfangslösung zeigt die anfänglichen Kommunalitäten, Eigenwerte und den Prozentwert der erklärten Varianz an.

Korrelationsmatrix. Die verfügbaren Optionen sind Koeffizienten, Signifikanzniveaus, Determinante, Inverse, Reproduziert, Anti-Image sowie KMO und Bartlett-Test auf Sphärizität.

- **KMO und Bartlett-Test auf Sphärizität.** Das Kaiser-Meyer-Olkin-Maß für Angemessenheit der Stichproben überprüft, ob die partiellen Korrelationen zwischen Variablen klein sind. Der Bartlett-Test auf Sphärizität prüft, ob die Korrelationsmatrix eine Einheitsmatrix ist, wobei das Faktorenmodell in diesem Fall ungeeignet wäre.
- **Reproduziert.** Die geschätzte Korrelationsmatrix aus der Faktorlösung. Residuen (Differenz zwischen geschätzten und beobachteten Korrelationen) werden ebenfalls angezeigt.
- **Anti-Image.** Die Anti-Image-Korrelationsmatrix enthält die negativen Werte der partiellen Korrelationskoeffizienten. Die Anti-Image-Kovarianzmatrix enthält die negativen Werte der partiellen Kovarianzen. In einem guten Faktorenmodell sind die meisten außerhalb der Diagonalen liegenden Elemente klein. Das Maß der Stichprobeneignung einer Variablen wird auf der Diagonalen der Anti-Image-Korrelationsmatrix angezeigt.

Faktorenanalyse: Extraktion

Abbildung 22-4
Dialogfeld "Faktorenanalyse: Extraktion"



Methode. Hier kann die Methode der Faktorenextraktion festgelegt werden. Folgende Methoden sind verfügbar: Hauptkomponenten, ungewichtete kleinste Quadrate, verallgemeinerte kleinste Quadrate, Maximum Likelihood, Hauptachsen-Faktorenanalyse, Alpha-Faktorisierung und Image-Faktorisierung.

- **Hauptkomponentenanalyse.** Eine Methode zur Faktorextraktion. Sie wird verwendet, um unkorrelierte Linearkombinationen der beobachteten Variablen zu bilden. Die erste Komponente besitzt den größten Varianzanteil. Nachfolgende Komponenten erklären stufenweise kleinere Anteile der Varianz. Sie sind alle miteinander unkorreliert. Die Hauptkomponentenanalyse wird zur Ermittlung der Anfangslösung der Faktorenanalyse verwendet. Sie kann verwendet werden, wenn die Korrelationsmatrix singulär ist.
- **Ungewichtete kleinste Quadrate.** Eine Faktorextraktionsmethode, welche die Summe der quadrierten Differenzen zwischen der beobachteten und der reproduzierten Korrelationsmatrix unter Nichtberücksichtigung der Diagonalen minimiert.
- **Verallgemeinerte Methode der kleinsten Quadrate.** Eine Faktorextraktionsmethode, welche die Summe der quadrierten Differenzen zwischen der beobachteten und der reproduzierten Korrelationsmatrix minimiert. Die Korrelationen werden mit dem inversen Wert der Eindeutigkeit gewichtet, sodass Variablen mit großer Eindeutigkeit schwach und solche mit kleiner Eindeutigkeit stärker gewichtet werden.
- **Maximum-Likelihood-Methode.** Eine Methode für die Faktorextraktion, die Parameterschätzer erzeugt, bei denen die Wahrscheinlichkeit am größten ist, dass sie die beobachtete Korrelationsmatrix erzeugt haben, wenn die Stichprobe aus einer multivariaten Normalverteilung stammt. Die Korrelationen werden durch die inverse Eindeutigkeit der Variablen gewichtet und es wird ein iterativer Algorithmus eingesetzt.
- **Hauptachsen-Faktorenanalyse.** Eine Methode der Faktorextraktion aus der ursprünglichen Korrelationsmatrix, bei der die auf der Diagonalen befindlichen quadrierten Korrelationskoeffizienten als Anfangsschätzer der Kommunalitäten verwendet werden. Diese

Faktorladungen werden benutzt, um neue Kommunalitäten zu schätzen, welche die alten Schätzer auf der Diagonalen ersetzen. Die Iterationen werden so lange fortgesetzt, bis die Änderungen in den Kommunalitäten von einer Iteration zur nächsten das Konvergenzkriterium der Extraktion erfüllen.

- **Alpha.** Eine Methode der Faktorextraktion, welche die Variablen in der Analyse als eine Stichprobe aus einer Grundgesamtheit aller potenziellen Variablen betrachtet. Dies vergrößert die Alpha-Reliabilität der Faktoren.
- **Image-Faktorisierung.** Eine Faktorextraktionsmethode, die von Guttman entwickelt wurde und auf der Imagetheorie basiert. Der gemeinsame Teil einer Variablen – partielles Image genannt – ist als ihre lineare Regression auf die verbleibenden Variablen definiert und nicht als eine Funktion von hypothetischen Faktoren.

Analysieren. Hier können Sie entweder eine Korrelationsmatrix oder eine Kovarianzmatrix festlegen.

- **Korrelationsmatrix.**Diese Funktion ist nützlich, wenn die Variablen in Ihrer Analyse anhand verschiedener Skalen gemessen werden.
- **Kovarianzmatrix.**Diese Funktion ist nützlich, wenn Sie die Faktorenanalyse auf mehrere Gruppen mit unterschiedlichen Varianzen für die einzelnen Variablen anwenden möchten.

Extrahieren. Sie können entweder alle Faktoren, deren Eigenwerte über einem festgelegten Wert liegen, oder eine festgelegte Anzahl von Faktoren beibehalten.

Anzeigen. Hier können Sie die nicht rotierte Faktorlösung und ein Screeplot der Eigenwerte anfordern.

- **Nicht rotierte Faktorlösung.** Zeigt unrotierte Faktorladungen (Faktormustermatrix), Kommunalitäten und Eigenwerte für die Faktorlösung an.
- **Screeplot.** Ein Diagramm der Varianz, die jedem Faktor zugeordnet ist. Es dient dazu, zu bestimmen, wie viele Faktoren beibehalten werden sollen. Normalerweise zeigt das Diagramm einen deutlichen Bruch zwischen der starken Steigung der großen Faktoren und dem graduellen Verlauf der restlichen Faktoren (der "Geröllhalde", engl. "Scree").

Maximalzahl der Iterationen für Konvergenz. Hier können Sie für den Algorithmus eine Maximalzahl von Schritten zum Schätzen der Lösung festlegen.

Faktorenanalyse: Rotation

Abbildung 22-5
Dialogfeld "Faktorenanalyse: Rotation"

Methode. Hier können Sie die Methode der Faktor-Rotation auswählen. Die verfügbaren Methoden sind Varimax, Quartimax, Equamax, Promax oder Oblimin, direkt.

- **Varimax-Rotation.** Eine orthogonale Rotationsmethode, die die Anzahl der Variablen mit hohen Ladungen für jeden Faktor minimiert. Sie vereinfacht die Interpretation der Faktoren.
- **Methode Oblimin, direkt.** Ein Verfahren zur schiefwinkligen (nichtorthogonalen) Rotation. Wenn Delta den Wert 0 annimmt (Standardeinstellung), sind die Ergebnisse am schiefsten. Mit zunehmendem negativem Wert von Delta werden die Faktoren weniger schiefwinklig. Um den Standardwert von 0 zu überschreiben, geben Sie eine Zahl kleiner gleich 0,8 ein.
- **Quartimax-Rotation.** Eine Rotationsmethode, welche die Zahl der Faktoren minimiert, die zum Erklären aller Variablen benötigt werden. Sie vereinfacht die Interpretation der beobachteten Variablen.
- **Equamax-Rotation.** Eine Rotationsmethode, die eine Kombination zwischen der Varimax-Methode (vereinfacht die Faktoren) und der Quartimax-Methode (vereinfacht die Variablen) darstellt. Die Anzahl der Variablen mit hohen Ladungen auf einen Faktor sowie die Anzahl der Faktoren, die benötigt werden, um eine Variable zu erklären, werden minimiert.
- **Promax-Rotation.** Eine schiefe Rotation, bei der Faktoren korreliert sein dürfen. Diese Rotation kann schneller berechnet werden als eine direkte Oblimin-Rotation und ist daher nützlich für große Daten-Sets.

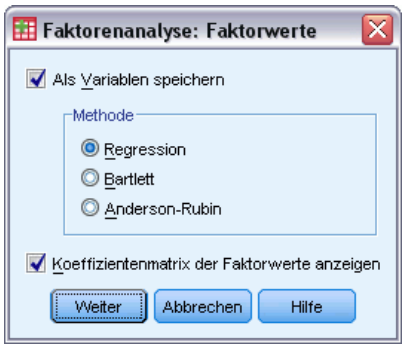
Anzeigen. Hiermit können Sie eine Ausgabe für die rotierte Lösung sowie Ladungsdiagramme für die ersten zwei oder drei Faktoren einbeziehen.

- **Rotierte Lösung.** Um eine rotierte Lösung zu erhalten, muss eine Rotationsmethode ausgewählt sein. Für orthogonale Rotationen werden die rotierte Mustermatrix und Faktortransformationsmatrix angezeigt. Für schiefe Rotationen werden Muster-, Struktur- und Faktorkorrelationsmatrix angezeigt.
- **Diagramm der Faktorladungen.** Dreidimensionales Diagramm der Faktorladungen für die ersten drei Faktoren. Für eine Lösung mit zwei Faktoren wird ein zweidimensionales Diagramm angezeigt. Das Diagramm wird nicht angezeigt, wenn nur ein Faktor extrahiert wird. Auf Wunsch zeigen die Diagramme rotierte Lösungen an.

Maximalzahl der Iterationen für Konvergenz. Hier können Sie eine Maximalzahl von Schritten zum Durchführen der Rotation für den Algorithmus festlegen.

Faktorenanalyse: Faktorwerte

Abbildung 22-6
Dialogfeld "Faktorenanalyse: Faktorwerte"



Als Variablen speichern. Hiermit wird für jeden Faktor in der endgültigen Lösung eine neue Variable erstellt.

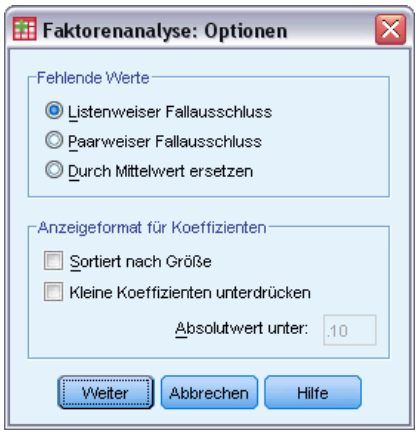
Methode. Alternative Methoden zur Berechnung der Faktorwerte sind Regression, Bartlett und Anderson-Rubin.

- **Regressionsmethode.** Eine Methode, um Koeffizienten für Faktorwerte zu schätzen. Die Faktorwerte haben einen Mittelwert von 0 und eine Varianz, die der quadrierten multiplen Korrelation zwischen den geschätzten und den wahren Faktorwerten entspricht. Die Scores können korreliert sein, selbst wenn die Faktoren orthogonal sind.
- **Barlett-Werte.** Eine Methode, um Koeffizienten für Faktorwerte zu schätzen. Eine Methode zum Schätzen von Koeffizienten für Faktorwerte. Die erzeugten Faktorwerte haben einen Mittelwert von 0. Die Quadratsumme der eindeutigen Faktoren über dem Variablenbereich wird minimiert.
- **Anderson-Rubin-Methode.** Eine Methode zur Berechnung der Koeffizienten von Faktorwerten; eine Modifizierung der Bartlett-Methode, die die Orthogonalität der geschätzten Faktoren gewährleistet. Die berechneten Werte haben einen Mittelwert von 0 und eine Standardabweichung von 1 und sind unkorreliert.

Koeffizientenmatrix der Faktorwerte anzeigen. Hiermit werden die Koeffizienten angezeigt, mit denen die Variablen multipliziert werden, um Faktorwerte zu erhalten. Hiermit werden auch die Korrelationen zwischen Faktorwerten angezeigt.

Faktorenanalyse: Optionen

Abbildung 22-7
Dialogfeld "Faktorenanalyse: Optionen"



Fehlende Werte. Hier können Sie festlegen, wie fehlende Werte behandelt werden. Es stehen zur Verfügung: "**Listenweiser** Fallausschluss", "**Paarweiser** Fallausschluss" und "Durch Mittelwert ersetzen".

Anzeigeformat für Koeffizienten. Hiermit können Sie Einstellungen für Aspekte der Ausgabematrix vornehmen. Sie können die Koeffizienten nach Größe sortieren lassen und Koeffizienten mit absoluten Werten unterdrücken, die kleiner als der festgelegte Wert sind.

Zusätzliche Funktionen beim Befehl FACTOR

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Angeben von Konvergenzkriterien für die Iteration während der Extraktion und Rotation.
- Angeben von einzelnen rotierten Faktordiagrammen.
- Angeben der Anzahl der zu speichernden Faktorwerte.
- Angeben der Diagonalwerte für die Hauptachsen-Faktorenanalyse.
- Schreiben der Korrelationsmatrizen oder der Faktorladungs-Matrizen auf die Festplatte für eine spätere Analyse.
- Einlesen und Analysieren von Korrelationsmatrizen oder Faktorladungs-Matrizen.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
23

Auswählen einer Prozedur zum Durchführen einer Clusteranalyse

Clusteranalysen können mit den Prozeduren “Two-Step-Clusteranalyse”, “Hierarchische Clusteranalyse” oder “Clusterzentrenanalyse” durchgeführt werden. In jeder Prozedur wird ein anderer Algorithmus zum Erstellen von Clustern eingesetzt, und jede Prozedur verfügt über Optionen, die in den jeweils anderen Prozeduren nicht verfügbar sind.

Two-Step-Clusteranalyse. In vielen Fällen ist die Prozedur “Two-Step-Clusteranalyse” die beste Wahl. Sie bietet die folgenden speziellen Funktionen:

- Automatische Auswahl der optimalen Anzahl von Clustern sowie Maße, die bei der Auswahl des Cluster-Modells helfen
- Gleichzeitiges Erstellen von Cluster-Modellen mit kategorialen und stetigen Variablen
- Speichern des Cluster-Modells in einer externen XML-Datei und anschließendem Einlesen dieser Datei und Aktualisieren des Cluster-Modells mit neuen Daten.

Außerdem können von der Prozedur “Two-Step-Clusteranalyse” auch umfangreiche Datendateien analysiert werden.

Hierarchische Clusteranalyse. Die Prozedur “Hierarchische Clusteranalyse” ist auf kleinere Datendateien begrenzt (mehrere Hundert zu gruppierende Objekte), bietet jedoch die folgenden speziellen Funktionen:

- Möglichkeit der Zusammenfassung von Fällen oder Variablen in Clustern
- Funktion zum Berechnen eines Bereichs möglicher Lösungen und zum Speichern der Cluster-Zugehörigkeiten für jede dieser Lösungen
- Verschiedene Methoden zur Clusterbildung, Transformation von Variablen und Messung der Unähnlichkeit zwischen Clustern

Mit der Prozedur “Hierarchische Clusteranalyse” können Intervallvariablen (stetige Variablen), Zählvariablen oder binäre Variablen analysiert werden, wobei alle für die Prozedur ausgewählten Variablen jeweils denselben Typ aufweisen müssen.

Clusterzentrenanalyse. Die Prozedur “Clusterzentrenanalyse” ist auf stetige Daten beschränkt und setzt eine Festlegung der Cluster-Anzahl voraus, bietet jedoch die folgenden speziellen Funktionen:

- Funktion zum Speichern der Distanz vom Clusterzentrum für jedes Objekt
- Funktion zum Einlesen der anfänglichen Clusterzentren aus einer externen IBM® SPSS® Statistics-Datei und zum Speichern der endgültigen Clusterzentren in dieser Datei

Außerdem können von der Prozedur “Clusterzentrenanalyse” auch umfangreiche Datendateien analysiert werden.

Kapitel
24

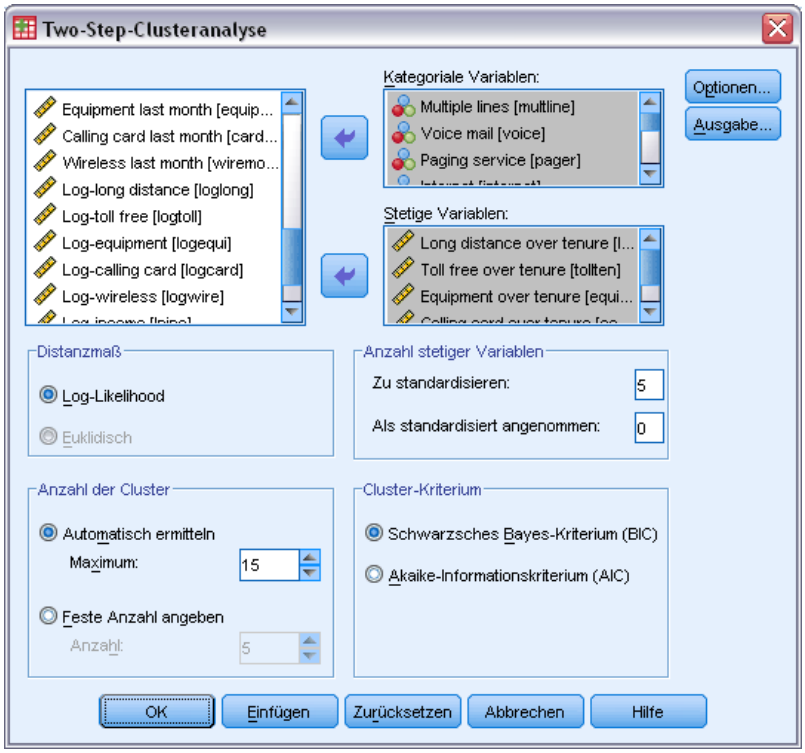
Two-Step-Clusteranalyse

Bei der Two-Step-Clusteranalyse handelt es sich um eine explorative Prozedur zum Ermitteln von natürlichen Gruppierungen (Clustern) innerhalb eines Daten-Sets, die anderenfalls nicht erkennbar wären. Der von der Prozedur verwendete Algorithmus verfügt über vielfältige nützliche Funktionen, durch die er sich von traditionellen Cluster-Methoden unterscheidet:

- **Verarbeitung von kategorialen und stetigen Variablen.**Die Annahme der Unabhängigkeit der Variablen ermöglicht eine kombinierte multinomiale Normalverteilung für kategoriale und stetige Variablen.
- **Automatische Auswahl der Cluster-Anzahl.**Durch den Vergleich der Werte eines Modellauswahlkriteriums in verschiedenen Clusteranalysen kann die optimale Anzahl der Cluster von der Prozedur automatisch bestimmt werden.
- **Skalierbarkeit.**Durch das Zusammenfassen der Datensätze in einem Clusterfunktionsbaum (CF-Baum) können mit dem Two-Step-Algorithmus sehr große Datendateien analysiert werden.

Beispiel. In Einzel- und Fachhandel werden Cluster-Methoden regelmäßig auf Daten angewendet, die Kaufgewohnheiten, Geschlecht, Alter und Einkommensniveau der Kundschaft beschreiben. Ziel der Analyse ist eine Ausrichtung der unternehmenseigenen Marketing- und Produktentwicklungsstrategien auf einzelne Konsumentengruppen, um Umsatzsteigerungen und Markentreue zu erreichen.

Abbildung 24-1
Dialogfeld "Two-Step-Clusteranalyse"



Distanzmaß. Mit dieser Auswahl legen Sie fest, wie Ähnlichkeiten zwischen zwei Clustern verarbeitet werden.

- **Log-Likelihood.**Mit dem Likelihood-Maß wird eine Wahrscheinlichkeitsverteilung für die Variablen vorgenommen. Bei stetigen Variablen wird von einer Normalverteilung, bei kategorialen Variablen von einer multinomialen Verteilung ausgegangen. Bei allen Variablen wird davon ausgegangen, dass sie unabhängig sind.
- **Euklidisch.**Das Euklidische Maß bezeichnet die "gerade" Distanz zwischen zwei Clustern. Es kann nur dann verwendet werden, wenn es sich bei sämtlichen Variablen um stetige Variablen handelt.

Anzahl der Cluster. Mit dieser Auswahl können Sie angeben, wie die Anzahl der Cluster bestimmt werden soll.

- **Automatisch ermitteln.**Mit dieser Prozedur wird das im Gruppenfeld "Cluster-Kriterium" angegebene Kriterium verwendet, um automatisch die "beste" Anzahl der Cluster zu ermitteln. Sie haben die Möglichkeit, eine positive Ganzzahl für die Höchstzahl der Cluster anzugeben, die von der Prozedur berücksichtigt werden sollen.
- **Feste Anzahl angeben.**Ermöglicht das Festlegen der Anzahl der Cluster für die Analyse. Geben Sie eine positive ganze Zahl ein.

Anzahl stetiger Variablen. Dieses Gruppenfeld enthält eine Zusammenfassung der Standardeinstellungen, die im Dialogfeld "Optionen" für stetige Variablen vorgenommen wurden.
Für weitere Informationen siehe Thema Two-Step-Clusteranalyse: Optionen auf S. 181.

Cluster-Kriterium. Mit dieser Auswahl legen Sie fest, wie die Anzahl der Cluster vom automatischen Cluster-Algorithmus bestimmt wird. Angegeben werden kann entweder das Bayes-Informationskriterium (BIC) oder das Akaikes-Informationskriterium (AIC).

Daten. Mit dieser Prozedur können sowohl stetige als auch kategoriale Variablen analysiert werden. Die Fälle bilden dabei die Objekte, die gruppiert werden sollen, während die Variablen die Attribute darstellen, auf deren Grundlage die Gruppierung erfolgt.

Fallreihenfolge. Beachten Sie, dass der Cluster-Funktionsbaum und die endgültige Lösung ggf. von der Reihenfolge der Fälle abhängig sein können. Um die Auswirkungen der Reihenfolge zu minimieren, mischen Sie die Fälle in zufälliger Reihenfolge. Prüfen Sie daher die Stabilität einer bestimmten Lösung, indem Sie verschiedene Lösungen abrufen, bei denen die Fälle in einer unterschiedlichen, zufällig ausgewählten Reihenfolge sortiert sind. In schwierigen Situationen mit äußerst umfangreichen Dateien führen Sie statt dessen mehrere Läufe aus, bei denen eine Stichprobe der Fälle in unterschiedlicher, zufälliger Reihenfolge angeordnet ist.

Annahmen. Das Likelihood-Distanzmaß geht davon aus, dass die Variablen im Clustermodell unabhängig sind. Außerdem wird für stetige Variablen eine Normal- bzw. Gauß-Verteilung und für kategoriale Variable eine multinomiale Verteilung vorausgesetzt. Empirische interne Tests zeigen, dass die Prozedur wenig anfällig gegenüber Verletzungen hinsichtlich der Unabhängigkeitsannahme und der Verteilungsannahme ist. Dennoch sollten Sie darauf achten, wie genau diese Voraussetzungen erfüllt sind.

Mit der Prozedur Bivariate Korrelationen können Sie die Unabhängigkeit zwischen zwei stetigen Variablen überprüfen. Mit der Prozedur Kreuztabellen können Sie die Unabhängigkeit zwischen zwei kategorialen Variablen überprüfen. Mit der Prozedur Mittelwerte können Sie die Unabhängigkeit zwischen einer stetigen und einer kategorialen Variablen überprüfen. Mit der Prozedur Explorative Datenanalyse prüfen Sie die Normalverteilung einer stetigen Variablen. Mit der Prozedur Chi-Quadrat-Test überprüfen Sie, ob eine kategoriale Variable eine bestimmte multinomiale Verteilung aufweist.

So lassen Sie eine Two-Step-Clusteranalyse berechnen:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Klassifizieren > Two-Step-Clusteranalyse...

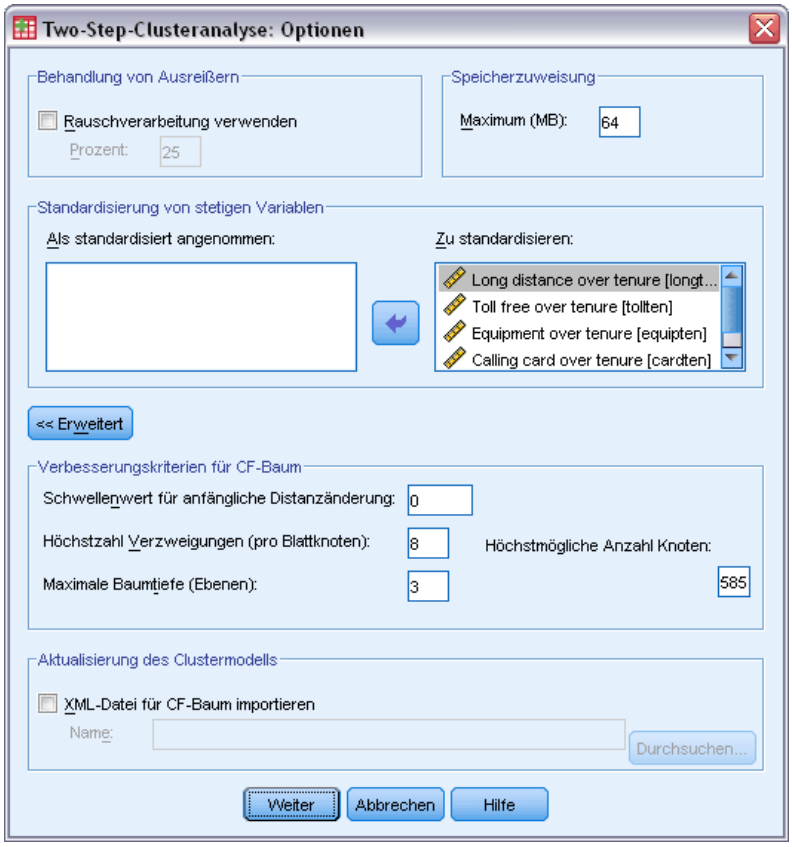
▶ Wählen Sie mindestens eine kategoriale oder stetige Variable aus.

Die folgenden Optionen sind verfügbar:

- Anpassen der Kriterien für die Erstellung der Cluster
- Auswählen der Einstellungen für die Rauschverarbeitung, Speicherzuweisung, Variablenstandardisierung und Eingabe des Clustermodells
- Ausgabe der Modellanzeige anfordern.
- Speichern der Modellergebnisse in der Arbeitsdatei oder in einer externen XML-Datei

Two-Step-Clusteranalyse: Optionen

Abbildung 24-2
Dialogfeld "Two-Step-Clusteranalyse: Optionen"



Behandlung von Ausreißern. Mit diesem Gruppenfeld können Sie Ausreißer während des Füllvorgangs des CF-Baums bei der Clusteranalyse gesondert behandeln. Der CF-Baum ist vollständig, wenn keine weiteren Fälle in einem Blattknoten aufgenommen werden können und kein Blattknoten mehr aufgeteilt werden kann.

- Wenn während des Füllvorgangs des CF-Baums eine Rauschverarbeitung stattfinden soll, wird der CF-Baum neu gebildet, nachdem Fälle von wenig besetzten Blättern auf einem "Rauschblatt" positioniert worden sind. Ein Blatt wird als wenig besetzt betrachtet, wenn es weniger Fälle als den angegebenen Prozentsatz der maximalen Blattgröße enthält. Nach der Neubildung des Baums können gegebenenfalls noch Ausreißer im CF-Baum positioniert werden. Andernfalls werden die Ausreißer verworfen.
- Wenn während des Füllvorgangs des CF-Baums keine Rauschverarbeitung stattfinden soll, wird der Baum unter Verwendung eines größeren Schwellenwerts für die Distanzänderung neu gebildet. Nach der abschließenden Clusteranalyse werden die Werte, die keinem Cluster zugewiesen werden konnten, als Ausreißer bezeichnet. Der Ausreißer-Cluster erhält die Identifikationsnummer –1 und wird nicht in die Auszählung der Anzahl von Clustern aufgenommen.

Speicherzuweisung. In diesem Gruppenfeld können Sie den maximalen Speicherplatz in MB angeben, der vom Cluster-Algorithmus verwenden soll. Wenn der für die Prozedur erforderliche Speicherplatz den maximalen Speicherplatz übersteigt, wird die Festplatte zum Speichern der Daten verwendet, die nicht in den Arbeitsspeicher passen. Geben Sie eine Zahl größer oder gleich 4 ein.

- Den größtmöglichen Wert, den Sie für Ihr System angeben können, erfahren Sie bei Ihrem Systemadministrator.
- Wenn dieser Wert zu niedrig ist, kann die Anzahl der Cluster unter Umständen nicht ordnungsgemäß ermittelt werden.

Variablenstandardisierung. Mit dem Cluster-Algorithmus werden standardisierte stetigen Variablen analysiert. Alle stetigen Variablen, die nicht standardisiert sind, sollten in der Liste "Zu standardisieren" verbleiben. Um Zeit und Verarbeitungsaufwand zu sparen, können Sie alle bereits standardisierten stetigen Variablen in der Liste "Als standardisiert angenommen" auswählen.

Erweiterte Optionen

Verbesserungskriterien für CF-Baum. Die folgenden Einstellungen für den Cluster-Algorithmus gelten insbesondere für den CF-Baum und sollten nur nach sorgfältiger Prüfung geändert werden:

- **Schwellenwert für anfängliche Distanzänderung.**Hierbei handelt es sich um den anfänglichen Schwellenwert, der zum Erstellen des CF-Baums verwendet wird. Wenn das Hinzufügen eines gegebenen Falls zu einem Blatt des CF-Baums eine Dichte unterhalb dieses Schwellenwerts ergibt, wird das Blatt nicht geteilt. Wenn die Dichte den Schwellenwert überschreitet, wird das Blatt geteilt.
- **Höchstzahl Verzweigungen (pro Blattknoten).**Hierbei handelt es sich um die maximale Anzahl an untergeordneten Knoten, über die ein Blattknoten verfügen kann.
- **Maximale Baumtiefe.**Die maximale Anzahl an Ebenen, über die ein CF-Baum verfügen kann.
- **Höchstmögliche Anzahl Knoten.**Gibt die maximale Anzahl an CF-Baumknoten an, die von der Prozedur anhand der Gleichung $(b^{d+1} - 1) / (b - 1)$ potenziell erstellt werden können, wobei b für die Höchstzahl der Verzweigungen und d für die maximale Baumtiefe steht. Beachten Sie, dass ein extrem großer CF-Baum die Systemressourcen stark belastet und somit die Prozedurleistung beinträchtigen kann. Die Mindestanforderung pro Knoten beträgt 16 Bytes.

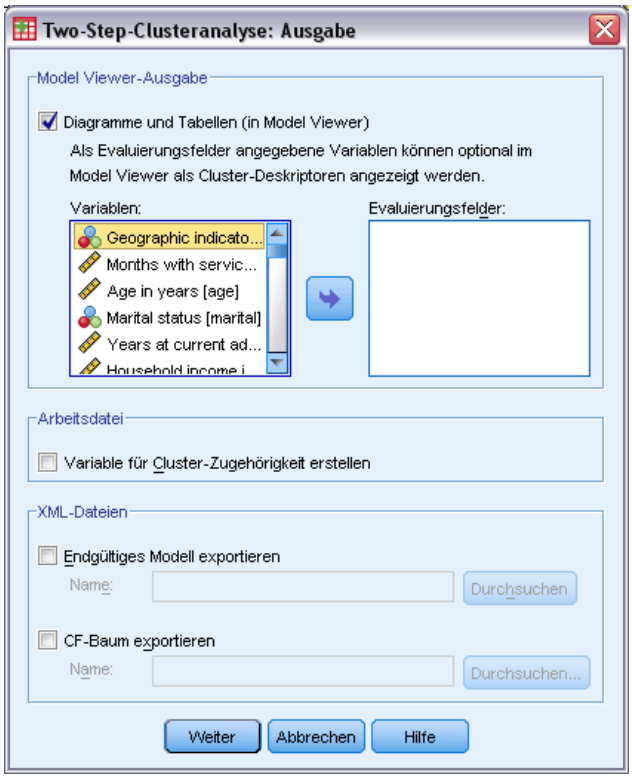
Aktualisierung des Clustermodells. Mit diesem Gruppenfeld können Sie ein Clustermodell importieren und aktualisieren, das in einer vorangegangenen Analyse erstellt wurde. Die Eingabedatei enthält den CF-Baum im XML-Format. Das Modell wird dann mit den Daten der aktiven Datei aktualisiert. Die Variablennamen müssen im Hauptdialogfeld in der Reihenfolge ausgewählt werden, in der sie in der vorangegangenen Analyse angegeben wurden. Die XML-Datei bleibt unverändert, es sei denn, Sie speichern die neuen Modelldaten unter demselben Dateinamen. Für weitere Informationen siehe Thema Two-Step-Clusteranalyse: Ausgabe auf S. 183.

Bei einer Aktualisierung des Clustermodells werden zur Erstellung des CF-Baums dieselben Optionen verwendet, die für das ursprüngliche Modell gelten. Genauer gesagt werden die Optionen für Distanzmaß, Rauschverarbeitung, Speicherzuweisung und Verbesserungskriterien für den CF-Baum aus dem gespeicherten Modell übernommen, wobei die in den Dialogfeldern für diese Optionen vorgenommenen Einstellungen ignoriert werden.

Hinweis: Beim Ausführen einer Aktualisierung des Clustermodells wird von der Prozedur vorausgesetzt, dass keiner der ausgewählten Fälle in der Arbeitsdatei für die Erstellung des ursprünglichen Clustermodells verwendet wurde. Außerdem gilt die Annahme, dass die Fälle für die Modellaktualisierung der gleichen Grundgesamtheit entstammen wie die Fälle, die zur Erstellung des ursprünglichen Modells verwendet wurden. Das heißt, es wird angenommen, dass die Mittelwerte und Varianzen der stetigen Variablen sowie die Ebenen der kategorialen Variablen in beiden Fallgruppen identisch sind. Wenn Ihre "neuen" und "alten" Fallgruppen aus heterogenen Grundgesamtheiten stammen, müssen Sie die Two-Step-Clusteranalyse für eine Kombination der beiden Fallgruppen ausführen, um optimale Ergebnisse zu erzielen.

Two-Step-Clusteranalyse: Ausgabe

Abbildung 24-3
Dialogfeld "Two-Step-Clusteranalyse: Ausgabe"



Ausgabe der Modellanzeige. In diesem Gruppenfeld können Sie Optionen für die Anzeige der Ergebnisse der Clusteranalyse einstellen.

- **Diagramme und Tabellen.** Enthält modellbezogene Ausgaben einschließlich Tabellen und Diagrammen. Tabellen in der Modellansicht enthalten eine Modellzusammenfassung und ein Raster mit Clustern nach Funktionen. Die grafische Ausgabe in der Modellansicht

enthält ein Diagramm zur Cluster-Qualität, Cluster-Größen, die Variablenwichtigkeit, Cluster-Vergleichsraster und Zelleninformationen.

- **Evaluierungsfelder.** Mit dieser Option werden Cluster-Daten für Variablen berechnet, die bei der Cluster-Erstellung nicht verwendet wurden. Evaluierungsfelder können zusammen mit den Eingabefunktionen in der Modellanzeige angezeigt werden, indem sie im untergeordneten Dialogfeld "Anzeigen" ausgewählt werden. Felder mit fehlenden Werten werden ignoriert.

Arbeitsdatei. Mit diesem Gruppenfeld können Sie Variablen in der Arbeitsdatei speichern.

- **Variable für Cluster-Zugehörigkeit erstellen.**Diese Variable enthält für jeden Fall eine Cluster-Identifikationsnummer. Der Name dieser Variablen lautet *tsc_n*, wobei *n* eine positive Ganzzahl ist, die auf die Ordinalzahl der Arbeitsdatei hinweist, die von dieser Prozedur in einer gegebenen Sitzung gespeichert wurde.

XML-Dateien. Das endgültige Clustermodell und der CF-Baum sind zwei Arten von Ausgabedateien, die als XML-Format exportiert werden können.

- **Endgültiges Modell exportieren.**Das endgültige Clustermodell wird in die angegebene Datei exportiert. Anhand dieser Modelldatei können Sie die Modellinformationen zu Bewertungszwecken auf andere Datendateien anwenden.
- **CF-Baum exportieren.**Mit dieser Option können Sie den aktuellen Stand des Cluster-Baums speichern und zu einem späteren Zeitpunkt mit neuen Daten aktualisieren.

Die Clusteranzeige

Clustermodelle werden üblicherweise verwendet, um Gruppen (oder Cluster) ähnlicher Datensätze zu finden, die auf den untersuchten Variablen basieren, wobei die Ähnlichkeit zwischen Elementen derselben Gruppe hoch und die Ähnlichkeit zwischen Elementen verschiedener Gruppen niedrig ist. Die Ergebnisse können zur Identifizierung von Zusammenhängen verwendet werden, die ansonsten nicht offensichtlich wären. So kann es zum Beispiel die Clusteranalyse von Kundenpräferenzen, Einkommensniveau und Kaufgewohnheiten ermöglichen, die Kundentypen zu identifizieren, die mit größerer Wahrscheinlichkeit auf eine bestimmte Marketingkampagne ansprechen.

Es gibt zwei Ansätze bei der Interpretierung der Ergebnisse in einer Cluster-Darstellung:

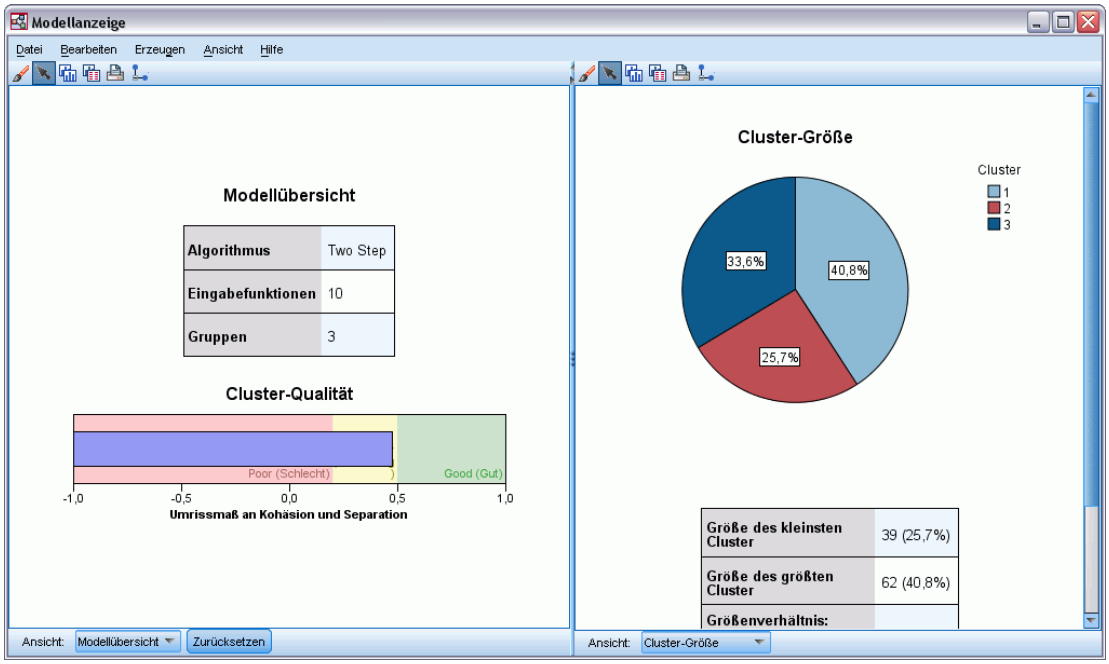
- Untersuchen der Cluster, um die Merkmale zu bestimmen, die in einem Cluster eindeutig sind. *Enthält ein Cluster sämtliche Käufer mit hohem Einkommen? Enthält dieser Cluster mehr Datensätze als die anderen?*
- Untersuchen von Feldern in allen Clustern, um zu bestimmen, wie die Werte in den Clustern verteilt sind. *Ist der Bildungsstand entscheidend für die Zugehörigkeit zu einem Cluster? Spielt ein hoher Kreditrahmen eine Rolle bei der Zugehörigkeit zu einem Cluster oder einem anderen?*

Wenn Sie die Hauptansicht und die zahlreichen verknüpften Ansichten in der Clusteranzeige nutzen, lassen sich diese Fragen beantworten.

Um Informationen über das Clustermodell anzuzeigen, aktivieren Sie (durch Doppelklicken) das Objekt Modellanzeige in der Clusteranzeige.

Clusteranzeige

Abbildung 24-4
Clusteranzeige mit Standardanzeige



Die Clusteranzeige besteht aus zwei Bereichen, der Hauptansicht im linken Bereich und der verknüpften oder Hilfsansicht im rechten Bereich. Es gibt zwei Hauptansichten:

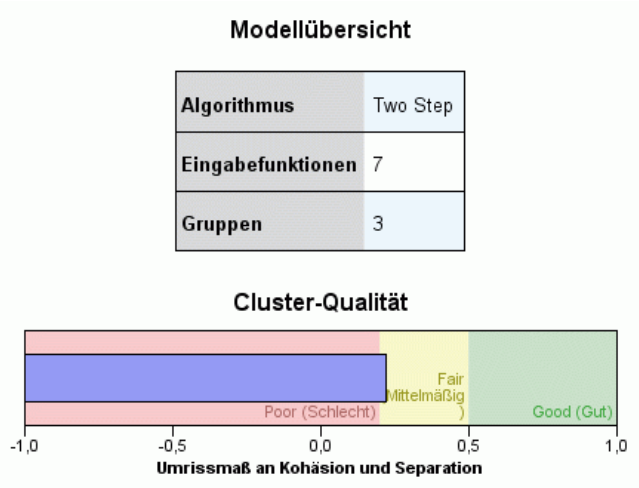
- Modellübersicht (Standard). Für weitere Informationen siehe Thema Ansicht Modellübersicht auf S. 186.
- Cluster. Für weitere Informationen siehe Thema Clusteransicht auf S. 187.

Es gibt vier verknüpfte/Hilfsansichten:

- Bedeutsamkeit des Prädiktors. Für weitere Informationen siehe Thema Ansicht “Bedeutsamkeit des Prädiktors” für Cluster auf S. 190.
- Clustergrößen (Standard). Für weitere Informationen siehe Thema Clustergrößenansicht auf S. 191.
- Zellenverteilung. Für weitere Informationen siehe Thema Ansicht Zellverteilung auf S. 192.
- Cluster-Vergleich. Für weitere Informationen siehe Thema Ansicht Clustervergleich auf S. 193.

Ansicht Modellübersicht

Abbildung 24-5
Ansicht "Modellübersicht" im Hauptpanel



Die Ansicht "Modellübersicht" zeigt eine Momentaufnahme oder eine Übersicht des Clustermodells einschließlich eines schattierten Umrissmaßes der Cluster-Kohäsion und Cluster-Separation, um schlechte, mittelmäßige und gute Ergebnisse anzuzeigen. Anhand dieser Momentaufnahme erkennen Sie schnell, ob die Qualität schlecht ist, so dass Sie dann gegebenenfalls zum Modellierungsknoten zurückkehren und die Clustermodell-Einstellungen ändern können, um ein besseres Ergebnis zu erzielen.

Die Ergebnisse "schlecht", "mittelmäßig" oder "gut" basieren auf der Arbeit von Kaufman und Rousseeuw (1990) zur Interpretation von Clusterstrukturen. In der Ansicht "Modellübersicht" entspricht ein gutes Ergebnis Daten, die von Kaufman und Rousseeuw als annehmbarer oder starker Hinweis auf eine Clusterstruktur eingestuft werden, "mittelmäßig" entspricht ihrer Einstufung als schwacher Hinweis und "schlecht" entspricht ihrer Einstufung als kein signifikanter Hinweis.

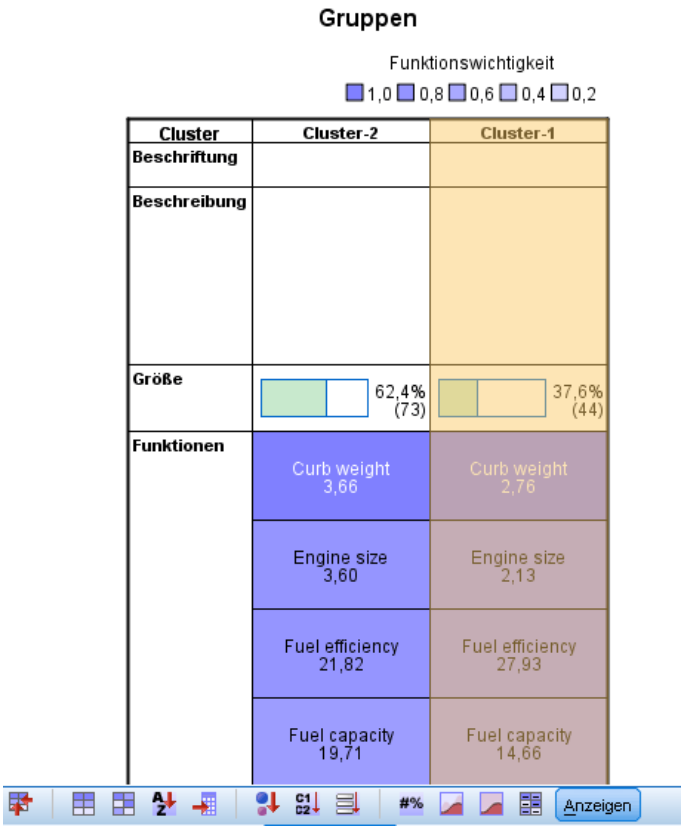
Das Umrissmaß ist ein Durchschnitt aller Datensätze (B−A) / max(A,B), wobei A der Abstand des Datensatzes zu seinem Clusterzentrum und B der Abstand des Datensatzes zu dem am nächsten liegenden, nicht zugehörigen Clusterzentrum ist. Ein Umrisskoeffizient von 1 würde bedeuten, dass alle Fälle direkt in ihren Clusterzentren liegen. Ein Wert von −1 würde bedeuten, dass alle Fälle in den Clusterzentren anderer Cluster liegen. Ein Wert von 0 bedeutet, dass die Fälle im Durchschnitt gleich weit entfernt von ihrem eigenen Clusterzentrum und dem nächsten benachbarten Cluster liegen.

Die Übersicht beinhaltet eine Tabelle, die folgende Daten enthält:

- **Algorithmus.** Der verwendete Clustering-Algorithmus, zum Beispiel "TwoStep".
- **Eingabefunktionen.** Die Anzahl der Felder, auch bekannt als **Eingaben** oder **Einflussgrößen**.
- **Cluster.** Die Anzahl der Cluster in der Lösung.

Clusteransicht

Abbildung 24-6
Ansicht "Clusterzentrum" im Hauptpanel



Die Clusteransicht enthält ein Cluster-nach-Funktionen-Raster mit Clusternamen, -größen und -profilen für jeden Cluster.

Die Spalten in der Tabelle enthalten die folgenden Informationen:

- **Cluster.** Die Clusternummern werden von dem Algorithmus erstellt.
- **Bezeichnung.** Bezeichnungen für jeden Cluster (ist standardmäßig leer). Doppelklicken Sie in die Zelle, um eine Bezeichnung einzugeben, die den Clusterinhalt beschreibt; zum Beispiel "Käufer von Luxusautos".
- **Beschreibung.** Beschreibung des Clusterinhalts (ist standardmäßig leer). Doppelklicken Sie in die Zelle, um eine Beschreibung des Clusters einzugeben, zum Beispiel "Alter 55+, Berufstätige, Einkommen über $100.000".

- **Größe.** Die Größe jedes Clusters als Prozentsatz der gesamten Cluster-Stichprobe. Jede Größenzelle in der Tabelle zeigt einen vertikalen Balken, der den Größenprozentsatz innerhalb des Clusters, einen Größenprozentsatz in numerischem Format und die Cluster-Fallzahl anzeigt.
- **Merkmale.** Die einzelnen Eingaben oder Einflussgrößen, standardmäßig nach Gesamtwichtigkeit sortiert. Wenn Spalten die gleiche Größe aufweisen, werden sie in aufsteigender Sortierfolge ihrer Clusternummern angezeigt.

  Die Gesamtwichtigkeit des Merkmals wird von der Farbe der Zellenhintergrundschattierung angezeigt; das wichtigste Merkmal ist am dunkelsten, das am wenigsten wichtige Merkmal ist unschattiert. Ein Hinweis oberhalb der Tabelle erläutert die Wichtigkeit, die jeder Merkmalszellfarbe zugewiesen ist.

Wenn Sie mit der Maus über eine Zelle fahren, wird der volle Name/die Bezeichnung des Merkmals und der Wichtigkeitswert der Zelle angezeigt. Je nach Anzeige- und Merkmalstyp können auch weitere Informationen angezeigt werden. In der Ansicht “Clusterzentrum” zählen die Zellenstatistik und der Zellenwert dazu; zum Beispiel: “Mittelwert: 4.32”. Bei kategorischen Merkmalen zeigt die Zelle den Namen der häufigsten (typischen) Kategorie und deren Prozentsatz.

In der Ansicht “Cluster” können Sie verschiedene Anzeigearten für die Clusterinformationen auswählen:

- Cluster und Funktionen transponieren. Für weitere Informationen siehe Thema Cluster und Merkmale transponieren auf S. 188.
- Merkmale sortieren. Für weitere Informationen siehe Thema Merkmale sortieren auf S. 189.
- Cluster sortieren. Für weitere Informationen siehe Thema Cluster sortieren. auf S. 189.
- Zelleninhalte auswählen. Für weitere Informationen siehe Thema Zelleninhalt auf S. 189.

Cluster und Merkmale transponieren

Standardmäßig werden Cluster als Spalten und Merkmale als Zeilen angezeigt. Um die Anzeige umzudrehen, klicken Sie auf die Schaltfläche Cluster und Merkmale transponieren links von der Schaltfläche Merkmale sortieren nach. Dies kann zum Beispiel wünschenswert sein, wenn zahlreiche Cluster angezeigt werden, um den horizontalen Bildlauf bei der Datenansicht zu verringern.

Abbildung 24-7
Transponierte Cluster im Hauptpanel

Cluster	Beschriftung	Beschreibung	Größe	
cluster-1			45,0% (91)	BP HIGH (41,8%)
cluster-3			35,0% (70)	BP NORMAL (51,4%)
cluster-2			19,0% (39)	BP HIGH (100,0%)

Merkmale sortieren

Die Schaltflächen Merkmale sortieren nach ermöglichen Ihnen die Auswahl, wie Merkmalzellen angezeigt werden:

- **Gesamtwichtigkeit.** Das ist die standardmäßige Sortierfolge. Die Merkmale werden in absteigender Sortierfolge der Gesamtwichtigkeit sortiert, und die Sortierfolge ist dieselbe bei allen Clustern. Wenn Merkmale gebundene Wichtigkeitswerte aufweisen, sind die gebundenen Merkmale in aufsteigender Sortierfolge der Merkmalnamen aufgelistet.
- **Wichtigkeit innerhalb der Cluster.** Die Merkmale werden hinsichtlich ihrer Wichtigkeit für jeden Cluster sortiert. Wenn Merkmale gebundene Wichtigkeitswerte aufweisen, sind die gebundenen Merkmale in aufsteigender Sortierfolge der Merkmalnamen aufgelistet. Wenn diese Option ausgewählt wird, variiert üblicherweise die Sortierfolge in den Clustern.
- **Name.** Die Merkmale werden nach Namen in alphabetischer Reihenfolge sortiert.
- **Datenfolge.** Die Merkmale werden nach ihrer Reihenfolge im Datensatz sortiert.

Cluster sortieren.

Standardmäßig werden Cluster ihrer Größe nach absteigend sortiert. Mit den Schaltflächen Cluster sortieren nach können Sie die Cluster nach Namen in alphabetischer Reihenfolge sortieren, oder, wenn Sie eindeutige Bezeichnungen erstellt haben, stattdessen auch in alphanumerischer Bezeichnungsreihenfolge.

Merkmale mit derselben Bezeichnung werden nach Clustername sortiert. Wenn die Cluster nach Bezeichnung sortiert sind und Sie die Bezeichnung eines Clusters bearbeiten, wird die Sortierfolge automatisch aktualisiert.

Zelleninhalt

Mit den Schaltflächen Zellen können Sie die Anzeige der Zelleninhalte für Merkmale- und Evaluationsfelder ändern.

- **Clusterzentren.** Standardmäßig zeigen Zellen Namen/Bezeichnungen und das Maß der Zentraltendenz für jede Cluster/Merkmal-Kombination an. Für kontinuierliche Felder wird der Mittelwert angezeigt und für kategorische Felder der Modus (die am häufigsten auftretende Kategorie) mit Kategorieprozentsatz.
- **Absolute Verteilungen.** Zeigt die Merkmalnamen/-bezeichnungen und die absoluten Verteilungen der Merkmale in jedem Cluster. Bei kategorischen Merkmalen werden Balkendiagramme angezeigt, mit überlagerter Anzeige der Kategorien, die nach ihren Datenwerten aufsteigend geordnet sind. Bei kontinuierlichen Mekmalen stellt die Anzeige ein gleichmäßiges Dichtediagramm dar, bei dem die gleichen Endpunkte und Intervalle für jeden Cluster verwendet werden.

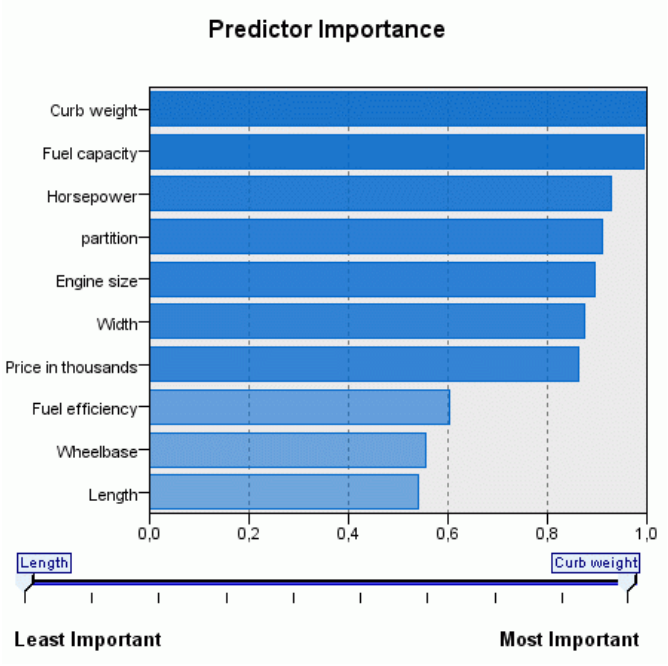
  Die intensiv rote Anzeige stellt die Clusterverteilung dar, wogegen die blassere Anzeige die Gesamtdaten repräsentiert.
- **Relative Verteilungen.** Zeigt die Merkmalnamen/-bezeichnungen und die relativen Verteilungen in den Zellen. Im Allgemeinen sind die Anzeigen vergleichbar mit denen für absolute Verteilungen, nur dass stattdessen die relativen Verteilungen dargestellt sind.

Die intensiv rote Anzeige stellt die Clusterverteilung dar, wogegen die blassere Anzeige die Gesamtdaten repräsentiert.

- **Basisansicht.** Bei sehr vielen Clustern kann es schwierig sein, sämtliche Details ohne Bildlauf zu sehen. Wählen Sie diese Ansicht, um den Bildlauf einzuschränken und die Anzeige auf eine kompaktere Version der Tabelle zu ändern.

Ansicht "Bedeutsamkeit des Prädiktors" für Cluster

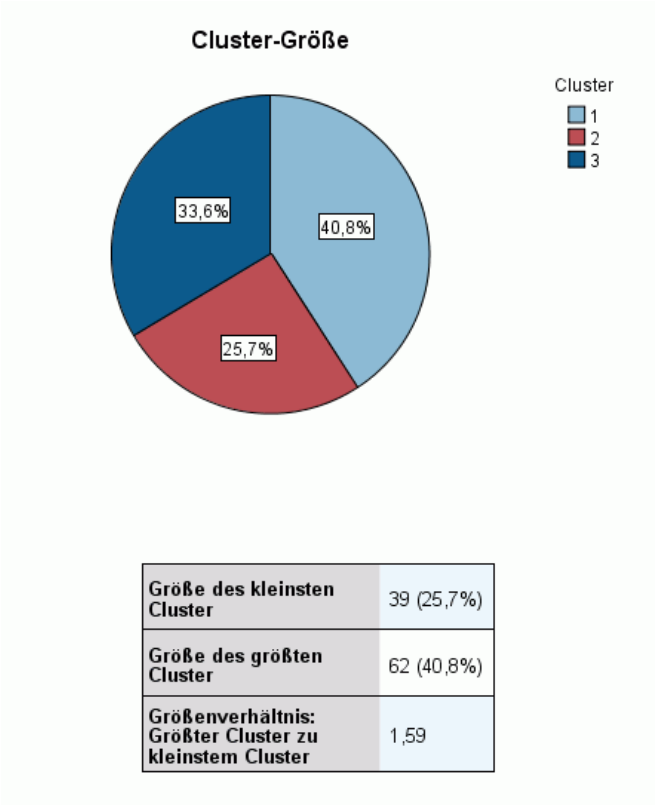
Abbildung 24-8
Ansicht "Bedeutsamkeit des Prädiktors" für Cluster im Verknüpfungspanel



Die Ansicht "Bedeutsamkeit des Prädiktors" zeigt die relative Wichtigkeit jedes Felds bei Schätzung des Modells.

Clustergrößenansicht

Abbildung 24-9
Ansicht "Clustergrößen" im Verknüpfungspanel



Größe des kleinsten Cluster	39 (25,7%)
Größe des größten Cluster	62 (40,8%)
Größenverhältnis: Größter Cluster zu kleinstem Cluster	1,59

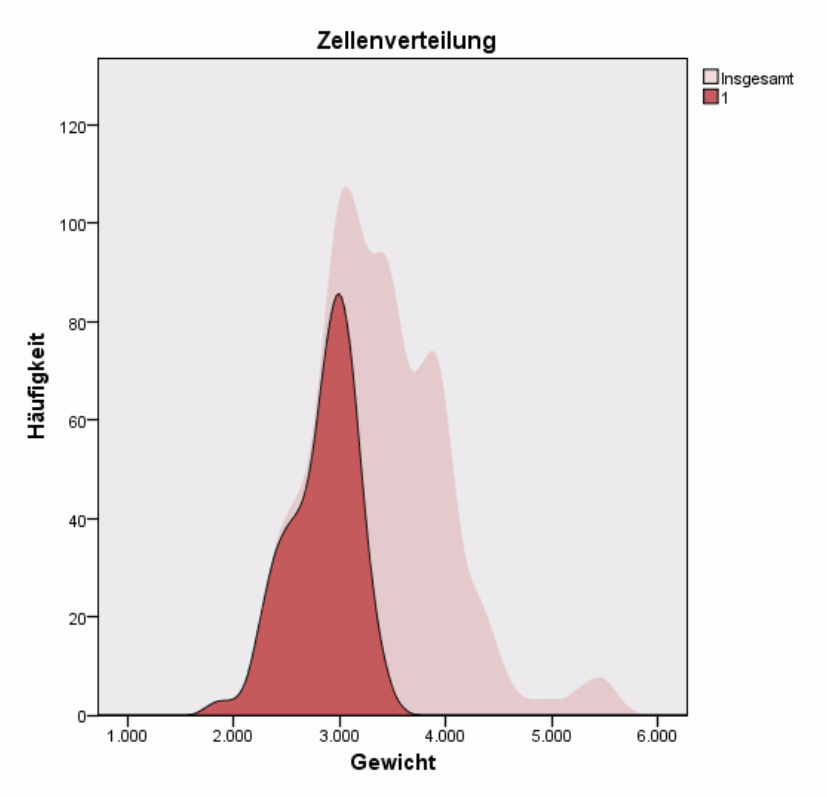
Die Ansicht "Clustergrößen" zeigt ein Tortendiagramm, das sämtliche Cluster enthält. In jedem Stückchen wird die prozentuale Größe des Clusters angezeigt; fahren Sie mit der Maus über ein Stückchen, um den Zahlwert in diesem Stück anzuzeigen.

Unterhalb des Diagramms sind in einer Tabelle die folgenden Informationen aufgelistet:

- Größe des kleinsten Clusters (als Zahlwert und Prozentsatz des Ganzen).
- Größe des größten Clusters (als Zahlwert und Prozentsatz des Ganzen).
- Verhältnis der Größe des größten Clusters zum kleinsten Cluster.

Ansicht Zellverteilung

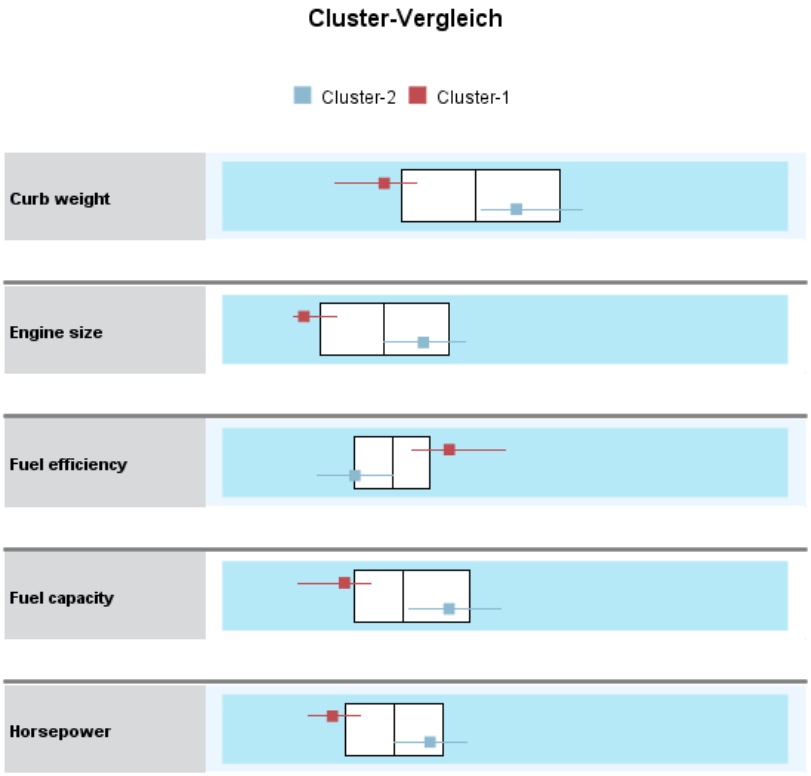
Abbildung 24-10
Ansicht "Zellverteilung" im Verknüpfungspanel



Die Ansicht "Zellverteilung" zeigt ein erweitertes, detaillierteres Diagramm der Datenverteilung für jede Merkmalszelle, die Sie in der Tabelle im Cluster-Hauptpanel auswählen.

Ansicht Clustervergleich

Abbildung 24-11
Ansicht "Clustervergleich" im Verknüpfungspanel



Die Ansicht "Clustervergleich" ist eine tabellarische Grafik, bei der die Merkmale in den Zeilen und die ausgewählten Cluster in den Spalten dargestellt werden. Mit dieser Ansicht lassen sich die Faktoren besser verstehen, die die Cluster ausmachen; außerdem hilft sie dabei, die Unterschiede zwischen den Clustern zu erkennen – nicht nur im Vergleich zum Gesamtdatensatz, sondern auch untereinander.

Zum Auswählen der Cluster für die Ansicht klicken Sie oben auf die Clusterspalte im Cluster-Hauptpanel. Wenn Sie die Strg-Taste oder die Umschalttaste beim Klicken gedrückt halten, können Sie mehrere Cluster zum Vergleich auswählen oder wieder aus der Auswahl entfernen.

Hinweis: Sie können bis zu fünf Cluster für die Anzeige auswählen.

Die Cluster werden in der Reihenfolge ihrer Auswahl angezeigt, während die Reihenfolge der Felder mit der Option Merkmale sortieren nach festgelegt wird. Wenn Sie Wichtigkeit innerhalb der Cluster auswählen, werden die Felder immer nach ihrer Gesamtwichtigkeit sortiert.

Die Hintergrunddiagramme zeigen die Gesamtverteilungen der Merkmale:

- Kategorische Merkmale sind als Punktdiagramme dargestellt, wobei die Größe des Punktes die häufigste/typische Kategorie für jeden Cluster (nach Merkmal) anzeigt.
- Kontinuierliche Merkmale sind als Boxplots angezeigt, der die Gesamtmediane und die Interquartilbereiche anzeigt.

Vor diesen Hintergrundansichten sind Boxplots für ausgewählte Cluster dargestellt:

- Bei kontinuierlichen Merkmalen zeigen quadratische Punktmarkierungen und horizontale Linien den Median und den Interquartilbereich für jeden Cluster an.
- Jeder Cluster ist mit einer anderen Farbe gekennzeichnet, die oben an der Ansicht angezeigt wird.

Navigieren in der Clusteranzeige

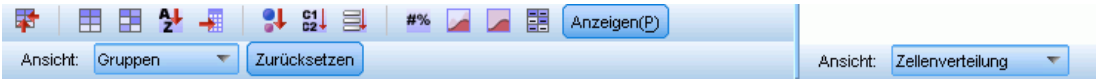
Bei der Clusteranzeige handelt es sich um eine interaktive Anzeige. Sie verfügen über folgende Möglichkeiten:

- Auswählen eines Felds oder eines Clusters für weitere Details
- Vergleichen von Clustern, um die Objekte von Interesse auszuwählen
- Verändern der Anzeige
- Transponierenvon Achsen

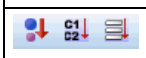
Verwendung der Symbolleisten

Sie können die Informationen, die in den Panels links und rechts angezeigt werden, mithilfe der Symbolleistenoptionen steuern. Mit der Symbolleistensteuerung können Sie die Ausrichtung der Anzeige ändern (oben-unten, links-rechts oder rechts-links). Außerdem können Sie die Clusteranzeige auf die Standardeinstellungen zurücksetzen und ein Dialogfeld öffnen, um den Inhalt der Clusteransicht im Hauptpanel zu spezifizieren.

Abbildung 24-12
Symbolleisten zum Steuern der in der Clusteranzeige angezeigten Daten



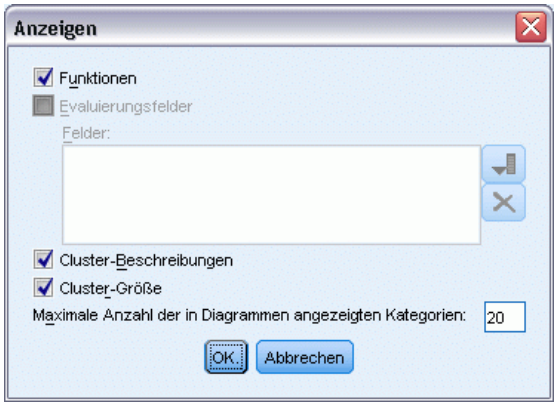
Die Optionen Merkmale sortieren nach, Cluster sortieren nach, Zellen und Anzeige sind nur verfügbar, wenn Sie die Ansicht Cluster im Hauptpanel auswählen. Für weitere Informationen siehe Thema Clusteransicht auf S. 187.

	Siehe Cluster und Merkmale transponieren auf S. 188
	Siehe Merkmale sortieren nach auf S. 189
	Siehe Cluster sortieren nach auf S. 189
	Siehe Zellen auf S. 189

Anzeige "Clusteransicht steuern"

Um zu steuern, was in der Clusteransicht im Hauptpanel angezeigt wird, klicken Sie auf die Schaltfläche Anzeige. Der Anzeige-Dialog wird geöffnet.

Abbildung 24-13
Clusteranzeige- Anzeigeoptionen



Merkmale. Standardmäßig ausgewählt. Deaktivieren Sie das Kästchen, um alle Eingabemerkmale auszublenden.

Evaluierungsfelder. Wählen Sie die anzuzeigenden Evaluierungsfelder aus (Felder, die nicht für die Erstellung des Clustermodells verwendet, sondern an die Modellanzeige zur Evaluierung der Cluster gesendet werden); standardmäßig werden keine angezeigt. *Hinweis*: Dieses Kontrollkästchen ist nicht verfügbar, wenn keine Evaluierungsfelder verfügbar sind.

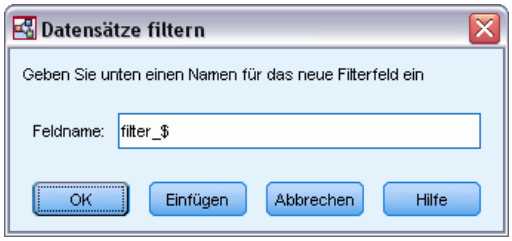
Clusterbeschreibungen. Standardmäßig ausgewählt. Deaktivieren Sie das Kontrollkästchen, um alle Clusterbeschreibungszellen auszublenden.

Clustergröße. Standardmäßig ausgewählt. Deaktivieren Sie das Kontrollkästchen, um alle Clustergrößenzellen auszublenden.

Maximale Anzahl an Kategorien. Geben Sie die maximale Anzahl an Kategorien an, die in den Diagrammen der kategorischen Merkmale angezeigt werden sollen; der Standard ist 20.

Datensätze filtern

Abbildung 24-14
Clusteranzeige - Fälle filtern



Wenn Sie weitere Informationen zu den Fällen in einem bestimmten Cluster oder einer Clustergruppe benötigen, können Sie eine Untergruppe an Datensätzen für die weitere Analyse auf der Grundlage der ausgewählten Cluster auswählen.

- Wählen Sie die Cluster in der Clusteransicht der Clusteranzeige aus. Sollen mehrere Knoten ausgewählt werden, halten Sie beim Klicken die Strg-Taste gedrückt.
- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Erzeugen > Datensätze filtern...
- Geben Sie einen Namen für die Filtervariable an. Die Datensätze aus den ausgewählten Clustern erhalten den Wert 1 für dieses Feld. Alle anderen Datensätze erhalten den Wert 0 und werden aus den nachfolgenden Analysen ausgeschlossen, bis Sie den Filterstatus ändern.
- Klicken Sie auf OK.

Kapitel
25

Hierarchische Clusteranalyse

Mit diesem Verfahren wird anhand ausgewählter Merkmale versucht, relativ homogene Fallgruppen oder Variablen zu identifizieren. Dabei wird ein Algorithmus eingesetzt, der für jeden Fall oder für jede Variable, einen separaten Cluster bildet und die Cluster so lange kombiniert, bis nur noch einer zurückbleibt. Sie können einfache Variablen analysieren oder eine Auswahl aus einer Vielfalt von Transformationen zur Standardisierung treffen. Distanz- oder Ähnlichkeitsmaße werden durch die Prozedur "Ähnlichkeiten" erzeugt. Für jeden Schritt werden Statistiken angezeigt, um Sie bei der Auswahl der besten Lösung zu unterstützen.

Beispiel.Können Gruppen von verschiedenen Fernseh-Shows identifiziert werden, die ein ähnliches Publikum ansprechen? Mithilfe der hierarchischen Clusteranalyse können Sie die Fernseh-Shows (Fälle) anhand der Merkmale der Zuschauer in homogene Gruppen (Cluster) aufteilen. Damit lassen sich beispielsweise Marktsegmente identifizieren. Sie können außerdem Städte (Fälle) in homogene Gruppen clustern, sodass vergleichbare Städte zum Testen verschiedener Marketingstrategien ausgewählt werden können.

Statistiken. Zuordnungsübersicht, Distanz- oder Ähnlichkeitsmatrix und Cluster-Zugehörigkeit für eine einzelne Lösung oder einen Bereich von Lösungen. Diagramme: Dendrogramme und Eiszapfendiagramme.

Daten. Bei den Variablen kann es sich um quantitative Daten, binäre Daten oder Häufigkeitsdaten handeln. Die Skalierung der Variablen spielt eine wichtige Rolle. Unterschiede in der Skalierung können sich auf Ihre Cluster-Lösung(en) auswirken. Wenn Ihre Variablen sehr unterschiedlich skaliert sind, eine also beispielsweise in Dollar und die andere in Jahren angegeben wird, empfiehlt sich die Standardisierung. (Die Prozedur "Hierarchische Clusteranalyse" kann dies automatisch durchführen.)

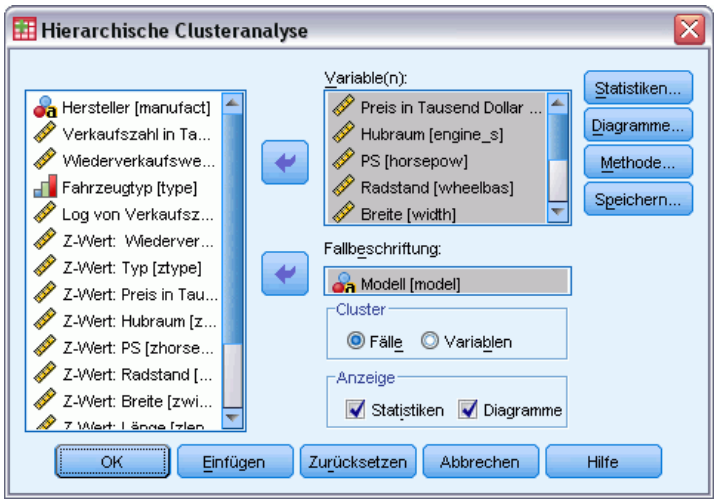
Fallreihenfolge. Wenn gebundene Distanzen oder Ähnlichkeiten in den Eingabedaten vorliegen (oder beim Verbinden in den aktualisierten Clustern auftreten), ist die resultierende Cluster-Lösung ggf. abhängig von der Reihenfolge der Fälle in der Datei. Prüfen Sie daher die Stabilität einer bestimmten Lösung, indem Sie verschiedene Lösungen abrufen, bei denen die Fälle in einer unterschiedlichen, zufällig ausgewählten Reihenfolge sortiert sind.

Annahmen. Die verwendeten Distanz- und Ähnlichkeitsmaße müssen für die analysierten Daten geeignet sein. Weitere Informationen zur Auswahl der Distanz- und Ähnlichkeitsmaße finden Sie unter der Prozedur "Ähnlichkeiten". Außerdem sollten Sie alle relevanten Variablen in Ihre Analyse einschließen. Das Weglassen einflußreicher Variablen kann zu irreführenden Lösungen führen. Da es sich bei der hierarchischen Clusteranalyse um eine explorative Methode handelt, sollten die Ergebnisse als vorläufig gelten, bis diese durch eine unabhängige Stichprobe bestätigt werden.

So führen Sie eine hierarchische Clusteranalyse durch:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Klassifizieren > Hierarchische Cluster...

Abbildung 25-1
Dialogfeld "Hierarchische Clusteranalyse"

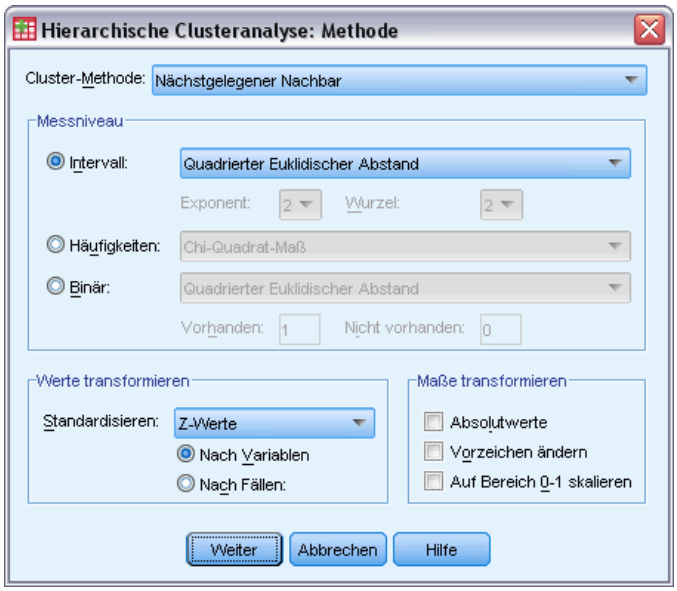


- Beim Clustern von Fällen müssen Sie mindestens eine numerische Variable auswählen. Beim Clustern von Variablen müssen Sie mindestens drei numerische Variablen auswählen.

  Sie haben auch die Möglichkeit, eine Variable für die Beschriftung der Fälle auszuwählen.

Hierarchische Clusteranalyse: Methode

Abbildung 25-2
Dialogfeld "Hierarchische Clusteranalyse: Methode"



Cluster-Methode. Verfügbar sind Linkage zwischen den Gruppen, Linkage innerhalb der Gruppen, nächstgelegener Nachbar, entferntester Nachbar, Zentroid-Clustering, Median-Clustering und die Ward-Methode.

Maß. Hiermit können Sie das Distanz- oder Ähnlichkeitsmaß bestimmen, das beim Clustern verwendet wird. Wählen Sie den Typ der Daten sowie das geeignete Distanz- oder Ähnlichkeitsmaß aus.

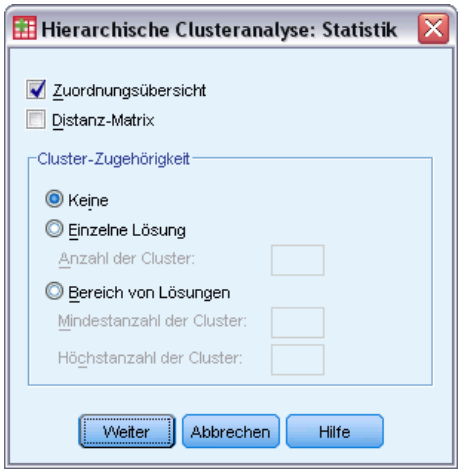
- **Intervall.**Verfügbar sind euklidische Distanz, quadrierte euklidische Distanz, Kosinus, Pearson-Korrelation, Tschebyscheff, Block, Minkowski und die Option Benutzerdefiniert.
- **Häufigkeiten.**Verfügbar sind Chi-Quadratmaß und Phi-Quadratmaß.
- **Binär.**Verfügbar sind euklidische Distanz, quadrierte euklidische Distanz, Größendifferenz, Musterdifferenz, Varianz, Streuung, Form, einfache Übereinstimmung, Phi-4-Punkt-Korrelation, Lambda, Anderberg-*D*, Würfel, Hamann, Jaccard, Kulczynski 1, Kulczynski 2, Distanzmaß nach Lance und Williams, Ochiai, Ähnlichkeitsmaß nach Rogers und Tanimoto, Russel und Rao, Ähnlichkeitsmaße nach Sokal und Sneath 1 bis 5, Yule-*Y* und Yule-*Q*.

Werte transformieren. Hier können Sie festlegen, ob die Datenwerte für Fälle oder Werte vor dem Berechnen von Ähnlichkeiten standardisiert werden (nicht für binäre Daten verfügbar). Die verfügbaren Standardisierungsmethoden sind "*Z*-Scores", "Bereich −1 bis 1", "Bereich 0 bis 1", "Maximale Größe von 1", "Mittelwert 1" und "Standardabweichung 1".

Maße transformieren. Hier können Sie festlegen, ob die durch das Distanzmaß erzeugten Werte transformiert werden. Dies erfolgt, nachdem das Distanzmaß berechnet wurde. Zu den verfügbaren Alternativen zählen Absolutwerte, Ändern des Vorzeichens und Skalieren auf den Bereich 0–1.

Hierarchische Clusteranalyse: Statistik

Abbildung 25-3
Dialogfeld "Hierarchische Clusteranalyse: Statistik"



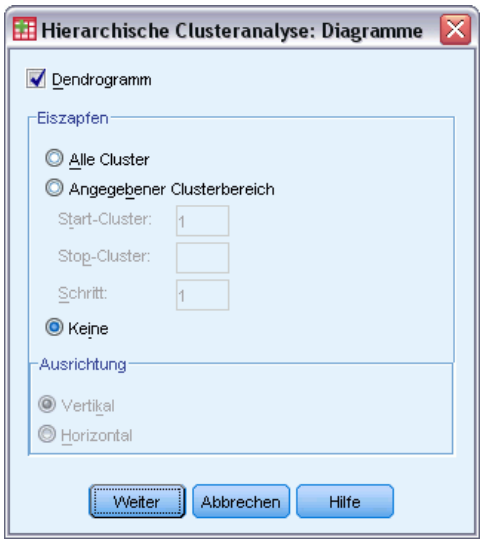
Zuordnungsübersicht. Hier wird folgendes angezeigt: Welche Fälle bzw. Cluster in jedem Schritt kombiniert wurden, die Abstände zwischen den Fällen oder Clustern, die kombiniert werden, und der Cluster-Schritt, in dem ein Fall (oder eine Variable) in den Cluster aufgenommen wurde.

Distanz-Matrix. Zeigt die Distanzen oder Ähnlichkeiten zwischen den Objekten.

Cluster-Zugehörigkeit. Zeigt den Cluster an, dem alle Fälle beim Kombinieren der Cluster in einem oder mehreren Schritten zugeordnet wurden. Die Optionen "Einzelne Lösung" und "Bereich von Lösungen" stehen zur Verfügung.

Hierarchische Clusteranalyse: Diagramme

Abbildung 25-4
Dialogfeld "Hierarchische Clusteranalyse: Grafiken"

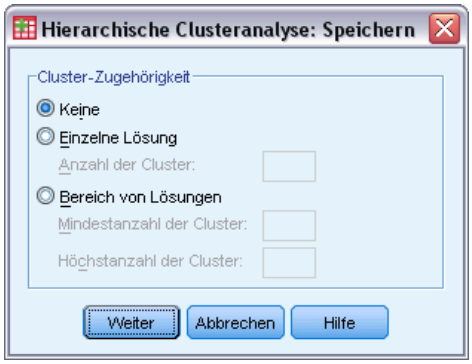


Dendrogramm. Zeigt ein **Dendrogramm** an. Dendrogramme können verwendet werden, um die Dichte der gebildeten Cluster zu bewerten. Sie enthalten Informationen über die angemessene Anzahl der Cluster, die beibehalten werden sollen.

Eiszapfen. Zeigt ein **Eiszapfendiagramm** an, das alle Cluster oder einen bestimmten Bereich von Clustern enthält. Eiszapfendiagramme zeigen an, wie Fälle bei jeder Iteration der Analyse in Clustern zusammengeführt werden. Unter Orientierung können Sie ein vertikales oder horizontales Diagramm auswählen.

Hierarchische Clusteranalyse: Neue Variablen

Abbildung 25-5
Dialogfeld "Hierarchische Clusteranalyse: Neue Variablen speichern"



Cluster-Zugehörigkeit. Hiermit können Sie die Cluster-Zugehörigkeit für eine einzelne Lösung oder einen Bereich von Lösungen speichern. Die gespeicherten Variablen können dann in nachfolgenden Analysen verwendet werden, um andere Differenzen zwischen Gruppen zu untersuchen.

Zusätzliche Funktionen beim Befehl CLUSTER

In der Prozedur "Hierarchische Clusteranalyse" wird die Befehlssyntax von `CLUSTER` verwendet. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Verwenden mehrerer Cluster-Methoden in einer einzigen Analyse
- Einlesen und Analysieren einer Distanzmatrix
- Schreiben einer Distanzmatrix auf die Festplatte für eine spätere Analyse
- Angeben aller Werte für den Exponenten und die Wurzel im benutzerdefinierten (exponentiellen) Distanzmaß
- Festlegen der Namen für gespeicherte Variablen

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
26

Clusterzentrenanalyse

Diese Prozedur kann relativ homogene Fallgruppen aufgrund ausgewählter Eigenschaften identifizieren, wobei ein Algorithmus verwendet wird, der eine große Anzahl von Fällen verarbeiten kann. Der Algorithmus erfordert jedoch, dass Sie die Anzahl der Cluster festlegen. Wenn Ihnen die anfänglichen Clusterzentren bekannt sind, können Sie diese angeben. Sie können eine der beiden Methoden zur Klassifikation der Fälle auswählen, entweder iteratives Aktualisieren der Clusterzentren oder nur Klassifizieren. Sie können Cluster-Zugehörigkeit, Informationen zur Distanz und endgültige Clusterzentren speichern. Wahlweise können Sie eine Variable festlegen, mit deren Werte fallweise Ausgaben beschriftet werden. Sie können außerdem eine *F*-Statistik zur Varianzanalyse anfordern. Während es sich bei dieser Statistik um eine opportunistische Statistik handelt (mit dieser Prozedur wird versucht, tatsächlich voneinander abweichende Gruppen zu bilden), lassen sich aus der relativen Größe der Statistik Informationen über den Beitrag jeder Variablen zu der Trennung der Gruppen gewinnen.

Beispiel. Wodurch können Gruppen von Fernseh-Shows identifiziert werden, die innerhalb jeder Gruppe ein ähnliches Publikum anziehen? Mit der Clusterzentrenanalyse könnten Sie Fernseh-Shows (Fälle) anhand der Merkmale der Zuschauer in *k* homogene Gruppen clustern. Damit lassen sich beispielsweise Marktsegmente identifizieren. Sie können außerdem Städte (Fälle) in homogene Gruppen clustern, sodass vergleichbare Städte zum Testen verschiedener Marketingstrategien ausgewählt werden können.

Statistiken. Vollständige Lösung: anfängliche Clusterzentren, ANOVA-Tabelle. Jeder Fall: Cluster-Informationen, Distanz vom Clusterzentrum.

Daten. Die Variablen müssen quantitativ sein, entweder auf dem Intervall- oder Verhältnisniveau. Wenn Ihre Variablen binär sind oder Häufigkeiten darstellen, verwenden Sie die Prozedur "Hierarchische Clusteranalyse".

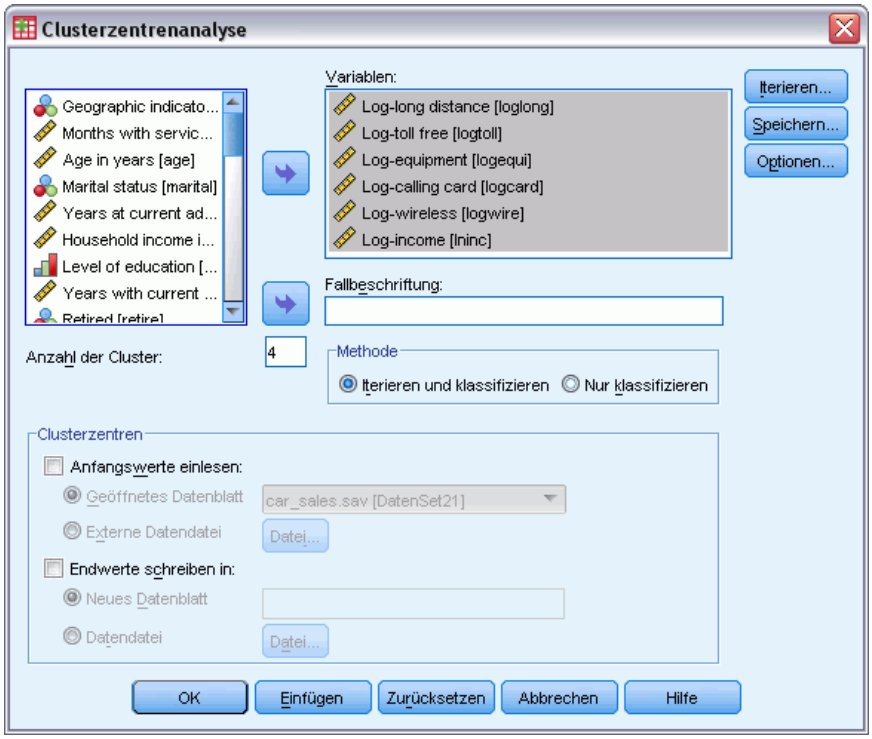
Reihenfolge der Fälle und der anfänglichen Clusterzentren. Der Standardalgorithmus zum Auswählen der anfänglichen Clusterzentren ist nicht invariant bezüglich der Fallreihenfolge. Mit der Option Gleitende Mittelwerte verwenden im Dialogfeld "Iterieren" wird die resultierende Lösung potenziell abhängig von der Reihenfolge der Fälle, unabhängig davon, auf welche Weise die anfänglichen Clusterzentren ausgewählt wurden. Wenn Sie eine dieser Methoden nutzen, prüfen Sie daher die Stabilität einer bestimmten Lösung, indem Sie verschiedene Lösungen abrufen, bei denen die Fälle in einer unterschiedlichen, zufällig ausgewählten Reihenfolge sortiert sind. Wenn Sie anfängliche Clusterzentren angeben und dabei nicht die Option Gleitende Mittelwerte verwenden aktivieren, vermeiden Sie so potentielle Probleme im Zusammenhang mit der Fallreihenfolge. Die Reihenfolge der anfänglichen Clusterzentren kann sich jedoch auf die Lösung auswirken, wenn gebundene Distanzen von Fällen zu Clusterzentren vorliegen. Um die Stabilität einer bestimmten Lösung zu bewerten, können Sie die Ergebnisse von Analysen mit verschiedenen Permutationen der Zentrumsanfangswerte vergleichen.

Annahmen. Distanzen werden unter Verwendung des einfachen euklidischen Abstands berechnet. Wenn Sie ein anderes Distanz- oder Ähnlichkeitsmaß verwenden möchten, verwenden Sie die Prozedur "Hierarchische Clusteranalyse". Die Skalierung der Variablen ist eine wichtige Überlegung. Wenn Ihre Variablen auf unterschiedlichen Skalen gemessen wurden (wenn zum Beispiel eine Variable in Dollar und eine andere in Jahren ausgedrückt wird), können die Ergebnisse irreführend sein. In solchen Fällen sollten Sie eine Standardisierung Ihrer Variablen in Betracht ziehen, bevor Sie die Clusterzentrenanalyse durchführen (mit der Prozedur "Deskriptive Statistiken"). Diese Prozedur setzt voraus, dass Sie die passende Anzahl von Clustern ausgewählt und alle relevanten Variablen eingeschlossen haben. Wenn Sie eine ungeeignete Anzahl von Clustern ausgewählt oder wichtige Variablen ausgelassen haben, können Ihre Ergebnisse irreführend sein.

So lassen Sie eine Clusterzentrenanalyse berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Klassifizieren > Clusterzentrenanalyse...

Abbildung 26-1
Dialogfeld "Clusterzentrenanalyse"



- Wählen Sie die Variablen für die Clusteranalyse aus.
- Legen Sie die Anzahl der Cluster fest. (Die Anzahl der Cluster muss mindestens 2 betragen und darf nicht größer als die Anzahl der Fälle in der Datendatei sein.)
- Wählen Sie als Methode entweder Iterieren und klassifizieren oder Nur klassifizieren.
- Wählen Sie optional eine Identifizierungsvariable zum Beschriften der Fälle aus.

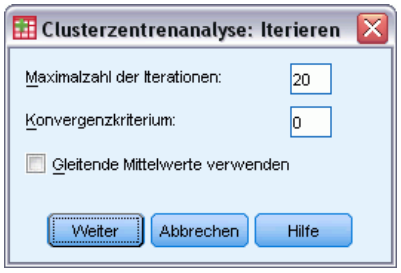
Clusterzentrenanalyse: Effizienz

Der Befehl "Clusterzentrenanalyse" ist in erster Linie deshalb so effizient, weil er nicht die Distanzen zwischen allen Paaren von Fällen berechnet. Dies wird in vielen Algorithmen zum Clustern, auch beim hierarchischen Clustern, durchgeführt.

Für größtmögliche Effizienz nehmen Sie eine Stichprobe von Fällen und bestimmen die Clusterzentren mit der Methode Iterieren und klassifizieren. Wählen Sie Endwerte schreiben in aus. Stellen Sie anschließend die gesamte Datendatei wieder her und wählen Sie als Methode Nur klassifizieren aus. Wählen Sie Anfangswerte einlesen, um die gesamte Datei anhand der aus der Stichprobe geschätzten Clusterzentren zu klassifizieren. Die Daten können in eine Datei oder in ein Daten-Set geschrieben und aus einer Datei oder einem Daten-Set ausgelesen werden. Daten-Sets sind für die anschließende Verwendung in der gleichen Sitzung verfügbar, werden jedoch nicht als Dateien gespeichert, sofern Sie diese nicht ausdrücklich vor dem Beenden der Sitzung speichern. Die Namen von Daten-Sets müssen den Regeln zum Benennen von Variablen entsprechen.

Clusterzentrenanalyse: Iterieren

Abbildung 26-2
Dialogfeld "Clusterzentrenanalyse: Iterieren"



Hinweis: Diese Optionen sind nur verfügbar, wenn Sie im Dialogfeld "Clusterzentrenanalyse" die Methode Iterieren und klassifizieren auswählen.

Maximalzahl der Iterationen. Begrenzt die Anzahl der Iterationen im Clusterzentren-Algorithmus. Die Iteration wird nach der vorgegebenen Anzahl der Iterationen beendet, auch wenn das Konvergenzkriterium noch nicht erreicht wurde. Diese Zahl muss zwischen 1 und 999 liegen.

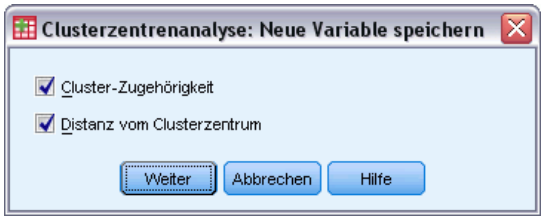
Um den Algorithmus zu verwenden, der beim Befehl "Quick Cluster" in IBM SPSS Statistics-Versionen vor Version 5.0 verwendet wurde, setzen Sie Anzahl der Iterationen auf 1.

Konvergenzkriterium. Bestimmt, wann die Iteration beendet ist. Das Konvergenzkriterium gibt einen Anteil der minimalen Distanz zwischen anfänglichen Clusterzentren wieder. Der Wert muss also größer als 0, darf aber nicht größer als 1 sein. Wenn das Kriterium zum Beispiel 0,02 lautet, ist die Iteration beendet, sobald eine vollständige Iteration keines der Clusterzentren um eine Distanz von mehr als 2 % der kleinsten Distanz zwischen beliebigen anfänglichen Clusterzentren bewegt.

Gleitende Mittelwerte verwenden. Mit dieser Funktion können Sie eine Aktualisierung der Clusterzentren veranlassen, nachdem jeder Fall zugeordnet wurde. Wenn Sie diese Option nicht auswählen, werden neue Clusterzentren berechnet, nachdem alle Fälle zugeordnet wurden.

Clusterzentrenanalyse: Neue Variablen

Abbildung 26-3
Dialogfeld "Clusterzentrenanalyse: Neue Variablen"



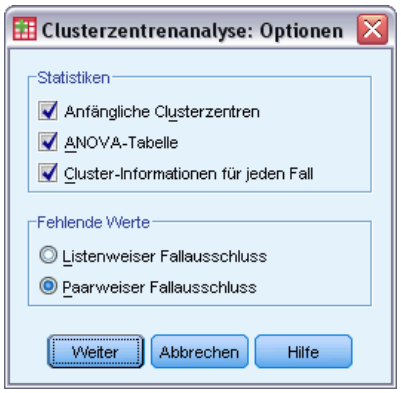
Sie können die Informationen über die Lösung als neue Variablen speichern, um diese in nachfolgenden Analysen zu verwenden:

Cluster-Zugehörigkeit. Erstellt eine neue Variable, welche die endgültige Cluster-Zugehörigkeit für jeden Fall anzeigt. Die Werte der neuen Variablen liegen in einem Bereich von 1 bis zur Anzahl der Cluster.

Distanz vom Clusterzentrum. Erstellt eine neue Variable, welche den euklidischen Abstand zwischen jedem Fall und seinem Klassifikationszentrum anzeigt.

Clusterzentrenanalyse: Optionen

Abbildung 26-4
Dialogfeld "Clusterzentrenanalyse: Optionen"



Statistiken. Sie können die folgenden Statistiken auswählen: anfängliche Clusterzentren, ANOVA-Tabelle und Cluster-Information für jeden Fall.

- **Anfängliche Clusterzentren.** Erster Schätzer der Mittelwerte der Variablen für jeden Cluster. In der Standardeinstellung werden zunächst so viele günstig gelegene Fälle aus den Daten ausgewählt, wie Cluster gebildet werden sollen. Die anfänglichen Clusterzentren werden für eine Ausgangsklassifizierung verwendet und dann aktualisiert.

- **ANOVA-Tabelle.** Zeigt eine Varianzanalysetabelle mit univariaten F-Tests für jede Cluster-Variable an. Die F-Tests haben nur beschreibenden Charakter und die daraus resultierenden Wahrscheinlichkeiten sind nicht zu interpretieren. Die ANOVA-Tabelle wird nicht angezeigt, wenn alle Fälle einem einzigen Cluster zugewiesen werden.
- **Cluster-Informationen für jeden Fall.** Zeigt für jeden Fall die endgültige Clusterzuordnung und den euklidischen Abstand zwischen dem Fall und dem Clusterzentrum, das zur Klassifizierung des Falles verwendet wird. Es werden auch die euklidischen Abstände zwischen den endgültigen Clusterzentren angezeigt.

Fehlende Werte. Die verfügbaren Optionen sind Listenweiser Fallausschluss oder Paarweiser Fallausschluss.

- **Listenweiser Fallausschluss.**Fälle, bei denen Werte einer beliebigen Clustervariable fehlen, werden aus der Analyse ausgeschlossen.
- **Paarweiser Fallausschluss.** Die Fälle werden den Clustern auf der Grundlage der aus allen Variablen mit nichtfehlenden Werten berechneten Distanzen zugewiesen.

Zusätzliche Funktionen beim Befehl QUICK CLUSTER

In der Prozedur "Clusterzentrenanalyse" wird die Befehlssyntax von `QUICK CLUSTER` verwendet. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Übernehmen der ersten *k* Fälle als anfängliche Clusterzentren. Dadurch wird der üblicherweise für deren Schätzung benötigte Verarbeitungsdurchlauf vermieden.
- Direktes Angeben der anfänglichen Clusterzentren als Teil der Befehlssyntax
- Festlegen der Namen für gespeicherte Variablen

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
27

Nichtparametrische Tests

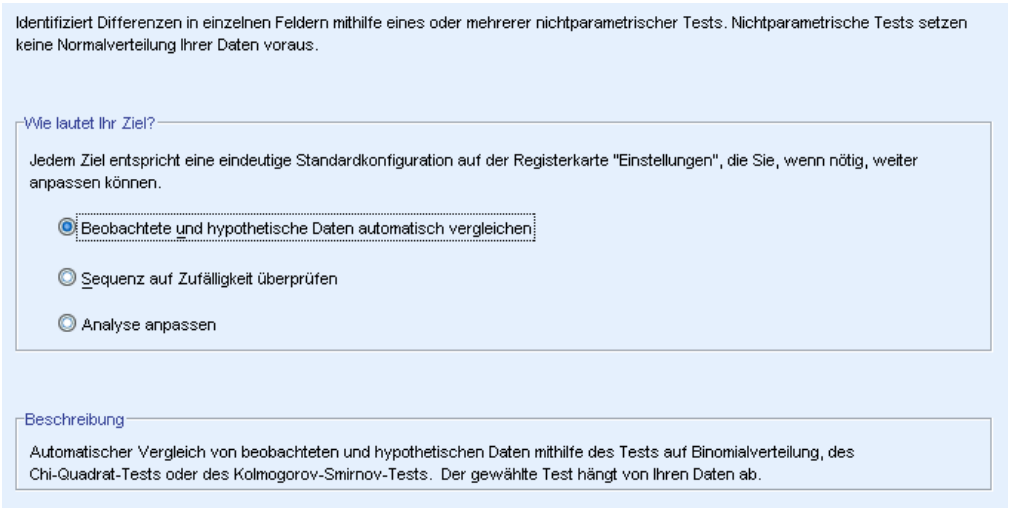
Nichtparametrische Tests machen minimale Annahmen über die zugrunde liegende Verteilung der Daten. Die in diesen Dialogfeldern verfügbaren Tests können anhand dessen, wie die Daten organisiert sind, in drei große Kategorien gruppiert werden:

- Ein Test bei einer Stichprobe analysiert ein Feld.
- Ein Test bei verbundenen Stichproben vergleicht zwei oder mehr Felder für das gleiche Fall-Set.
- Ein Test bei unabhängigen Stichproben analysiert ein Feld, das durch Kategorien eines anderen Felds gruppiert wurde.

Nichtparametrische Tests bei einer Stichprobe

Nichtparametrische Tests bei einer Stichprobe identifizieren Unterschiede in einzelnen Feldern mithilfe von einem oder mehreren nichtparametrischen Tests. Nichtparametrische Tests setzen keine Normalverteilung Ihrer Daten voraus.

Abbildung 27-1
Registerkarte "Ziel" in nichtparametrischen Tests bei einer Stichprobe



Wie lautet Ihr Ziel? Mit den Zielen können Sie schnell unterschiedliche, aber häufig genutzte Testeinstellungen angeben.

- **Beobachtete und hypothetische Daten automatisch vergleichen** Dieses Ziel wendet den Test auf Binomialverteilung auf kategoriale Felder mit nur zwei Kategorien, den Chi-Quadrat-Test auf alle anderen kategorialen Felder und den Kolmogorov-Smirnov-Test auf stetige Felder an.

- **Sequenz auf Zufälligkeit überprüfen** Dieses Ziel verwendet den Sequenztest, um die beobachtete Sequenz der Datenwerte auf Zufälligkeit zu prüfen.
- **Analyse anpassen** Wählen Sie diese Option, wenn Sie die Testeinstellungen auf der Registerkarte “Einstellungen” manuell ändern wollen. Beachten Sie, dass diese Einstellung automatisch ausgewählt wird, wenn Sie anschließend Änderungen auf der Registerkarte “Einstellungen” vornehmen, die mit dem aktuell ausgewählten Ziel nicht kompatibel sind.

So lassen Sie nichtparametrische Tests bei einer Stichprobe berechnen:

Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Nichtparametrische Tests > Eine Stichprobe...

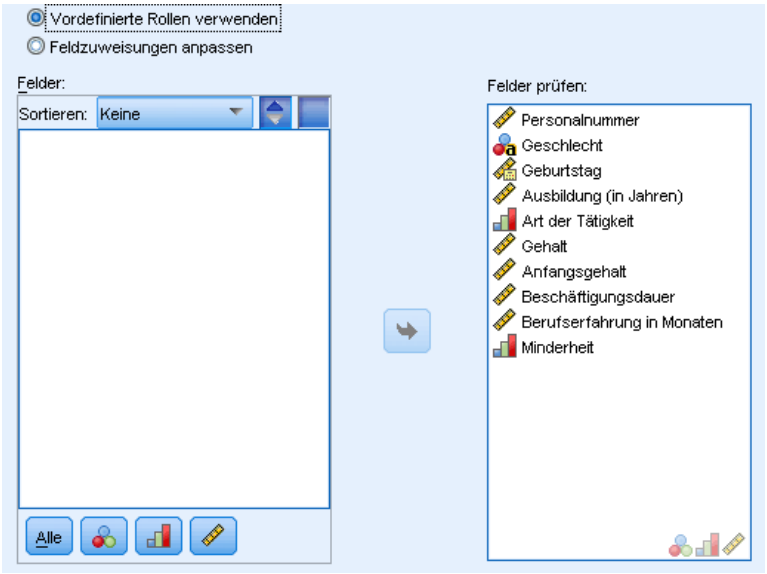
▶ Klicken Sie auf Ausführen.

Die folgenden Optionen sind verfügbar:

- Geben Sie ein Ziel auf der Registerkarte “Ziel” an.
- Geben Sie Feldzuweisungen auf der Registerkarte “Felder” an.
- Geben Sie Experteneinstellungen auf der Registerkarte “Einstellungen” an.

Registerkarte “Felder”

Abbildung 27-2
Registerkarte “Felder” in nichtparametrischen Tests bei einer Stichprobe



Die Registerkarte “Felder” gibt an, welche Felder getestet werden sollen.

Vordefinierte Rollen verwenden Diese Option greift auf bestehende Feldinformationen zurück. Alle Felder mit der vordefinierten Rolle “Eingabe”, “Ziel” oder “Beide” werden als Testfelder verwendet. Mindestens ein Testfeld ist erforderlich.

Benutzerdefinierte Feldzuweisungen verwenden Mit dieser Option können Sie Feldrollen überschreiben. Geben Sie nach Auswahl dieser Option die unten aufgeführten Felder an:

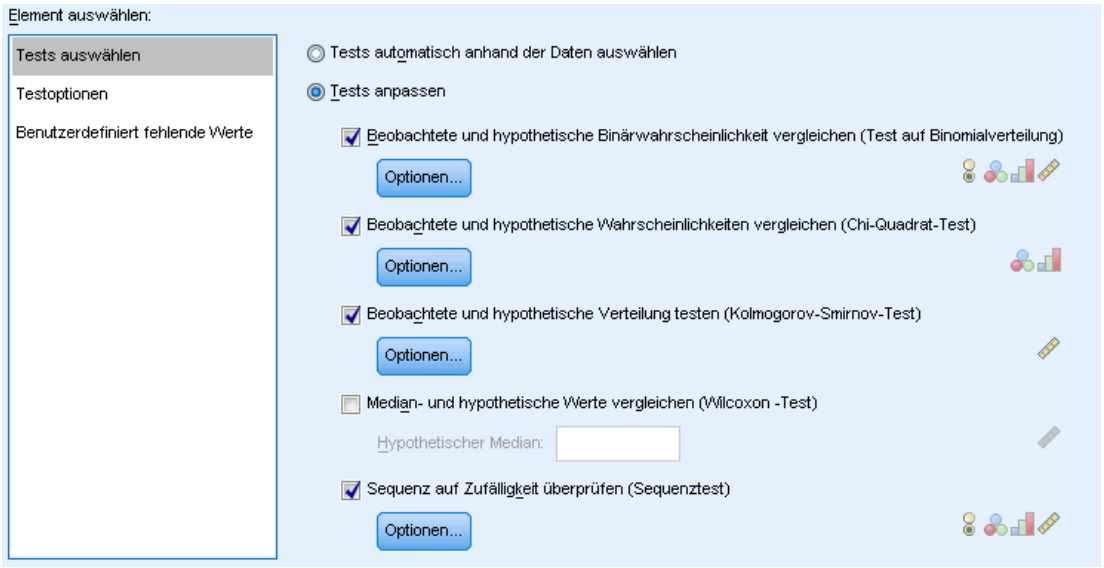
- **Testfelder.** Wählen Sie mindestens ein Feld aus.

Registerkarte "Einstellungen"

Die Registerkarte "Einstellungen" enthält mehrere unterschiedliche Gruppen von Einstellungen, die Sie ändern können, um genau festzulegen, wie der Algorithmus Ihre Daten verarbeiten soll. Wenn Sie an den Standardeinstellungen Änderungen vornehmen, die mit den aktuell ausgewählten Zielen nicht kompatibel sind, wird auf der Registerkarte "Ziel" automatisch die Option Analyse anpassen ausgewählt.

Tests auswählen

Abbildung 27-3
Einstellungen "Tests auswählen" in nichtparametrischen Tests bei einer Stichprobe



Diese Einstellungen geben die Tests an, die in den auf der Registerkarte "Felder" angegebenen Feldern durchgeführt werden.

Tests automatisch anhand der Daten auswählen Diese Einstellung wendet den Test auf Binomialverteilung auf kategoriale Felder mit nur zwei gültigen (nichtfehlenden) Kategorien, den Chi-Quadrat-Test auf alle anderen kategorialen Felder und den Kolmogorov-Smirnov-Test auf stetige Felder an.

Tests anpassen Mit dieser Einstellung können Sie bestimmte Tests auswählen, die durchgeführt werden sollen.

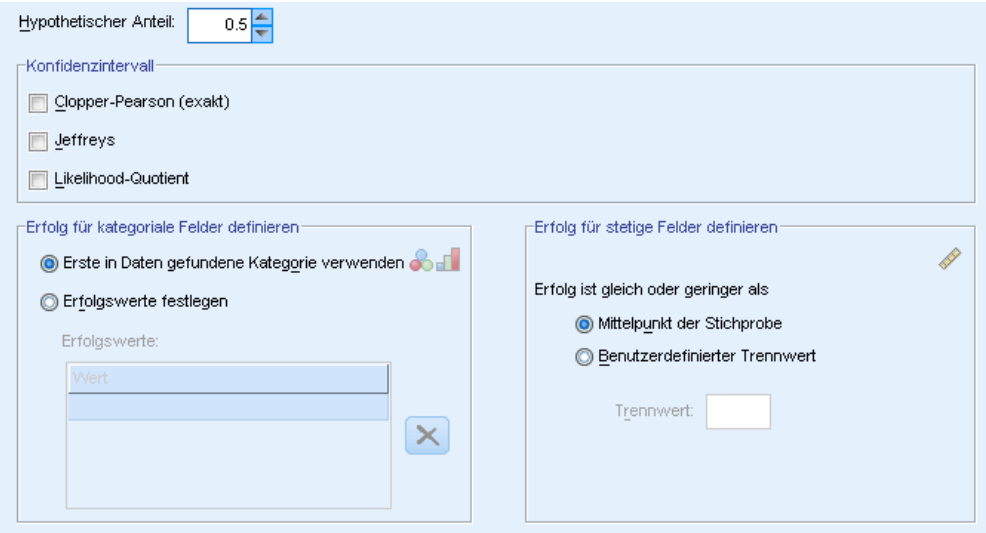
- **Beobachtete und hypothetische Binärwahrscheinlichkeit vergleichen (Test auf Binomialverteilung)** Der Test auf Binomialverteilung kann auf alle Felder angewendet werden. Mit dieser Option wird ein Test bei einer Stichprobe erstellt, der prüft, ob die

beobachtete Verteilung eines Flag-Feldes (ein kategoriales Feld mit nur zwei Kategorien) mit der erwarteten angegebenen Binomialverteilung übereinstimmt. Sie können außerdem Konfidenzintervalle anfordern. Unter Optionen für den Test auf Binomialverteilung finden Sie weitere Informationen über die Testeinstellungen.

- **Beobachtete und hypothetische Wahrscheinlichkeiten vergleichen (Chi-Quadrat-Test)** Der Chi-Quadrat-Test wird auf nominale und ordinale Felder angewendet. Mit dieser Option wird ein Test bei einer Stichprobe erstellt, der eine Chi-Quadrat-Statistik auf der Basis der Unterschiede zwischen den beobachteten und erwarteten Häufigkeiten an Kategorien eines Feldes berechnet. Unter Optionen für den Chi-Quadrat-Test finden Sie weitere Informationen über die Testeinstellungen.
- **Beobachtete und hypothetische Verteilung testen (Kolmogorov-Smirnov-Test)** Der Kolmogorov-Smirnov-Test wird auf stetige Felder angewendet. Mit dieser Option wird ein Test bei einer Stichprobe erstellt, der prüft, ob die kumulative Stichproben-Verteilungsfunktion für ein Feld homogen mit einer Gleich-, Normal-, Poisson- oder Exponentialverteilung ist. Unter Optionen für den Kolmogorov-Smirnov-Test finden Sie weitere Informationen über die Testeinstellungen.
- **Median- und hypothetische Werte vergleichen (Wilcoxon-Test)** Der Wilcoxon-Test wird auf stetige Felder angewendet. Mit dieser Option wird ein Test bei einer Stichprobe des Medianwerts eines Feldes erstellt. Geben Sie eine Zahl als hypothetischen Median an.
- **Sequenz auf Zufälligkeit überprüfen (Sequenztest)** Der Sequenztest wird auf alle Felder angewendet. Mit dieser Option wird ein Test bei einer Stichprobe erstellt, der prüft, ob die Sequenz der Werte eines dichotomisierten Feldes zufällig ist. Unter Optionen für den Sequenztest finden Sie weitere Informationen über die Testeinstellungen.

Optionen für den Test auf Binomialverteilung

Abbildung 27-4
Optionen für den Test auf Binomialverteilung in nichtparametrischen Tests bei einer Stichprobe



Der Test auf Binomialverteilung ist für Flag-Felder gedacht (kategoriale Felder mit nur zwei Kategorien), wird aber auf alle Felder angewendet, indem Regeln zur Definition von "Erfolg" aufgestellt werden.

Hypothetischer Anteil. Gibt den erwarteten Anteil der als "Erfolge" definierten Datensätze oder *p* an. Geben Sie einen Wert größer 0 und kleiner 1 ein. Der Standardwert ist 0,5.

Konfidenzintervall. Zur Berechnung von Konfidenzintervallen für binäre Daten stehen folgende Verfahren zur Verfügung:

- **Clopper-Pearson (exakt).** Ein exaktes Intervall auf der Basis der kumulativen Binomialverteilung.
- **Jeffreys.** Ein Bayes-Intervall auf der Basis der A-posteriori-Verteilung von *p* mithilfe des Jeffreys-Vorrangs.
- **Likelihood-Quotient.** Ein Intervall auf der Basis der Likelihood-Funktion für *p*.

Erfolg für kategoriale Felder definieren Gibt an, wie "Erfolg", der/die gegen den hypothetischen Anteil getestete(n) Datenwert(e), für kategoriale Felder definiert wird.

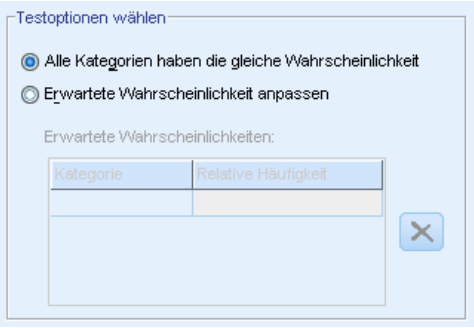
- Erste in Daten gefundene Kategorie verwenden führt den Test auf Binomialverteilung mithilfe des ersten in der Stichprobe gefundenen Werts durch, um "Erfolg" zu definieren. Diese Option ist nur für nominale oder ordinale Felder mit nur zwei Werten verfügbar; alle anderen in der Registerkarte "Felder" angegebenen kategorialen Felder, in denen diese Option verwendet wird, werden nicht getestet. Dies ist die Standardeinstellung.
- Erfolgswerte festlegen führt den Test auf Binomialverteilung mithilfe der angegebenen Werteliste durch, um "Erfolg" zu definieren. Geben Sie eine Liste von String- oder numerischen Werten an. Die Werte in der Liste müssen nicht in der Stichprobe vorhanden sein.

Erfolg für stetige Felder definieren Gibt an, wie "Erfolg", der/die gegen den Testwert getestete(n) Datenwert(e), für stetige Felder definiert wird. Erfolg wird in Form von Werten definiert, die kleiner oder gleich einem Trennwert sind.

- Mittelpunkt der Stichprobe setzt den Trennwert auf den durchschnittlichen Mindest- oder Höchstwert.
- Mit Trennwert anpassen können Sie einen eigenen Trennwert bestimmen.

Optionen für den Chi-Quadrat-Test

Abbildung 27-5
Optionen für den Chi-Quadrat-Test in nichtparametrischen Tests bei einer Stichprobe



Alle Kategorien haben die gleiche Wahrscheinlichkeit. Mit dieser Option werden unter allen Kategorien in der Stichprobe gleiche Häufigkeiten erstellt. Dies ist die Standardeinstellung.

Erwartete Wahrscheinlichkeit anpassen. Mit dieser Option können Sie für eine bestimmte Liste von Kategorien ungleiche Häufigkeiten angeben. Geben Sie eine Liste von String- oder numerischen Werten an. Die Werte in der Liste müssen nicht in der Stichprobe vorhanden sein. Geben Sie in der Spalte Kategorie Kategoriewerte an. Geben Sie in der Spalte Relative Häufigkeit einen Wert größer als 0 für jede Kategorie ein. Benutzerdefinierte Häufigkeiten werden als Verhältnisse behandelt, damit zum Beispiel die Angabe der Häufigkeiten 1, 2 und 3 der Angabe der Häufigkeiten 10, 20 und 30 entspricht und beide angeben, dass von 1/6 der Datensätze erwartet wird, dass sie in die erste Kategorie fallen, 1/3 in die zweite und 1/2 in die dritte. Wenn benutzerdefinierte erwartete Wahrscheinlichkeiten angegeben werden, müssen die benutzerdefinierten Kategoriewerte alle Feldwerte in den Daten enthalten, sonst wird der Test für dieses Feld nicht durchgeführt.

Optionen für den Kolmogorov-Smirnov-Test

Abbildung 27-6
Optionen für den Kolmogorov-Smirnov-Test in nichtparametrischen Tests bei einer Stichprobe

Dieses Dialogfeld gibt an, welche Verteilungen getestet werden sollten, sowie die Parameter der hypothetischen Verteilungen.

Normalverteilung Stichprobendaten verwenden verwendet den beobachteten Mittelwert und die Standardabweichung, mit Benutzerdefiniert können Sie eigene Werte bestimmen.

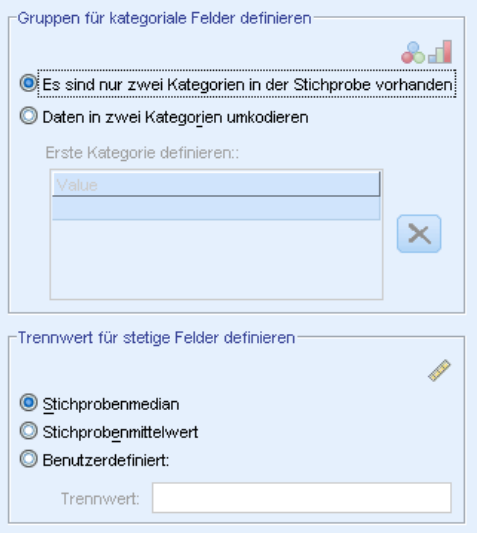
Gleichverteilung Stichprobendaten verwenden verwendet den beobachteten Mindest- und Höchstwert, mit Benutzerdefiniert können Sie eigene Werte bestimmen.

Exponentialverteilung Stichprobenmittelwert verwendet den beobachteten Mittelwert, mit Benutzerdefiniert können Sie eigene Werte bestimmen.

Poisson-Verteilung. Stichprobenmittelwert verwendet den beobachteten Mittelwert, mit Benutzerdefiniert können Sie eigene Werte bestimmen.

Optionen für den Sequenztest

Abbildung 27-7
Optionen für den Sequenztest in nichtparametrischen Tests bei einer Stichprobe



Der Sequenztest ist für Flag-Felder gedacht (kategoriale Felder mit nur zwei Kategorien), kann aber auf alle Felder angewendet werden, indem Regeln zur Definition der Gruppen aufgestellt werden.

Gruppen für kategoriale Felder definieren

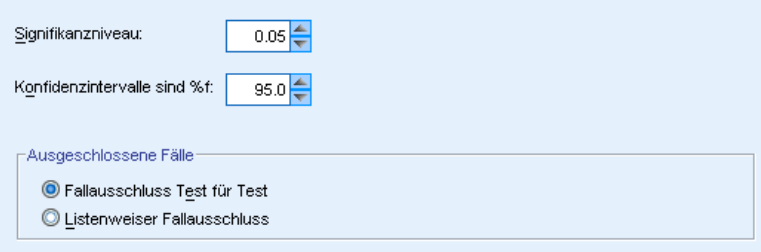
- Es sind nur zwei Kategorien in der Stichprobe vorhanden führt den Sequenztest mithilfe der in der Stichprobe gefundenen Daten durch, um die Gruppen zu definieren. Diese Option ist nur für nominale oder ordinale Felder mit nur zwei Werten verfügbar; alle anderen in der Registerkarte “Felder” angegebenen kategorialen Felder, in denen diese Option verwendet wird, werden nicht getestet.
- Daten in zwei Kategorien umkodieren führt den Sequenztest mithilfe der angegebenen Werteliste durch, um eine Gruppe zu definieren. Alle anderen Werte in der Stichprobe definieren die andere Gruppe. Nicht alle Werte in der Liste müssen in der Stichprobe vorhanden sein, aber es muss mindestens ein Datensatz in jeder Gruppe vorhanden sein.

Trennwert für stetige Felder definieren. Gibt an, wie Gruppen für stetige Felder definiert werden. Die erste Gruppe wird in Form von Werten definiert, die kleiner oder gleich einem Trennwert sind.

- Stichprobenmedian setzt den Trennwert auf den Stichprobenmedian.
- Stichprobenmittelwert setzt den Trennwert auf den Stichprobenmittelwert.
- Mit Benutzerdefiniert können Sie einen eigenen Trennwert bestimmen.

Testoptionen

Abbildung 27-8
Einstellungen "Testoptionen" in nichtparametrischen Tests bei einer Stichprobe



Signifikanzniveau. Gibt das Signifikanzniveau (Alpha) für alle Tests an. Geben Sie einen numerischen Wert zwischen 0 und 1 an. 0,05 ist die Standardeinstellung.

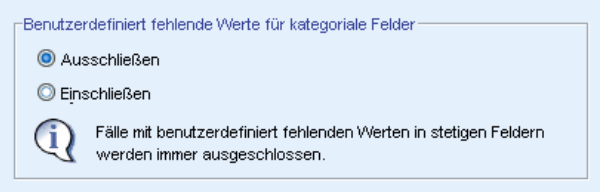
Konfidenzintervall (%). Gibt das Konfidenzniveau für alle erstellten Konfidenzintervalle an. Geben Sie einen numerischen Wert zwischen 0 und 100 an. 95 ist die Standardeinstellung.

Ausgeschlossene Fälle. Gibt an, wie die Fallbasis für Tests bestimmt wird.

- Listenweiser Fallausschluss bedeutet, dass Datensätze mit fehlenden Werten für ein beliebiges Feld, das auf der Registerkarte "Felder" genannt wurde, aus allen Analysen ausgeschlossen werden.
- Fallausschluss Test für Test bedeutet, dass Datensätze mit fehlenden Werten für ein Feld, das für einen bestimmten Test verwendet wird, aus diesem Test ausgeschlossen werden. Wenn in der Analyse mehrere Tests angegeben wurden, wird jeder Test getrennt ausgewertet.

Benutzerdefiniert fehlende Werte

Abbildung 27-9
Einstellungen "Benutzerdefiniert fehlende Werte" in nichtparametrischen Tests bei einer Stichprobe

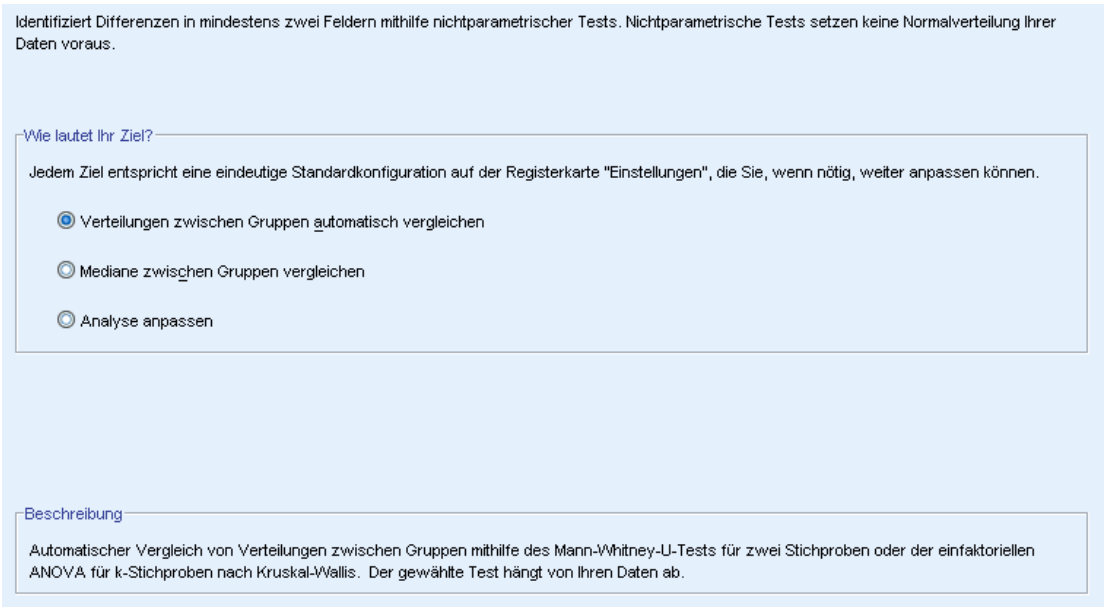


Benutzerdefiniert fehlende Werte für kategoriale Felder Kategoriale Felder müssen gültige Werte für einen Datensatz aufweisen, um in die Analyse aufgenommen zu werden. Mit diesen Steuerungen legen Sie fest, ob benutzerdefiniert fehlende Werte bei den kategorialen Feldern als gültige Werte behandelt werden sollen. Systemdefinierte fehlende Werte und fehlende Werte für stetige Felder werden immer als ungültige Werte behandelt.

Nichtparametrische Tests bei unabhängigen Stichproben

Nichtparametrische Tests bei unabhängigen Stichproben identifizieren Unterschiede zwischen zwei oder mehr Gruppen mithilfe von einem oder mehreren nichtparametrischen Tests. Nichtparametrische Tests setzen keine Normalverteilung Ihrer Daten voraus.

Abbildung 27-10
Registerkarte "Ziel" in nichtparametrischen Tests bei unabhängigen Stichproben



Wie lautet Ihr Ziel? Mit den Zielen können Sie schnell unterschiedliche, aber häufig genutzte Testeinstellungen angeben.

- **Verteilungen zwischen Gruppen automatisch vergleichen** Dieses Ziel wendet den Mann-Whitney-U-Test auf Daten mit zwei Gruppen oder die einfaktorielle ANOVA nach Kruskal-Wallis auf Daten mit *k* Gruppen an.
- **Mediane zwischen Gruppen vergleichen** Dieses Ziel verwendet den Mediantest, um die beobachteten Mediane zwischen Gruppen zu vergleichen.
- **Analyse anpassen** Wählen Sie diese Option, wenn Sie die Testeinstellungen auf der Registerkarte "Einstellungen" manuell ändern wollen. Beachten Sie, dass diese Einstellung automatisch ausgewählt wird, wenn Sie anschließend Änderungen auf der Registerkarte "Einstellungen" vornehmen, die mit dem aktuell ausgewählten Ziel nicht kompatibel sind.

So lassen Sie nichtparametrische Tests bei unabhängigen Stichproben berechnen:

Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Nichtparametrische Tests > Unabhängige Stichproben...

▶ Klicken Sie auf Ausführen.

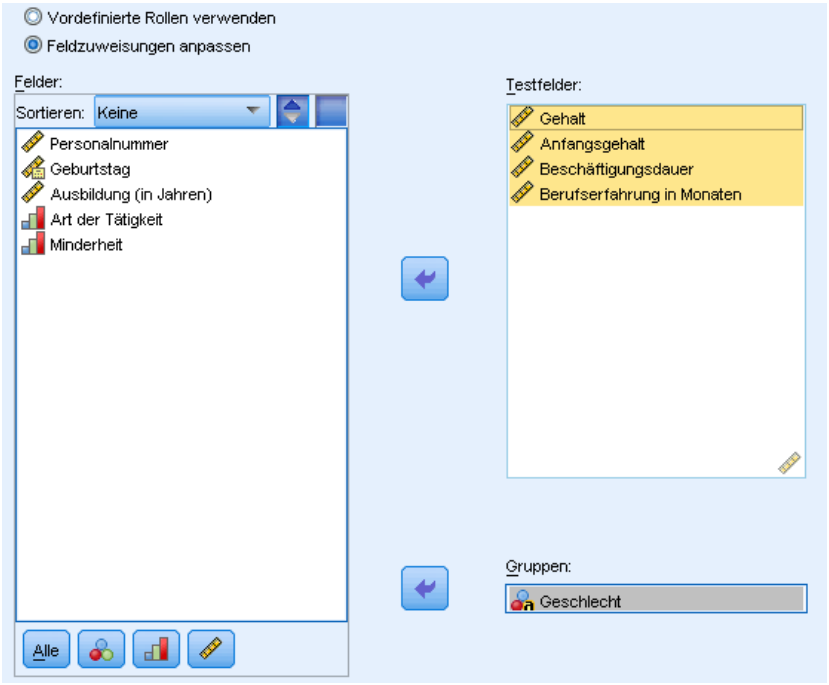
Die folgenden Optionen sind verfügbar:

- Geben Sie ein Ziel auf der Registerkarte "Ziel" an.

- Geben Sie Feldzuweisungen auf der Registerkarte “Felder” an.
- Geben Sie Experteneinstellungen auf der Registerkarte “Einstellungen” an.

Registerkarte “Felder”

Abbildung 27-11
Registerkarte “Felder” in nichtparametrischen Tests bei unabhängigen Stichproben



Die Registerkarte “Felder” gibt an, welche Felder getestet werden sollten, sowie das zur Definition von Gruppen verwendete Feld.

Vordefinierte Rollen verwenden Diese Option greift auf bestehende Feldinformationen zurück. Alle stetigen Felder mit der vordefinierten Rolle “Ziel” oder “Beide” werden als Testfelder verwendet. Falls ein einzelnes kategoriales Feld mit der vordefinierten Rolle “Eingabe” vorhanden ist, wird es als Gruppierungsfeld verwendet. Andernfalls wird standardmäßig kein anderes Gruppierungsfeld verwendet und Sie müssen benutzerdefinierte Feldzuweisungen verwenden. Es ist mindestens ein Testfeld und ein Gruppierungsfeld erforderlich.

Benutzerdefinierte Feldzuweisungen verwenden Mit dieser Option können Sie Feldrollen überschreiben. Geben Sie nach Auswahl dieser Option die unten aufgeführten Felder an:

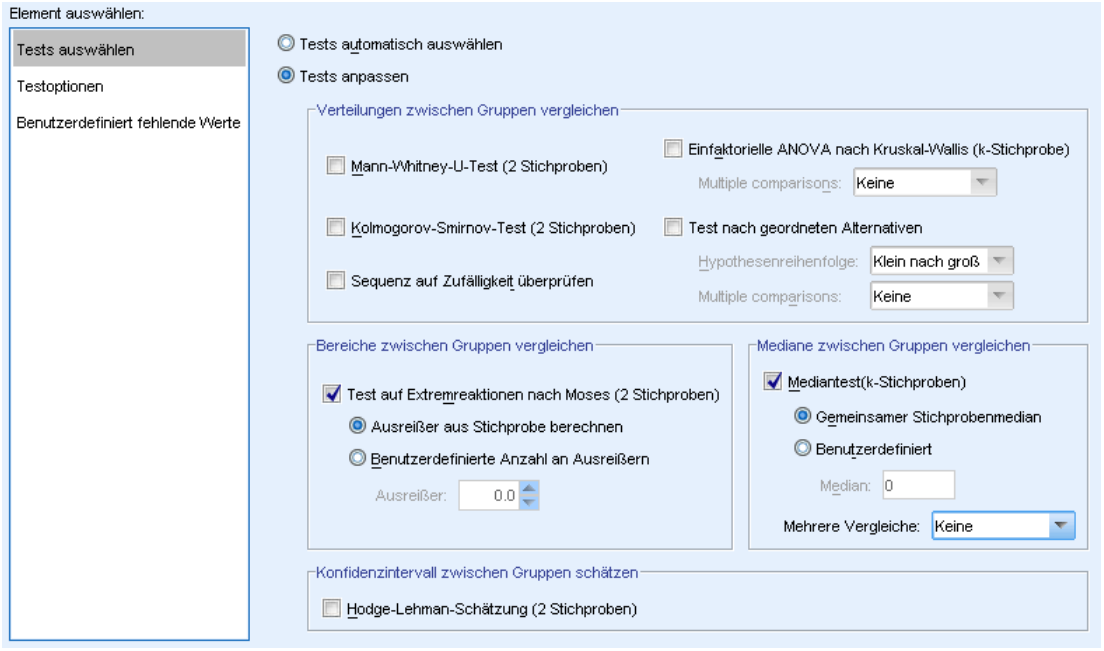
- **Testfelder.** Wählen Sie mindestens ein stetiges Feld aus.
- **Gruppen.** Wählen Sie ein kategoriales Feld aus.

Registerkarte "Einstellungen"

Die Registerkarte "Einstellungen" enthält mehrere unterschiedliche Gruppen von Einstellungen, die Sie ändern können, um genau festzulegen, wie der Algorithmus Ihre Daten verarbeiten soll. Wenn Sie an den Standardeinstellungen Änderungen vornehmen, die mit den aktuell ausgewählten Zielen nicht kompatibel sind, wird auf der Registerkarte "Ziel" automatisch die Option Analyse anpassen ausgewählt.

Tests auswählen

Abbildung 27-12
Einstellungen "Tests auswählen" in nichtparametrischen Tests bei unabhängigen Stichproben



Diese Einstellungen geben die Tests an, die in den auf der Registerkarte "Felder" angegebenen Feldern durchgeführt werden.

Tests automatisch anhand der Daten auswählen Diese Einstellung wendet den Mann-Whitney-U-Test auf Daten mit zwei Gruppen oder die einfaktorielle ANOVA nach Kruskal-Wallis auf Daten mit *k*-Gruppen an.

Tests anpassen Mit dieser Einstellung können Sie bestimmte Tests auswählen, die durchgeführt werden sollen.

- **Verteilungen zwischen Gruppen vergleichen** Damit werden Tests bei unabhängigen Stichproben durchgeführt, um zu testen, ob die Stichproben aus der gleichen Grundgesamtheit stammen.

  Der Mann-Whitney-U-Test (2 Stichproben) verwendet den Rang von jedem Fall, um zu prüfen, ob die Gruppen aus der gleichen Grundgesamtheit gezogen wurden. Der erste Wert im Gruppierungsfeld in aufsteigender Reihenfolge definiert die erste Gruppe und der zweite

definiert die zweite Gruppe. Dieser Test wird nicht durchgeführt, wenn das Gruppierungsfeld mehr als zwei Werte aufweist.

Der Kolmogorov-Smirnov-Test (2 Stichproben) reagiert auf unterschiedliche Mediane, Streuungen, Schiefegrade usw. zwischen den beiden Verteilungen. Dieser Test wird nicht durchgeführt, wenn das Gruppierungsfeld mehr als zwei Werte aufweist.

Bei Sequenz auf Zufälligkeit überprüfen (Wald-Wolfowitz-Test bei 2 Stichproben) wird ein Sequenztest mit Gruppenzugehörigkeit als Kriterium erzeugt. Dieser Test wird nicht durchgeführt, wenn das Gruppierungsfeld mehr als zwei Werte aufweist.

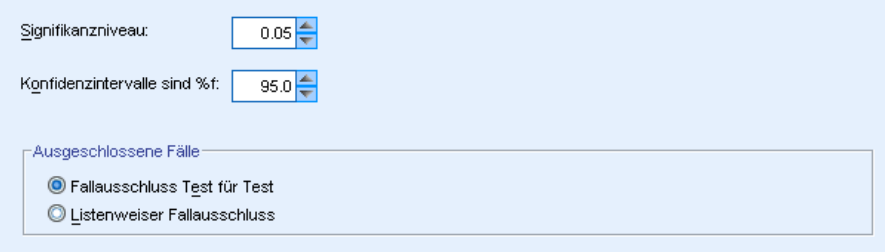
Die Einfaktorielle ANOVA nach Kruskal-Wallis (k-Stichproben) ist eine Erweiterung des Mann-Whitney-U-Tests und der nichtparametrischen Entsprechung der einfaktoriellen Varianzanalyse. Sie können optional Mehrfachvergleiche der *k*-Stichproben anfordern, entweder alle paarweisenMehrfachvergleiche oder schrittweise Step-Down-Vergleiche.

Der Test nach geordneten Alternativen (Jonckheere-Terpstra-Test bei k-Stichproben) ist eine leistungsfähigere Alternative zu Kruskal-Wallis, wenn die *k*-Stichproben eine natürliche Ordnung aufweisen. Die *k* Grundgesamtheiten könnten zum Beispiel *k* ansteigende Temperaturen darstellen. Die Hypothese, dass unterschiedliche Temperaturen die gleiche Verteilung von Antworten erzeugen, wird gegen die Alternative getestet, dass mit Zunahme der Temperatur die Größe der Antwort zunimmt. Hierbei ist die alternative Hypothese geordnet, deshalb ist der Jonckheere-Terpstra-Test für diesen Test am besten geeignet. Geben Sie die Ordnung der alternativen Hypothesen an; Klein nach groß legt eine alternative Hypothese fest, dass der Lageparameter der ersten Gruppe ungleich dem der zweiten Gruppe, der wiederum ungleich dem der dritten Gruppe ist usw.; Groß nach klein legt eine alternative Hypothese fest, dass der Lageparameter der ersten Gruppe ungleich dem der zweitletzten Gruppe ist, der wiederum ungleich dem der drittletzten Gruppe ist usw. Sie können optional Mehrfachvergleiche der *k*-Stichproben anfordern, entweder alle paarweisen Mehrfachvergleiche oder schrittweise Step-Down-Vergleiche.

- **Bereiche zwischen Gruppen vergleichen** Mit dieser Option wird ein Test bei unabhängigen Stichproben erstellt und geprüft, ob die Stichproben den gleichen Bereich aufweisen. Der Test auf Extremreaktionen nach Moses (2 Stichproben) prüft eine Kontrollgruppe gegen eine Vergleichsgruppe. Der erste Wert im Gruppierungsfeld in aufsteigender Reihenfolge definiert die Kontrollgruppe und der zweite definiert die Vergleichsgruppe. Dieser Test wird nicht durchgeführt, wenn das Gruppierungsfeld mehr als zwei Werte aufweist.
- **Mediane zwischen Gruppen vergleichen** Mit dieser Option wird ein Test bei unabhängigen Stichproben erstellt und geprüft, ob die Stichproben den gleichen Median aufweisen. Der Mediantest (k-Stichproben) kann entweder den gemeinsamen Stichprobenmedian (für alle Datensätze im Daten-Set berechnet) oder einen benutzerdefinierten Wert als hypothetischen Median verwenden. Sie können optional Mehrfachvergleiche der *k*-Stichproben anfordern, entweder alle paarweisen Mehrfachvergleiche oder schrittweise Step-Down-Vergleiche.
- **Konfidenzintervalle zwischen Gruppen schätzen** Die Hodges-Lehman-Schätzung (2 Stichproben) erstellt eine Schätzung und ein Konfidenzintervall bei unabhängigen Stichproben für die Differenz in den Medianen der zwei Gruppen. Dieser Test wird nicht durchgeführt, wenn das Gruppierungsfeld mehr als zwei Werte aufweist.

Testoptionen

Abbildung 27-13
Einstellungen "Testoptionen" in nichtparametrischen Tests bei unabhängigen Stichproben



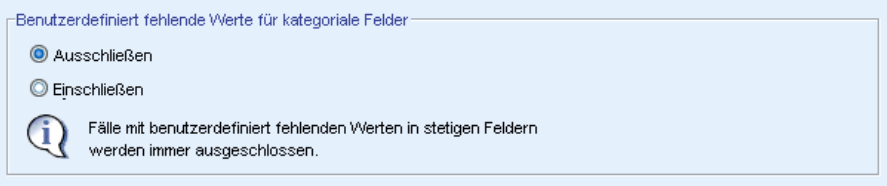
Signifikanzniveau. Gibt das Signifikanzniveau (Alpha) für alle Tests an. Geben Sie einen numerischen Wert zwischen 0 und 1 an. 0,05 ist die Standardeinstellung.

Konfidenzintervall (%). Gibt das Konfidenzniveau für alle erstellten Konfidenzintervalle an. Geben Sie einen numerischen Wert zwischen 0 und 100 an. 95 ist die Standardeinstellung.

Ausgeschlossene Fälle. Gibt an, wie die Fallbasis für Tests bestimmt wird. Listenweiser Fallausschluss bedeutet, dass Datensätze mit fehlenden Werten für ein beliebiges Feld, das in einem beliebigen Unterbefehl genannt wurde, aus allen Analysen ausgeschlossen werden. Fallausschluss Test für Test bedeutet, dass Datensätze mit fehlenden Werten für ein Feld, das für einen bestimmten Test verwendet wird, aus diesem Test ausgeschlossen werden. Wenn in der Analyse mehrere Tests angegeben wurden, wird jeder Test getrennt ausgewertet.

Benutzerdefiniert fehlende Werte

Abbildung 27-14
Einstellungen "Benutzerdefiniert fehlende Werte" in nichtparametrischen Tests bei unabhängigen Stichproben



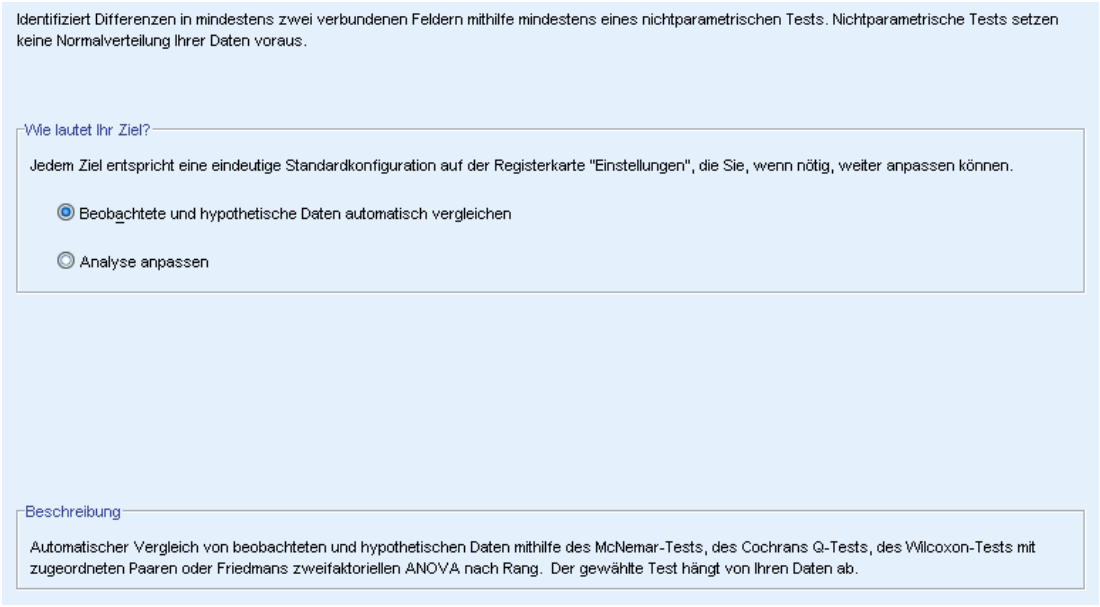
Benutzerdefiniert fehlende Werte für kategoriale Felder Kategoriale Felder müssen gültige Werte für einen Datensatz aufweisen, um in die Analyse aufgenommen zu werden. Mit diesen Steuerungen legen Sie fest, ob benutzerdefiniert fehlende Werte bei den kategorialen Feldern als gültige Werte behandelt werden sollen. Systemdefinierte fehlende Werte und fehlende Werte für stetige Felder werden immer als ungültige Werte behandelt.

Nichtparametrische Tests bei verbundenen Stichproben

Identifiziert Differenzen zwischen mindestens zwei verbundenen Feldern mithilfe mindestens eines nichtparametrischen Tests. Nichtparametrische Tests setzen keine Normalverteilung Ihrer Daten voraus.

Erläuterung der Daten Jeder Datensatz entspricht einem gegebenen Befragten, für den in separaten Feldern im Datensatz zwei oder mehr miteinander verbundene Messungen vorhanden sind. Beispielsweise kann eine Studie zur Wirksamkeit eines Diätplans mit nichtparametrischen Tests bei verbundenen Stichproben analysiert werden, falls das Gewicht jedes Befragten in regelmäßigen Abständen gemessen und in Feldern wie *Gewicht vor Diät*, *Zwischenzeitliches Gewicht* und *Gewicht nach Diät* gespeichert wird. Diese Felder sind "verbunden".

Abbildung 27-15
Registerkarte "Ziel" in nichtparametrischen Tests bei verbundenen Stichproben



Wie lautet Ihr Ziel? Mit den Zielen können Sie schnell unterschiedliche, aber häufig genutzte Testeinstellungen angeben.

- **Beobachtete und hypothetische Daten automatisch vergleichen.** Dieses Ziel wendet den McNemar-Test auf kategoriale Daten bei zwei angegebenen Feldern, Cochrans Q-Test auf kategoriale Daten bei mehr als zwei angegebenen Feldern, den Wilcoxon-Test mit zugeordneten Paaren auf stetige Daten bei zwei angegebenen Feldern und Friedmans zweifaktorielle ANOVA nach Rang auf stetige Daten bei mehr als zwei angegebenen Feldern an.
- **Analyse anpassen** Wählen Sie diese Option, wenn Sie die Testeinstellungen auf der Registerkarte "Einstellungen" manuell ändern wollen. Beachten Sie, dass diese Einstellung automatisch ausgewählt wird, wenn Sie anschließend Änderungen auf der Registerkarte "Einstellungen" vornehmen, die mit dem aktuell ausgewählten Ziel nicht kompatibel sind.

Wenn Felder mit unterschiedlichem Messniveau angegeben werden, werden sie zuerst nach Messniveau getrennt und anschließend wird für jede Gruppe der entsprechende Test durchgeführt. Wenn Sie beispielsweise Beobachtete und hypothetische Daten automatisch vergleichen als Ziel wählen und drei stetige und zwei nominale Felder angeben, wird der Friedman-Test auf die stetigen Felder und der McNemar-Test auf die nominalen Felder angewendet.

So lassen Sie nichtparametrische Tests bei verbundenen Stichproben berechnen:

Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Nichtparametrische Tests > Verbundene Stichproben...

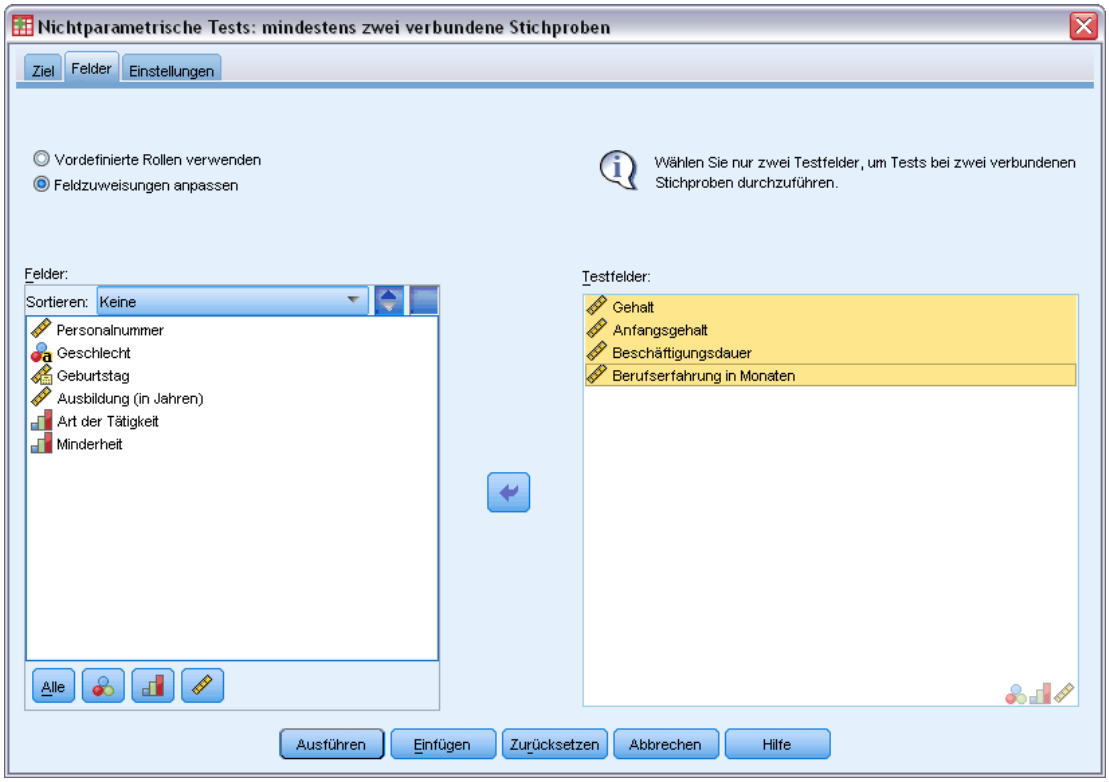
▶ Klicken Sie auf Ausführen.

Die folgenden Optionen sind verfügbar:

- Geben Sie ein Ziel auf der Registerkarte “Ziel” an.
- Geben Sie Feldzuweisungen auf der Registerkarte “Felder” an.
- Geben Sie Experteneinstellungen auf der Registerkarte “Einstellungen” an.

Registerkarte “Felder”

Abbildung 27-16
Registerkarte “Felder” in nichtparametrischen Tests bei verbundenen Stichproben



Die Registerkarte “Felder” gibt an, welche Felder getestet werden sollen.

Vordefinierte Rollen verwenden Diese Option greift auf bestehende Feldinformationen zurück. Alle Felder mit der vordefinierten Rolle “Ziel” oder “Beide” werden als Testfelder verwendet. Mindestens zwei Testfelder sind erforderlich.

Benutzerdefinierte Feldzuweisungen verwenden Mit dieser Option können Sie Feldrollen überschreiben. Geben Sie nach Auswahl dieser Option die unten aufgeführten Felder an:

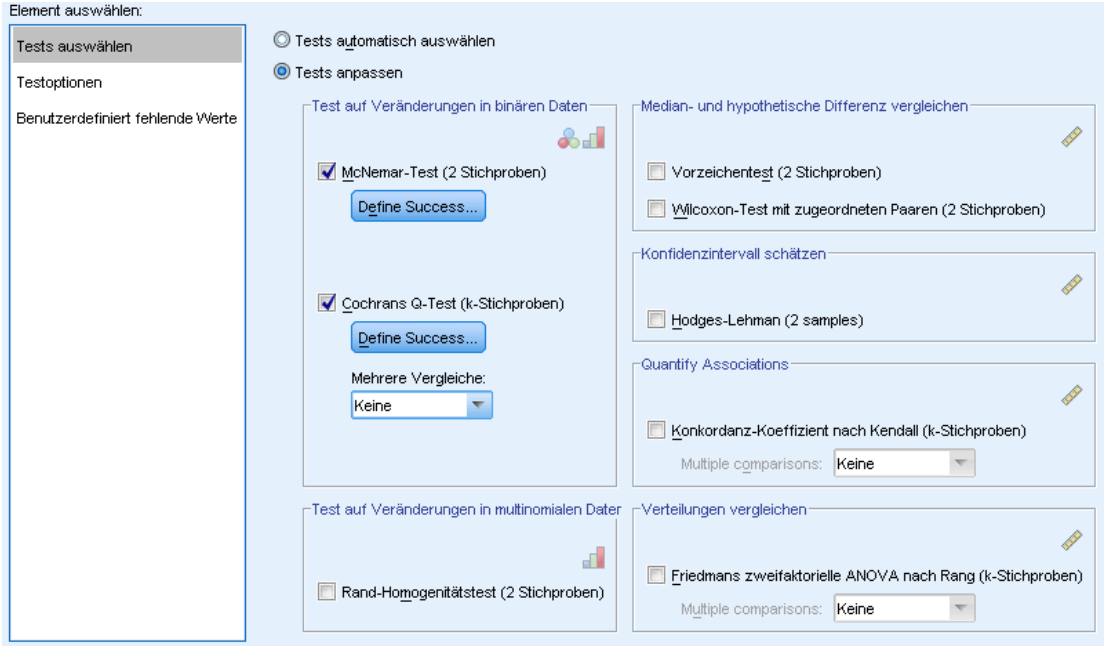
- **Testfelder.** Wählen Sie mindestens zwei Felder aus. Jedes Feld bezieht sich auf eine separate verbundene Stichprobe.

Registerkarte "Einstellungen"

Die Registerkarte "Einstellungen" enthält mehrere unterschiedliche Gruppen von Einstellungen, die Sie ändern können, um genau festzulegen, wie das Verfahren Ihre Daten verarbeiten soll. Wenn Sie an den Standardeinstellungen Änderungen vornehmen, die mit den anderen Zielen nicht kompatibel sind, wird auf der Registerkarte "Ziel" automatisch die Option Analyse anpassen ausgewählt.

Tests auswählen

Abbildung 27-17
Einstellungen "Tests auswählen" in nichtparametrischen Tests bei verbundenen Stichproben



Diese Einstellungen geben die Tests an, die in den auf der Registerkarte "Felder" angegebenen Feldern durchgeführt werden.

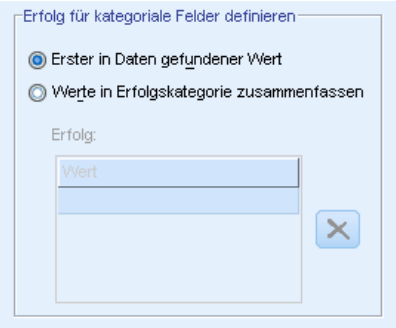
Tests automatisch anhand der Daten auswählen Diese Einstellung wendet den McNemar-Test auf kategoriale Daten bei zwei angegebenen Feldern, Cochrans Q-Test auf kategoriale Daten bei mehr als zwei angegebenen Feldern, den Wilcoxon-Test mit zugeordneten Paaren auf stetige Daten bei zwei angegebenen Feldern und Friedmans zweifaktorielle ANOVA nach Rang auf stetige Daten bei mehr als zwei angegebenen Feldern an.

Tests anpassen Mit dieser Einstellung können Sie bestimmte Tests auswählen, die durchgeführt werden sollen.

- **Test auf Veränderungen in binären Daten** Der McNemar-Test (2 Stichproben) kann auf kategoriale Felder angewendet werden. Mit dieser Option wird ein Test bei verbundenen Stichproben erstellt, der prüft, ob Wertekombinationen zwischen zwei Flag-Feldern (kategoriale Felder mit nur zwei Werten) gleich wahrscheinlich sind. Der Test wird nicht durchgeführt, wenn auf der Registerkarte "Felder" mehr als zwei Felder angegeben wurden. Unter McNemar-Test: Erfolg definieren finden Sie weitere Informationen über die Testeinstellungen. Cochrans Q-Test (k-Stichproben) kann auf kategoriale Felder angewendet werden. Mit dieser Option wird ein Test bei verbundenen Stichproben erstellt, der prüft, ob Wertekombinationen zwischen *k* Flag-Feldern (kategoriale Felder mit nur zwei Werten) gleich wahrscheinlich sind. Sie können optional Mehrfachvergleiche der *k*-Stichproben anfordern, entweder alle paarweisenMehrfachvergleiche oder schrittweise Step-Down-Vergleiche. Unter Cochrans Q-Test: Erfolg definieren finden Sie weitere Informationen über die Testeinstellungen.
- **Test auf Veränderungen in multinomialen Daten** Der Rand-Homogenitätstest (2 Stichproben) erstellt einen Test bei verbundenen Stichproben, der prüft, ob Wertekombinationen zwischen zwei gepaarten ordinalen Feldern gleich wahrscheinlich sind. Der Rand-Homogenitätstest wird üblicherweise bei Messwiederholungen verwendet. Dieser Test ist eine Erweiterung des McNemar-Tests von binären Variablen auf multinomiale Variablen. Der Test wird nicht durchgeführt, wenn auf der Registerkarte "Felder" mehr als zwei Felder angegeben wurden.
- **Median- und hypothetische Differenz vergleichen** Jeder dieser Tests erstellt einen Test bei verbundenen Stichproben, der prüft, ob die Mediandifferenzen zwischen zwei stetigen Feldern von 0 abweichen. Dieser Test wird nicht durchgeführt, wenn in der Registerkarte "Felder" mehr als zwei Felder angegeben wurden.
- **Konfidenzintervall schätzen** Mit dieser Option wird eine Schätzung und ein Konfidenzintervall bei verbundenen Stichproben für die Mediandifferenz zwischen zwei gepaarten stetigen Feldern erstellt. Der Test wird nicht durchgeführt, wenn auf der Registerkarte "Felder" mehr als zwei Felder angegeben wurden.
- **Zusammenhänge quantifizieren** Der Konkordanz-Koeffizient nach Kendall (k-Stichproben) erstellt ein Maß für die Übereinstimmung der Sachverständigen oder Prüfer, in dem jeder Datensatz der Bewertung eines Sachverständigen von mehreren Elementen (Feldern) entspricht. Sie können optional Mehrfachvergleiche der *k*-Stichproben anfordern, entweder alle paarweisen Mehrfachvergleiche oder schrittweise Step-Down-Vergleiche.
- **Verteilungen vergleichen** Friedmans zweifaktorielle ANOVA nach Rang (k-Stichproben) erstellt einen Test bei verbundenen Stichproben, der prüft, ob *k* verbundene Stichproben aus der gleichen Grundgesamtheit gezogen wurden. Sie können optional Mehrfachvergleiche der *k*-Stichproben anfordern, entweder alle paarweisen Mehrfachvergleiche oder schrittweise Step-Down-Vergleiche.

McNemar-Test: Erfolg definieren

Abbildung 27-18
Tests bei verbundenen Stichproben – McNemar-Test: Einstellungen "Erfolg definieren"



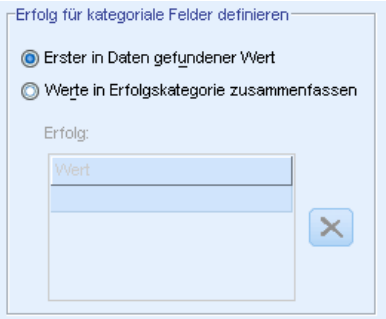
Der McNemar-Test ist für Flag-Felder gedacht (kategoriale Felder mit nur zwei Kategorien), wird aber auf alle kategorialen Felder angewendet, indem Regeln zur Definition von "Erfolg" aufgestellt werden.

Erfolg für kategoriale Felder definieren Gibt an, wie "Erfolg" für kategoriale Felder definiert wird.

- Erste in Daten gefundene Kategorie verwenden führt den Test mithilfe des ersten in der Stichprobe gefundenen Werts durch, um "Erfolg" zu definieren. Diese Option ist nur für nominale oder ordinale Felder mit nur zwei Werten verfügbar; alle anderen in der Registerkarte "Felder" angegebenen kategorialen Felder, in denen diese Option verwendet wird, werden nicht getestet. Dies ist die Standardeinstellung.
- Erfolgswerte festlegen führt den Test mithilfe der angegebenen Werteliste durch, um "Erfolg" zu definieren. Geben Sie eine Liste von String- oder numerischen Werten an. Die Werte in der Liste müssen nicht in der Stichprobe vorhanden sein.

Cochrans Q-Test: Erfolg definieren

Abbildung 27-19
Tests bei verbundenen Stichproben – Cochrans Q-Test: Erfolg definieren



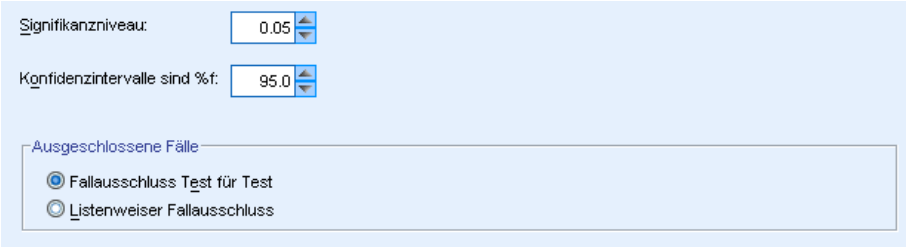
Cochrans Q-Test ist für Flag-Felder gedacht (kategoriale Felder mit nur zwei Kategorien), wird aber auf alle kategorialen Felder angewendet, indem Regeln zur Definition von "Erfolg" aufgestellt werden.

Erfolg für kategoriale Felder definieren Gibt an, wie "Erfolg" für kategoriale Felder definiert wird.

- Erste in Daten gefundene Kategorie verwenden führt den Test mithilfe des ersten in der Stichprobe gefundenen Werts durch, um "Erfolg" zu definieren. Diese Option ist nur für nominale oder ordinale Felder mit nur zwei Werten verfügbar; alle anderen in der Registerkarte "Felder" angegebenen kategorialen Felder, in denen diese Option verwendet wird, werden nicht getestet. Dies ist die Standardeinstellung.
- Erfolgswerte festlegen führt den Test mithilfe der angegebenen Werteliste durch, um "Erfolg" zu definieren. Geben Sie eine Liste von String- oder numerischen Werten an. Die Werte in der Liste müssen nicht in der Stichprobe vorhanden sein.

Testoptionen

Abbildung 27-20
Einstellungen "Testoptionen" in nichtparametrischen Tests bei verbundenen Stichproben



Signifikanzniveau. Gibt das Signifikanzniveau (Alpha) für alle Tests an. Geben Sie einen numerischen Wert zwischen 0 und 1 an. 0,05 ist die Standardeinstellung.

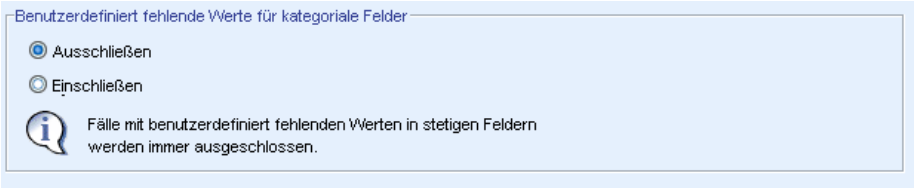
Konfidenzintervall (%). Gibt das Konfidenzniveau für alle erstellten Konfidenzintervalle an. Geben Sie einen numerischen Wert zwischen 0 und 100 an. 95 ist die Standardeinstellung.

Ausgeschlossene Fälle. Gibt an, wie die Fallbasis für Tests bestimmt wird.

- Listenweiser Fallausschluss bedeutet, dass Datensätze mit fehlenden Werten für ein beliebiges Feld, das in einem beliebigen Unterbefehl genannt wurde, aus allen Analysen ausgeschlossen werden.
- Fallausschluss Test für Test bedeutet, dass Datensätze mit fehlenden Werten für ein Feld, das für einen bestimmten Test verwendet wird, aus diesem Test ausgeschlossen werden. Wenn in der Analyse mehrere Tests angegeben wurden, wird jeder Test getrennt ausgewertet.

Benutzerdefiniert fehlende Werte

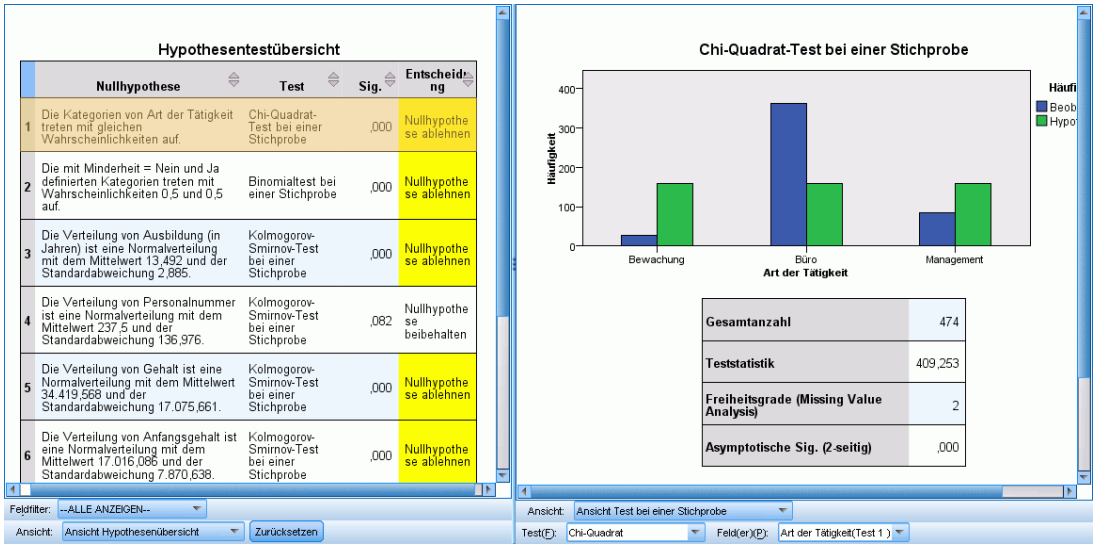
Abbildung 27-21
Einstellungen "Benutzerdefiniert fehlende Werte" in nichtparametrischen Tests bei verbundenen Stichproben



Benutzerdefiniert fehlende Werte für kategoriale Felder Kategoriale Felder müssen gültige Werte für einen Datensatz aufweisen, um in die Analyse aufgenommen zu werden. Mit diesen Steuerungen legen Sie fest, ob benutzerdefiniert fehlende Werte bei den kategorialen Feldern als gültige Werte behandelt werden sollen. Systemdefinierte fehlende Werte und fehlende Werte für stetige Felder werden immer als ungültige Werte behandelt.

Modellanzeige

Abbildung 27-22
Modellanzeige nichtparametrische Tests



Die Prozedur erstellt ein Modellansichts-Objekt im Viewer. Wenn Sie dieses Objekt durch einen Doppelklick aktivieren, erhalten Sie eine interaktive Ansicht des Modells. Das Fenster der Modellansicht setzt sich aus zwei Bereichen zusammen, der Hauptansicht im linken Bereich und der verknüpften oder Hilfsansicht im rechten Bereich.

Es gibt zwei Hauptansichten:

- Hypothesenübersicht Dies ist die Standardansicht.Für weitere Informationen siehe Thema Hypothesenübersicht auf S. 228.
- Konfidenzintervallübersicht Für weitere Informationen siehe Thema Konfidenzintervallübersicht auf S. 230.

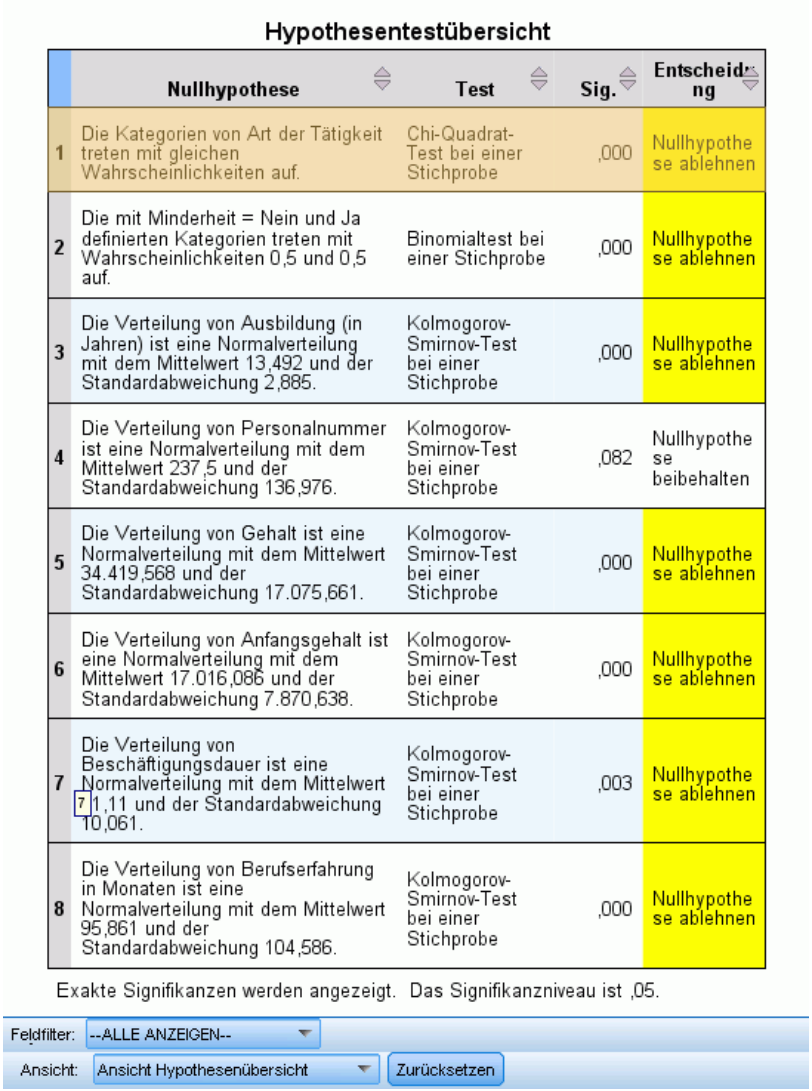
Es gibt sieben verknüpfte/Hilfsansichten:

- Ansicht Test bei einer Stichprobe Dies ist die Standardansicht, falls Tests bei einer Stichprobe angefordert wurden. Für weitere Informationen siehe Thema Test bei einer Stichprobe auf S. 230.
- Ansicht Test bei verbundenen Stichproben Dies ist die Standardansicht, falls keine Tests bei einer Stichprobe, sondern Tests bei mehreren verbundenen Stichproben angefordert wurden. Für weitere Informationen siehe Thema Test bei verbundenen Stichproben auf S. 235.

- Ansicht Test bei unabhängigen Stichproben Dies ist die Standardansicht, falls keine Tests bei mehreren verbundenen Stichproben oder Tests bei einer Stichprobe angefordert wurden. Für weitere Informationen siehe Thema Test bei unabhängigen Stichproben auf S. 242.
- Informationen über kategoriales Feld Für weitere Informationen siehe Thema Informationen über kategoriales Feld, auf S. 250.
- Informationen über stetiges Feld Für weitere Informationen siehe Thema Informationen über stetiges Feld, auf S. 251.
- Paarweise Vergleiche Für weitere Informationen siehe Thema Paarweise Vergleiche auf S. 252.
- Homogene Untergruppen Für weitere Informationen siehe Thema Homogene Untergruppen auf S. 253.

Hypothesenübersicht

Abbildung 27-23
Hypothesenübersicht



Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Kategorien von Art der Tätigkeit treten mit gleichen Wahrscheinlichkeiten auf.	Chi-Quadrat-Test bei einer Stichprobe	,000	Nullhypothese ablehnen
2	Die mit Minderheit = Nein und Ja definierten Kategorien treten mit Wahrscheinlichkeiten 0,5 und 0,5 auf.	Binomialtest bei einer Stichprobe	,000	Nullhypothese ablehnen
3	Die Verteilung von Ausbildung (in Jahren) ist eine Normalverteilung mit dem Mittelwert 13,492 und der Standardabweichung 2,885.	Kolmogorov-Smirnov-Test bei einer Stichprobe	,000	Nullhypothese ablehnen
4	Die Verteilung von Personalnummer ist eine Normalverteilung mit dem Mittelwert 237,5 und der Standardabweichung 136,976.	Kolmogorov-Smirnov-Test bei einer Stichprobe	,082	Nullhypothese beibehalten
5	Die Verteilung von Gehalt ist eine Normalverteilung mit dem Mittelwert 34.419,568 und der Standardabweichung 17.075,661.	Kolmogorov-Smirnov-Test bei einer Stichprobe	,000	Nullhypothese ablehnen
6	Die Verteilung von Anfangsgehalt ist eine Normalverteilung mit dem Mittelwert 17.016,086 und der Standardabweichung 7.870,638.	Kolmogorov-Smirnov-Test bei einer Stichprobe	,000	Nullhypothese ablehnen
7	Die Verteilung von Beschäftigungsdauer ist eine Normalverteilung mit dem Mittelwert 81,11 und der Standardabweichung 10,061.	Kolmogorov-Smirnov-Test bei einer Stichprobe	,003	Nullhypothese ablehnen
8	Die Verteilung von Berufserfahrung in Monaten ist eine Normalverteilung mit dem Mittelwert 95,861 und der Standardabweichung 104,586.	Kolmogorov-Smirnov-Test bei einer Stichprobe	,000	Nullhypothese ablehnen

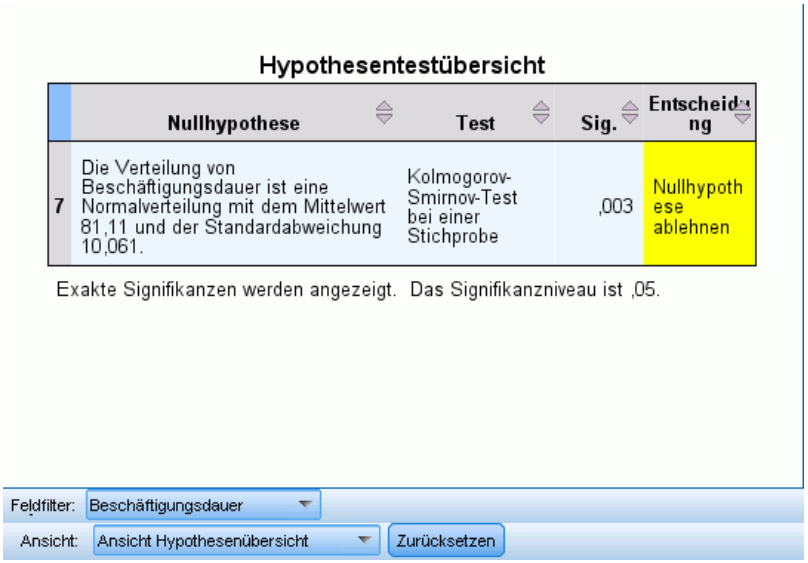
Exakte Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

Mit der Ansicht “Modellzusammenfassung” erhalten Sie eine momentane, übersichtliche Zusammenfassung der nichtparametrischen Tests. Sie hebt Nullhypothesen und Entscheidungen hervor und lenkt so die Aufmerksamkeit auf signifikante *p*-Werte.

- Jede Zeile entspricht einem separaten Test. Durch Klicken auf eine Zeile werden in der verknüpften Ansicht zusätzliche Informationen zum Test angezeigt.
- Durch Klicken auf eine Spaltenüberschrift werden die Zeilen nach den Werten in dieser Spalte sortiert.

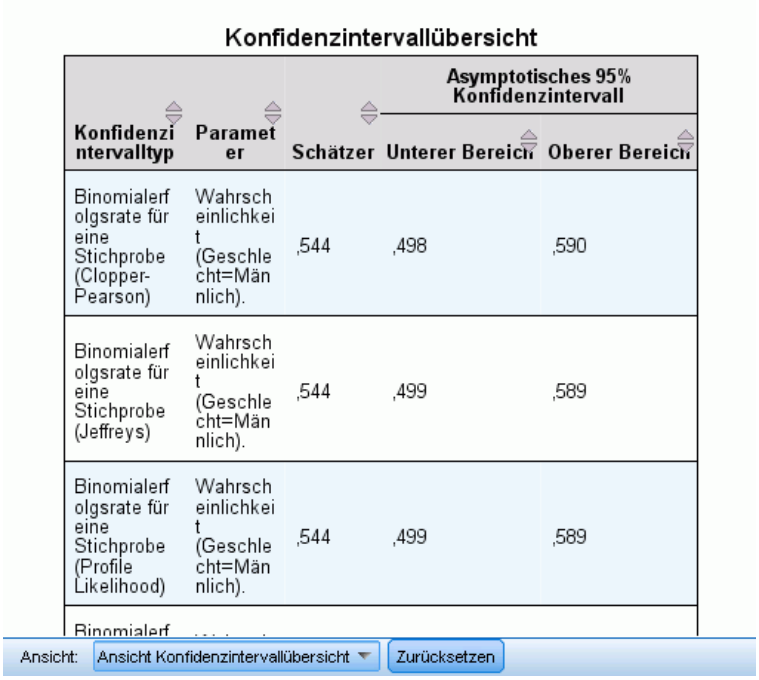
- Sie können die Modellanzeige über die Schaltfläche Zurücksetzen wieder in ihren Originalzustand versetzen.
- Die Dropdown-Liste Feldfilter ermöglicht es, nur diejenigen Tests anzuzeigen, die das ausgewählte Feld betreffen. Wenn in der Dropdown-Liste Feldfilter also beispielsweise *Anfangsgehalt* ausgewählt wurde, werden in der Hypothesenübersicht nur zwei Tests angezeigt.

Abbildung 27-24
Hypothesenübersicht, gefiltert nach Anfangsgehalt



Konfidenzintervallübersicht

Abbildung 27-25
Konfidenzintervallübersicht



Konfidenzintervallübersicht

Konfidenzintervalltyp	Parameter	Schätzer	Asymptotisches 95% Konfidenzintervall	
			Unterer Bereich	Oberer Bereich
Binomialerfolgsrate für eine Stichprobe (Clopper-Pearson)	Wahrscheinlichkeit (Geschlecht=Männlich).	,544	,498	,590
Binomialerfolgsrate für eine Stichprobe (Jeffreys)	Wahrscheinlichkeit (Geschlecht=Männlich).	,544	,499	,589
Binomialerfolgsrate für eine Stichprobe (Profile Likelihood)	Wahrscheinlichkeit (Geschlecht=Männlich).	,544	,499	,589
Binomialerf				

Ansicht: Ansicht Konfidenzintervallübersicht | Zurücksetzen

Die Konfidenzintervallübersicht zeigt alle Konfidenzintervalle an, die von den nichtparametrischen Tests erzeugt werden.

- Jede Zeile entspricht einem separaten Konfidenzintervall.
- Durch Klicken auf eine Spaltenüberschrift werden die Zeilen nach den Werten in dieser Spalte sortiert.

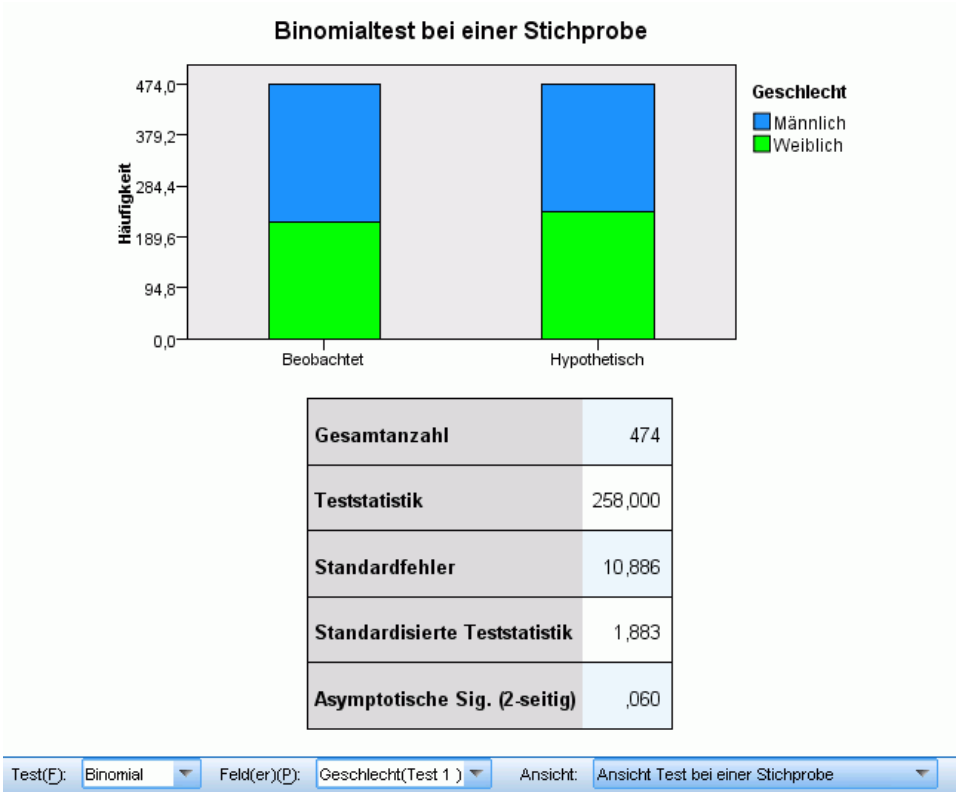
Test bei einer Stichprobe

Die Ansicht Test bei einer Stichprobe zeigt Details zu allen angeforderten nichtparametrischen Tests bei einer Stichprobe an. Die angezeigten Informationen hängen vom ausgewählten Test ab.

- Die Dropdown-Liste Test ermöglicht Ihnen die Auswahl eines bestimmten Tests bei einer Stichprobe.
- Die Dropdown-Liste Feld(er) ermöglicht Ihnen die Auswahl eines Felds, das mit dem in der Dropdown-Liste Test ausgewählten Test getestet wurde.

Test auf Binomialverteilung

Abbildung 27-26
Ansicht Test bei einer Stichprobe, Test auf Binomialverteilung

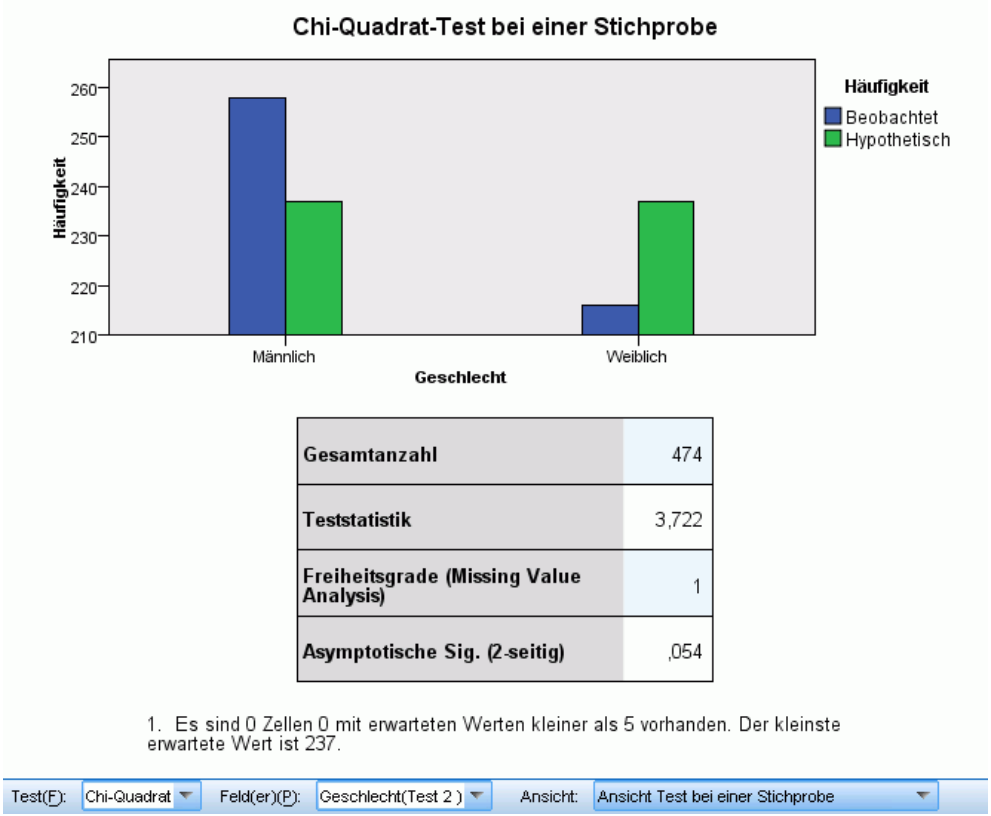


Der Test auf Binomialverteilung zeigt ein gestapeltes Balkendiagramm und eine Testtabelle an.

- Das gestapelte Balkendiagramm zeigt die beobachteten und hypothetischen Häufigkeiten der Kategorien “Erfolg” und “Fehlschlag” des Testfelds an, wobei “Fehlschläge” auf “Erfolge” gestapelt werden. Wenn Sie die Maus über einen Balken bewegen, werden in eine QuickInfo die Prozentwerte der Kategorien angezeigt. Sichtbare Unterschiede zwischen den Balken deuten darauf hin, dass das Testfeld unter Umständen nicht die hypothetische Binomialverteilung aufweist.
- Die Tabelle zeigt Details zum Test an.

Chi-Quadrat-Test

Abbildung 27-27
Ansicht Test bei einer Stichprobe, Chi-Quadrat-Test



Gesamtanzahl	474
Teststatistik	3,722
Freiheitsgrade (Missing Value Analysis)	1
Asymptotische Sig. (2-seitig)	,054

1. Es sind 0 Zellen 0 mit erwarteten Werten kleiner als 5 vorhanden. Der kleinste erwartete Wert ist 237.

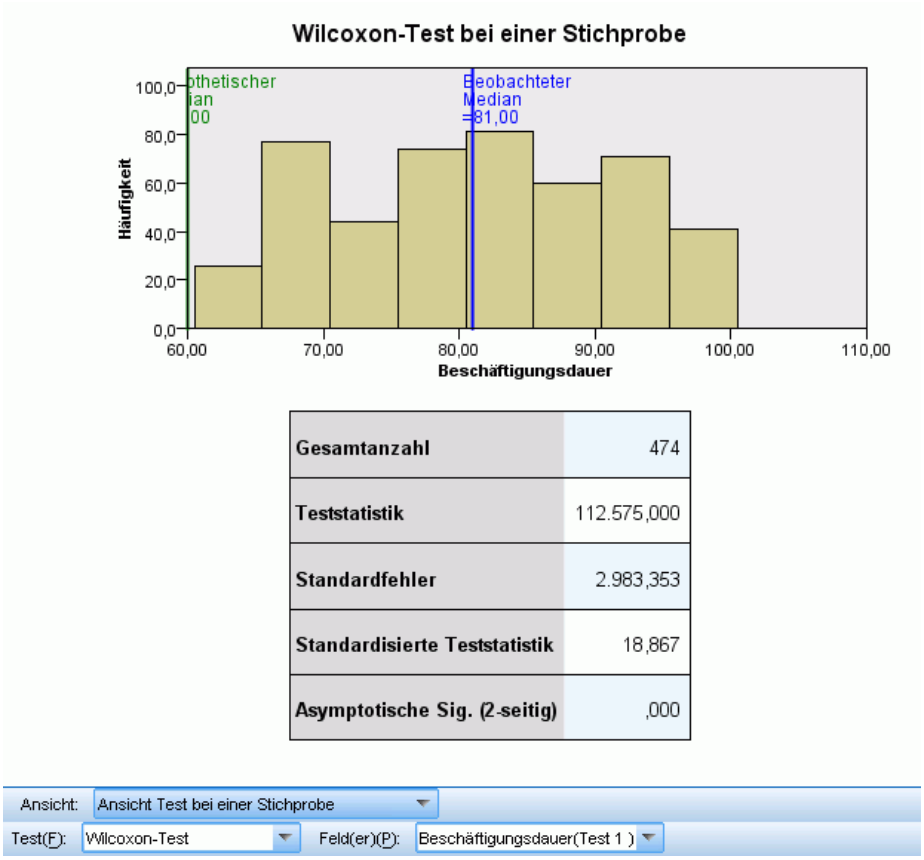
Test(F): Chi-Quadrat | Feld(er)(P): Geschlecht(Test 2) | Ansicht: Ansicht Test bei einer Stichprobe

Der Chi-Quadrat-Test zeigt ein gruppiertes Balkendiagramm und eine Testtabelle an.

- Das gruppierte Balkendiagramm zeigt die beobachteten und hypothetischen Häufigkeiten für jede Kategorie des Testfelds an. Wenn Sie die Maus über einen Balken bewegen, werden in einer QuickInfo die beobachteten und hypothetischen Häufigkeiten sowie ihre Abweichungen (Residuen) angezeigt. Sichtbare Unterschiede zwischen den Balken der beobachteten und der hypothetischen Häufigkeiten deuten darauf hin, dass das Testfeld unter Umständen nicht die hypothetische Verteilung aufweist.
- Die Tabelle zeigt Details zum Test an.

Wilcoxon-Test

Abbildung 27-28
Ansicht Test bei einer Stichprobe, Wilcoxon-Test

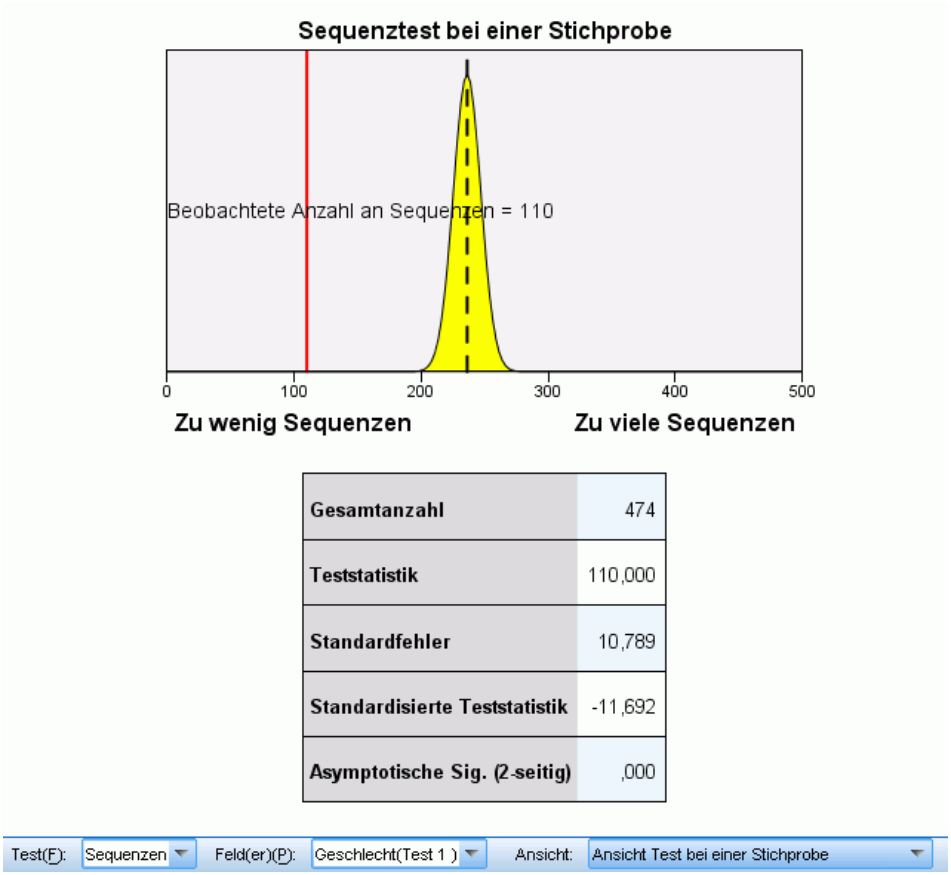


Der Wilcoxon-Test zeigt ein Histogramm und eine Testtabelle an.

- Das Histogramm enthält vertikale Linien, die die beobachteten und hypothetischen Mediane anzeigen.
- Die Tabelle zeigt Details zum Test an.

Sequenzentest

Abbildung 27-29
Ansicht Test bei einer Stichprobe, Sequenztest



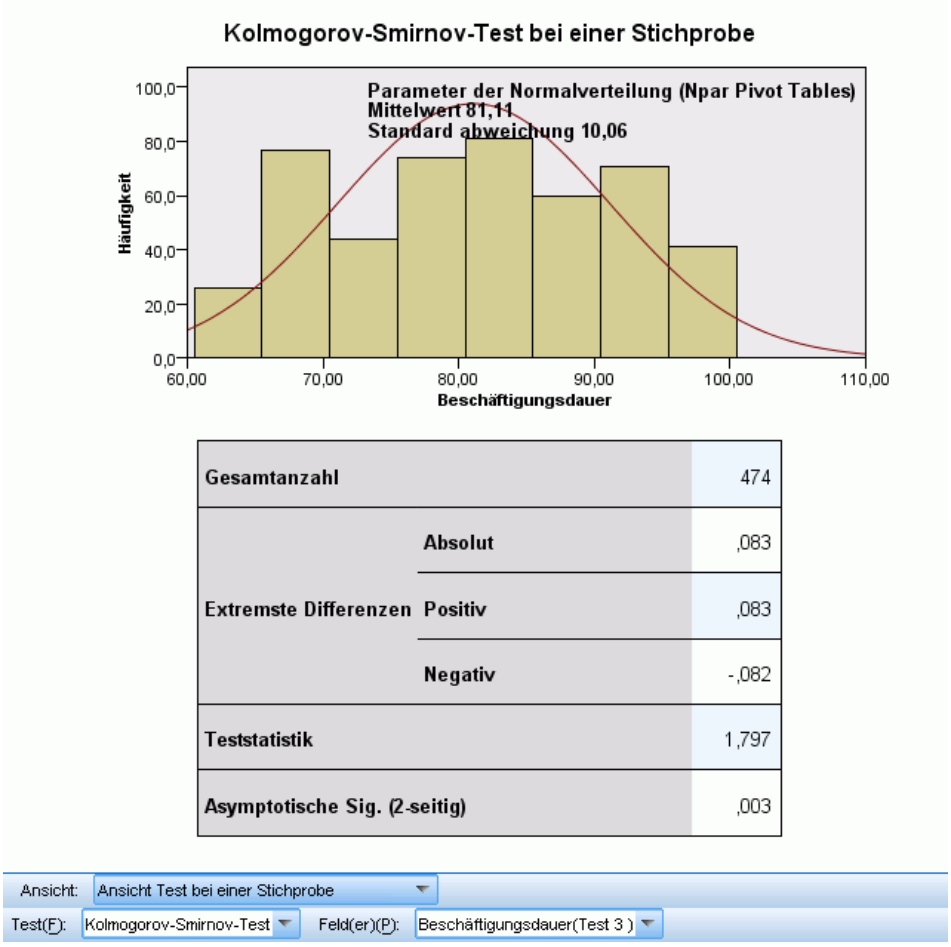
Gesamtanzahl	474
Teststatistik	110,000
Standardfehler	10,789
Standardisierte Teststatistik	-11,692
Asymptotische Sig. (2-seitig)	,000

Der Sequenztest zeigt ein Diagramm und eine Testtabelle an.

- Das Diagramm zeigt eine Normalverteilung an, in der die beobachtete Anzahl an Sequenzen durch eine vertikale Linie gekennzeichnet ist. Beachten Sie, dass der Test bei der exakten Durchführung nicht auf der Normalverteilung basiert.
- Die Tabelle zeigt Details zum Test an.

Kolmogorov-Smirnov-Test

Abbildung 27-30
Ansicht Test bei einer Stichprobe, Kolmogorov-Smirnov-Test



Der Kolmogorov-Smirnov-Test zeigt ein Histogramm und eine Testtabelle an.

- Das Histogramm enthält eine Überlagerung der Wahrscheinlichkeitsdichtefunktion für die hypothetische Gleich-, Normal-, Poisson- oder Exponentialverteilung. Beachten Sie, dass der Test auf kumulativen Verteilungen basiert und die in der Tabelle angegebenen extremsten Differenzen in Bezug auf kumulative Verteilungen interpretiert werden sollten.
- Die Tabelle zeigt Details zum Test an.

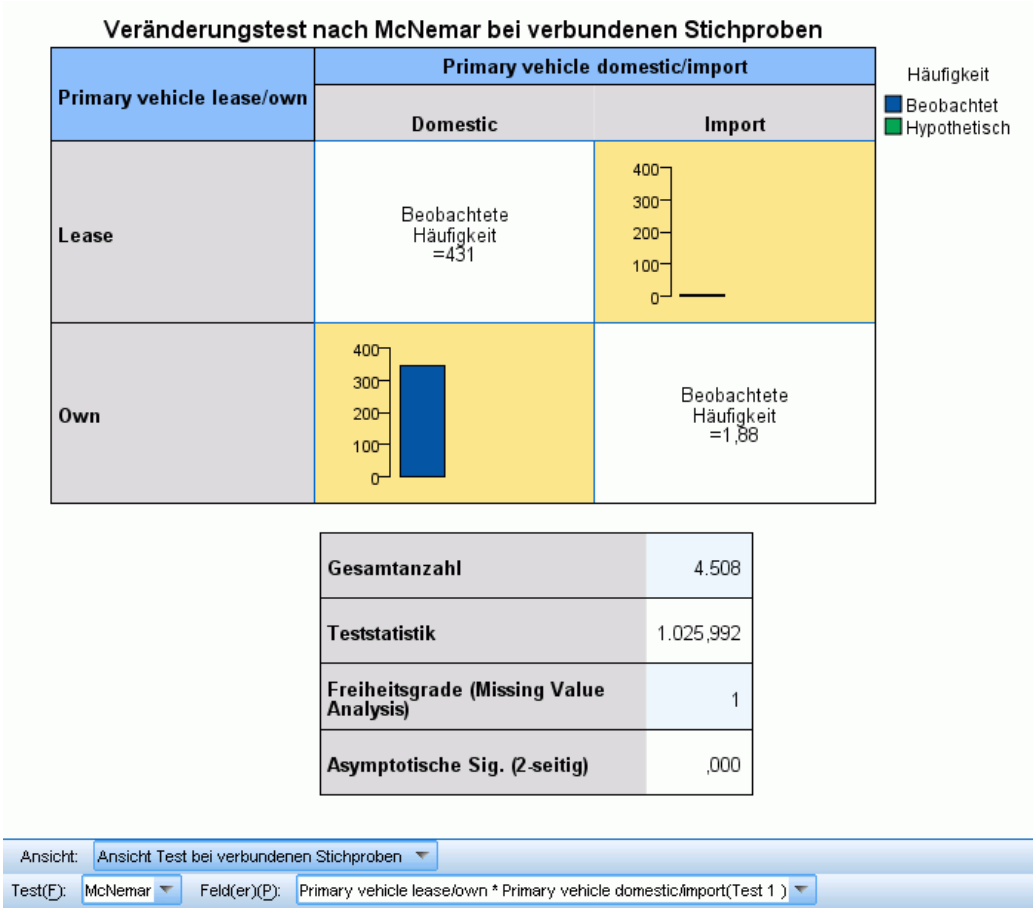
Test bei verbundenen Stichproben

Die Ansicht Test bei einer Stichprobe zeigt Details zu allen angeforderten nichtparametrischen Tests bei einer Stichprobe an. Die angezeigten Informationen hängen vom ausgewählten Test ab.

- Die Dropdown-Liste Test ermöglicht Ihnen die Auswahl eines bestimmten Tests bei einer Stichprobe.
- Die Dropdown-Liste Feld(er) ermöglicht Ihnen die Auswahl eines Felds, das mit dem in der Dropdown-Liste Test ausgewählten Test getestet wurde.

McNemar-Test

Abbildung 27-31
Ansicht Test bei verbundenen Stichproben, McNemar-Test

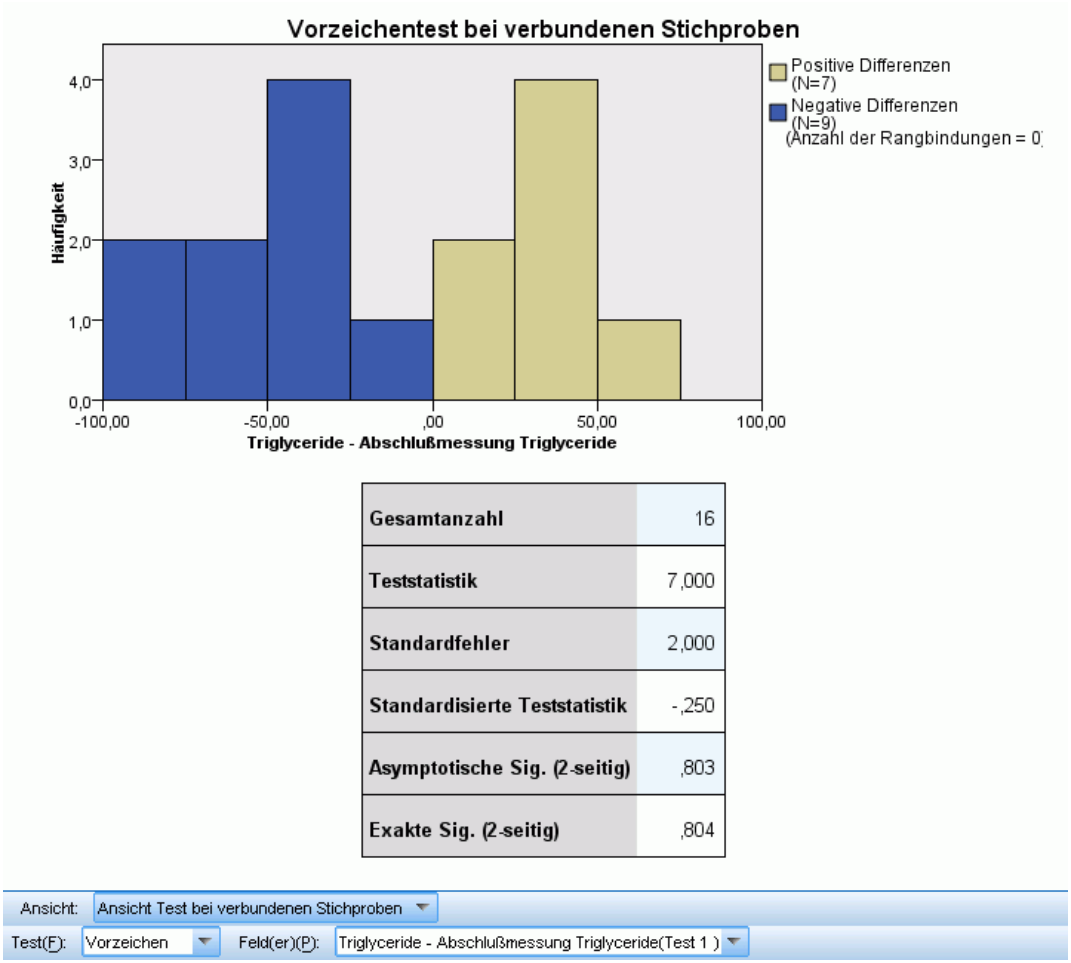


Der McNemar-Test zeigt ein gruppiertes Balkendiagramm und eine Testtabelle an.

- Das gruppierte Balkendiagramm zeigt die beobachteten und hypothetischen Häufigkeiten für die nicht auf der Diagonalen liegenden Zellen der von den Testfeldern definierten 2×2-Tabelle an.
- Die Tabelle zeigt Details zum Test an.

Vorzeichentest

Abbildung 27-32
Ansicht Test bei verbundenen Stichproben, Vorzeichentest



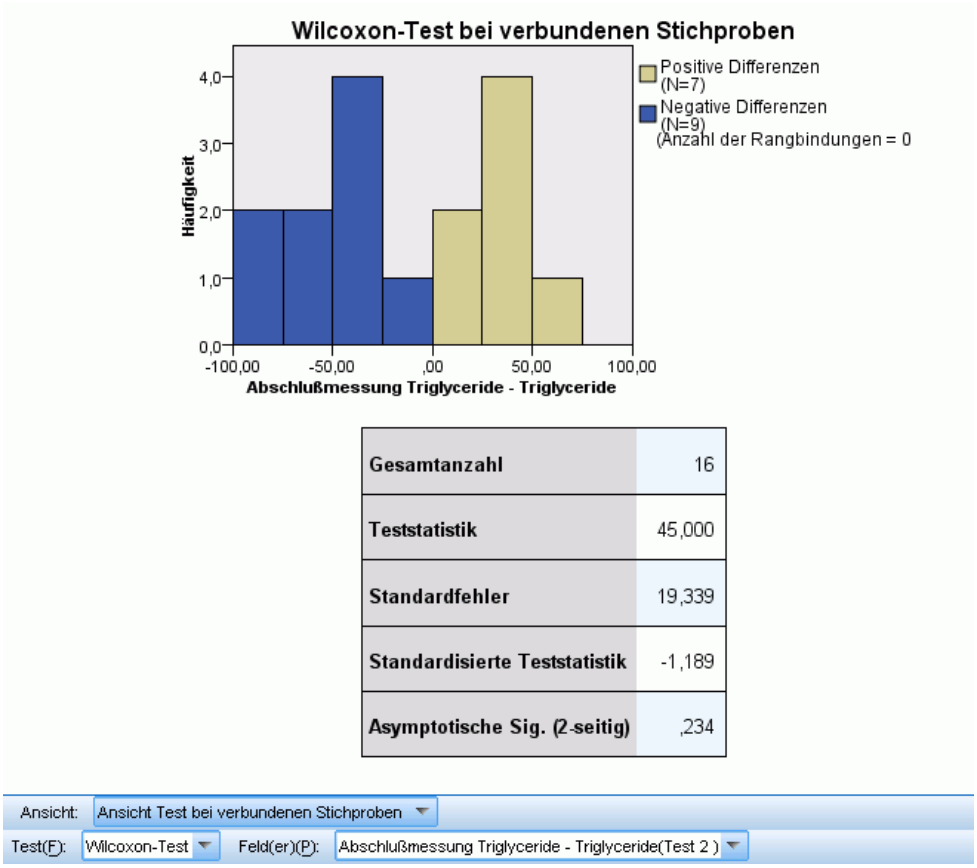
Gesamtanzahl	16
Teststatistik	7,000
Standardfehler	2,000
Standardisierte Teststatistik	-,250
Asymptotische Sig. (2-seitig)	,803
Exakte Sig. (2-seitig)	,804

Der Vorzeichentest zeigt ein gestapeltes Histogramm und eine Testtabelle an.

- Das gestapelte Histogramm zeigt die Differenzen zwischen den Feldern an und verwendet dabei das Vorzeichen der Differenz als stapelndes Feld.
- Die Tabelle zeigt Details zum Test an.

Wilcoxon-Test

Abbildung 27-33
Ansicht Test bei verbundenen Stichproben, Wilcoxon-Test



Gesamtanzahl	16
Teststatistik	45,000
Standardfehler	19,339
Standardisierte Teststatistik	-1,189
Asymptotische Sig. (2-seitig)	,234

Der Wilcoxon-Test zeigt ein gestapeltes Histogramm und eine Testtabelle an.

- Das gestapelte Histogramm zeigt die Differenzen zwischen den Feldern an und verwendet dabei das Vorzeichen der Differenz als stapelndes Feld.
- Die Tabelle zeigt Details zum Test an.

Rand-Homogenitätstest

Abbildung 27-34
Ansicht Test bei verbundenen Stichproben, Rand-Homogenitätstest

Rand-Homogenitätstest bei verbundenen Stichproben

Level of education	Spouse level of education				
	Did not complete high school	High school degree	Some college	College degree	Post-undergraduate degree
Did not complete high school	Beobachtete Häufigkeit =291				
High school degree		Beobachtete Häufigkeit =344			
Some college			Beobachtete Häufigkeit =159		
College degree				Beobachtete Häufigkeit =174	
Post-undergraduate degree					Beobachtete Häufigkeit =43

Gesamtanzahl	2.441
Teststatistik	2.660,000
Standardfehler	25,466
Standardisierte Teststatistik	10,642
Asymptotische Sig. (2-seitig)	,000

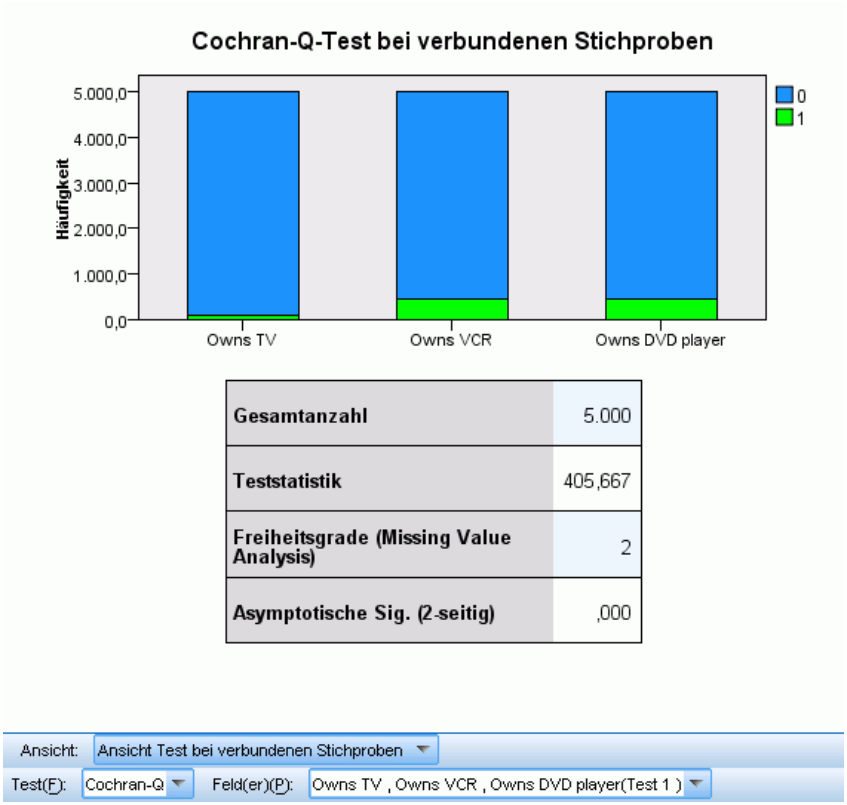
Test(F): Rand-Homogenität Feld(er)(P): Level of education * Spouse level of education(Test 1) Ansicht: Ansicht Test bei verbundenen Stichproben

Der Rand-Homogenitätstest zeigt ein gruppiertes Balkendiagramm und eine Testtabelle an.

- Das gruppierte Balkendiagramm zeigt die beobachteten Häufigkeiten für die nicht auf der Diagonalen liegenden Zellen der von den Testfeldern definierten Tabelle an.
- Die Tabelle zeigt Details zum Test an.

Cochrans Q-Test

Abbildung 27-35
Ansicht Test bei verbundenen Stichproben, Cochrans Q-Test

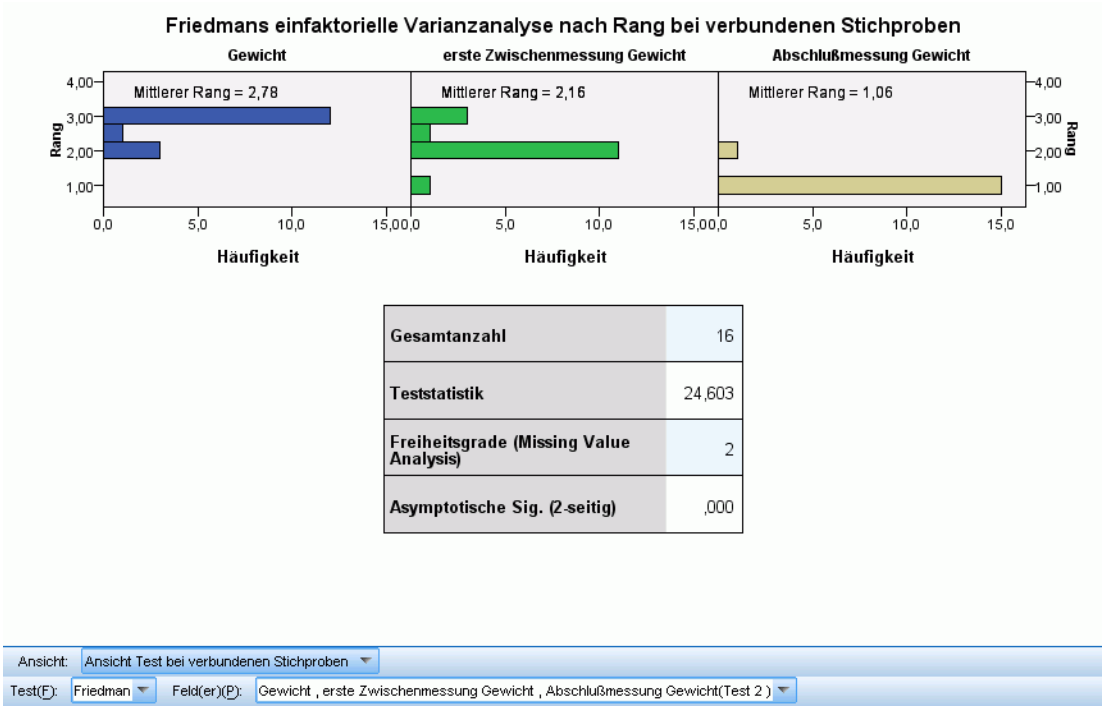


Cochrans Q-Test zeigt ein gestapeltes Balkendiagramm und eine Testtabelle an.

- Das gestapelte Balkendiagramm zeigt die beobachteten Häufigkeiten der Kategorien “Erfolg” und “Fehlschlag” der Testfelder an, wobei “Fehlschläge” auf “Erfolge” gestapelt werden. Wenn Sie die Maus über einen Balken bewegen, werden in eine QuickInfo die Prozentwerte der Kategorien angezeigt.
- Die Tabelle zeigt Details zum Test an.

Friedmans zweifaktorielle Varianzanalyse nach Rang

Abbildung 27-36
Ansicht Test bei verbundenen Stichproben, Friedmans zweifaktorielle Varianzanalyse nach Rang

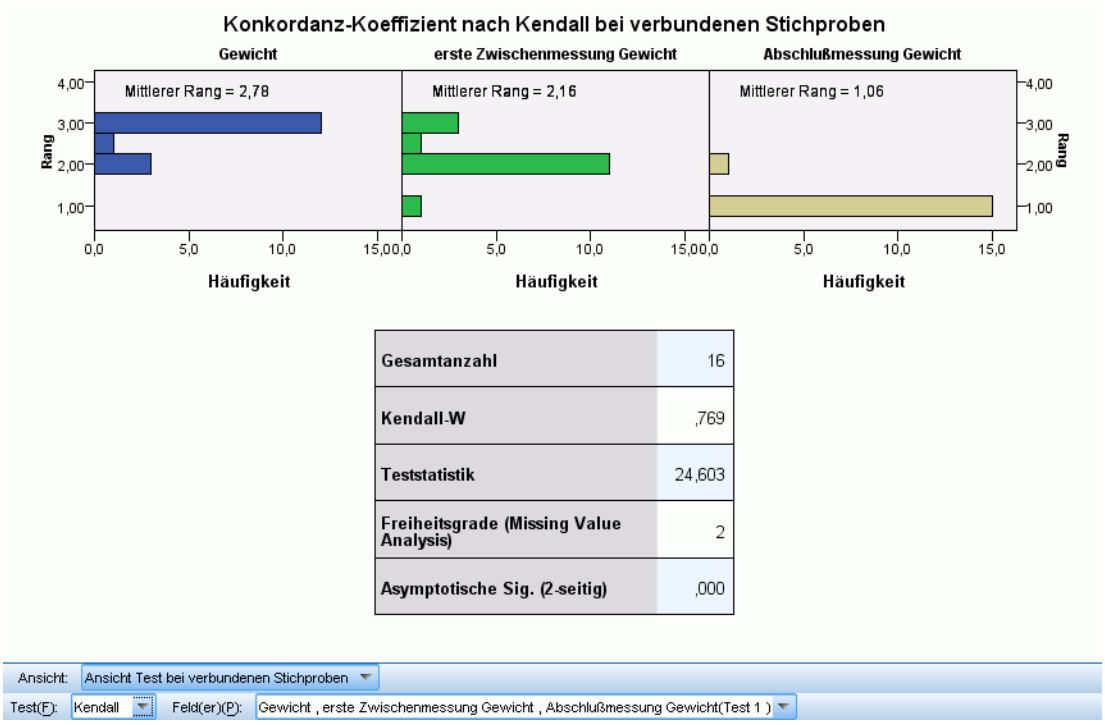


Friedmans zweifaktorielle Varianzanalyse nach Rang zeigt unterteilte Histogramme und eine Testtabelle an.

- Die Histogramme zeigen die beobachtete Verteilung von Rängen unterteilt nach den Testfeldern an.
- Die Tabelle zeigt Details zum Test an.

Konkordanz-Koeffizient nach Kendall

Abbildung 27-37
Ansicht Test bei verbundenen Stichproben, Konkordanz-Koeffizient nach Kendall



Die Ansicht Konkordanz-Koeffizient nach Kendall zeigt unterteilte Histogramme und eine Testtabelle an.

- Die Histogramme zeigen die beobachtete Verteilung von Rängen unterteilt nach den Testfeldern an.
- Die Tabelle zeigt Details zum Test an.

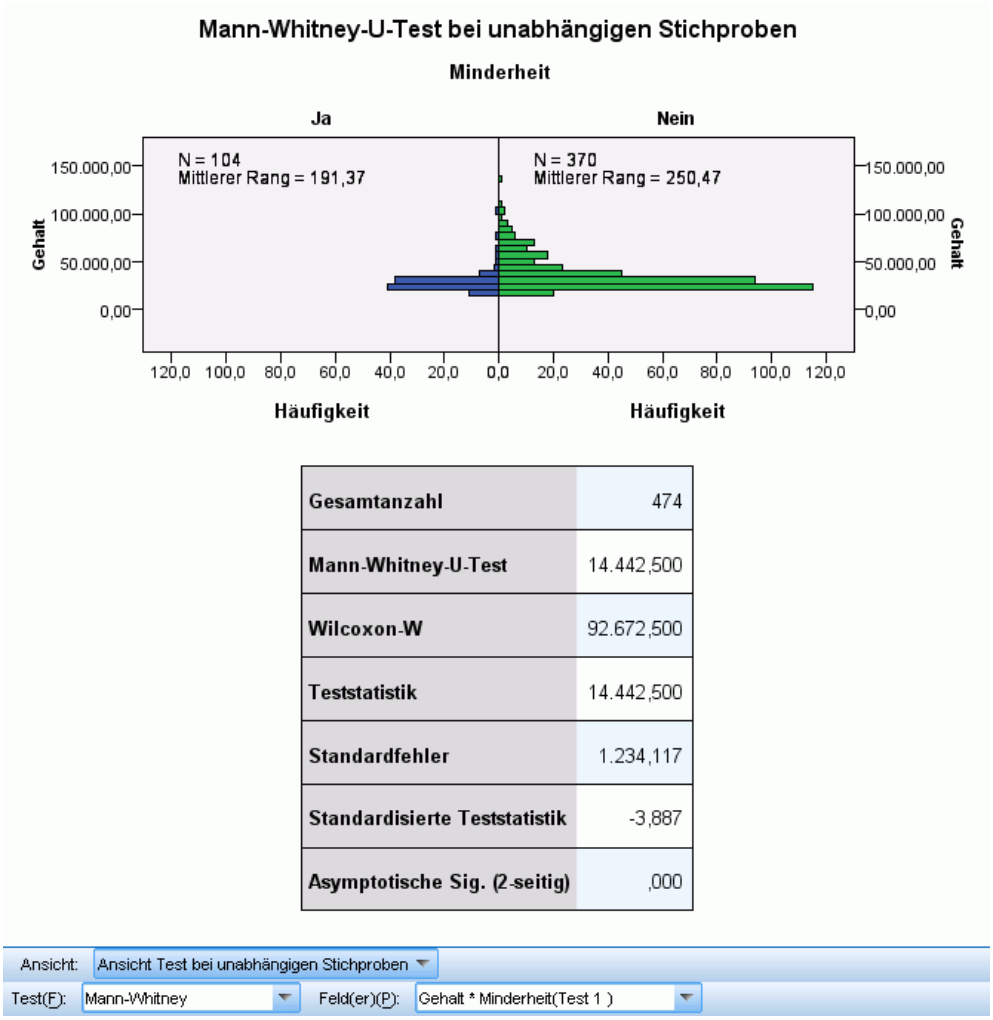
Test bei unabhängigen Stichproben

Die Ansicht Test bei unabhängigen Stichproben zeigt Details zu allen angeforderten nichtparametrischen Tests bei unabhängigen Stichproben an. Die angezeigten Informationen hängen vom ausgewählten Test ab.

- Die Dropdown-Liste Test ermöglicht Ihnen die Auswahl eines bestimmten Tests bei unabhängigen Stichproben.
- Die Dropdown-Liste Feld(er) ermöglicht Ihnen die Auswahl einer Kombination aus Test- und Gruppierungsfeld, die mit dem in der Dropdown-Liste Test ausgewählten Test getestet wurde.

Mann-Whitney-Test

Abbildung 27-38
Ansicht Test bei unabhängigen Stichproben, Mann-Whitney-Test



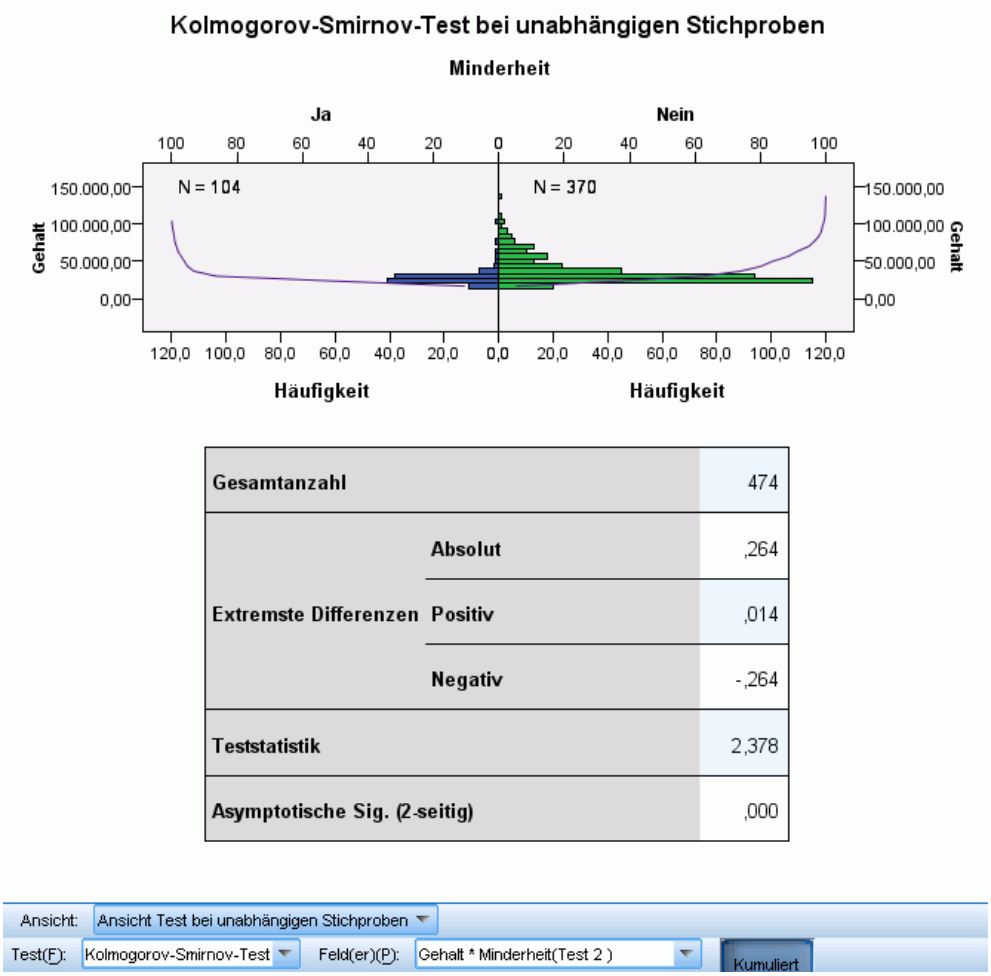
Gesamtanzahl	474
Mann-Whitney-U-Test	14.442,500
Wilcoxon-W	92.672,500
Teststatistik	14.442,500
Standardfehler	1.234,117
Standardisierte Teststatistik	-3,887
Asymptotische Sig. (2-seitig)	,000

Der Mann-Whitney Test zeigt eine Populationspyramide und eine Testtabelle an.

- Die Populationspyramide zeigt Back-to-Back-Histogramme nach den Kategorien der Gruppierungsfelder an, wobei die Anzahl der Datensätze in jeder Gruppe und der mittlere Rank der Gruppe angegeben werden.
- Die Tabelle zeigt Details zum Test an.

Kolmogorov-Smirnov-Test

Abbildung 27-39
Ansicht Test bei unabhängigen Stichproben, Kolmogorov-Smirnov-Test



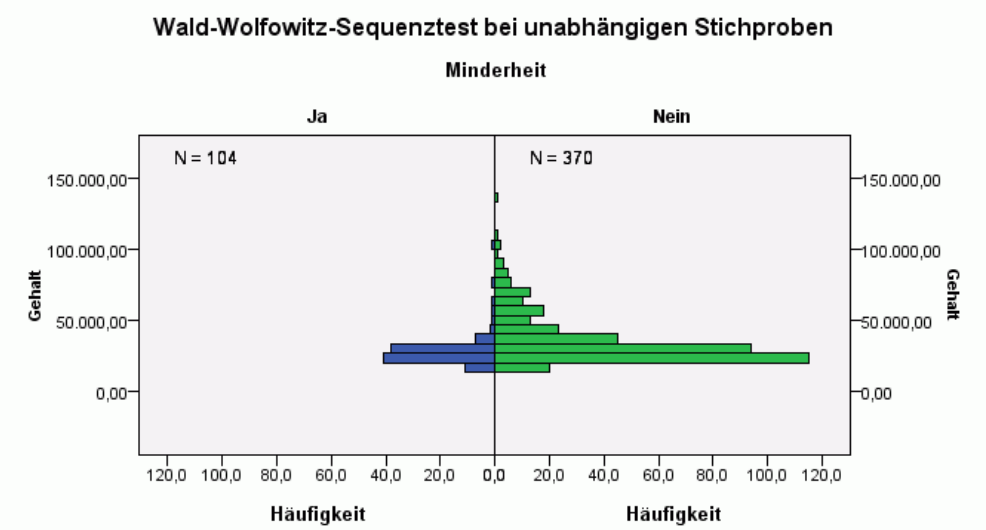
Gesamtanzahl		474
Extremste Differenzen	**Absolut**	,264
	Positiv	,014
	Negativ	-,264
Teststatistik		2,378
Asymptotische Sig. (2-seitig)		,000

Der Kolmogorov-Smirnov-Test zeigt eine Populationspyramide und eine Testtabelle an.

- Die Populationspyramide zeigt Back-to-Back-Histogramme nach den Kategorien der Gruppierungsfelder an, wobei die Anzahl der Datensätze in jeder Gruppe angegeben werden. Die beobachteten kumulativen Verteilungslinien können angezeigt oder ausgeblendet werden, indem Sie auf die Schaltfläche Kumulativ klicken.
- Die Tabelle zeigt Details zum Test an.

Sequenztest nach Wald-Wolfowitz

Abbildung 27-40
Ansicht Test bei unabhängigen Stichproben, Wald-Wolfowitz-Sequenztest



Gesamtanzahl		474
Minimal möglich (Pivot Table NPAR)	**Teststatistik**[1]	97,000
	Standardfehler	7,442
	Standardisierte Teststatistik	-8,917
	Asymptotische Sig. (2-seitig)	,000
Maximal mögliche (Pivot Table NPAR)	**Teststatistik**[1]	199,000
	Standardfehler	7,442
	Standardisierte Teststatistik	4,788
	Asymptotische Sig. (2-seitig)	1,000

[1]The test statistic is the number of runs.

1. There are 55 inter-group ties involving 228 records.

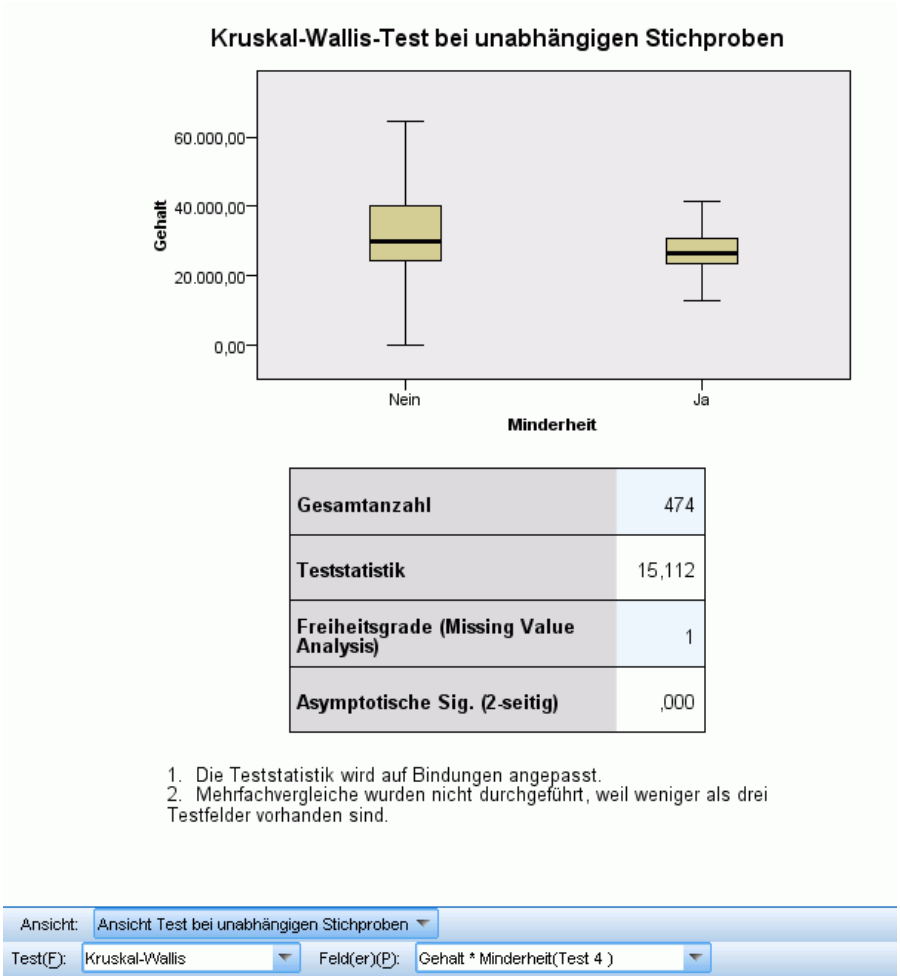
Ansicht: Ansicht Test bei unabhängigen Stichproben
Test(F): Wald-Wolfowitz Feld(er)(P): Gehalt * Minderheit(Test 3)

Der Wald-Wolfowitz-Sequenztest zeigt ein gestapeltes Balkendiagramm und eine Testtabelle an.

- Die Populationspyramide zeigt Back-to-Back-Histogramme nach den Kategorien der Gruppierungsfelder an, wobei die Anzahl der Datensätze in jeder Gruppe angegeben werden.
- Die Tabelle zeigt Details zum Test an.

Kruskal-Wallis-Test

Abbildung 27-41
Ansicht Test bei unabhängigen Stichproben, Kruskal-Wallis-Test

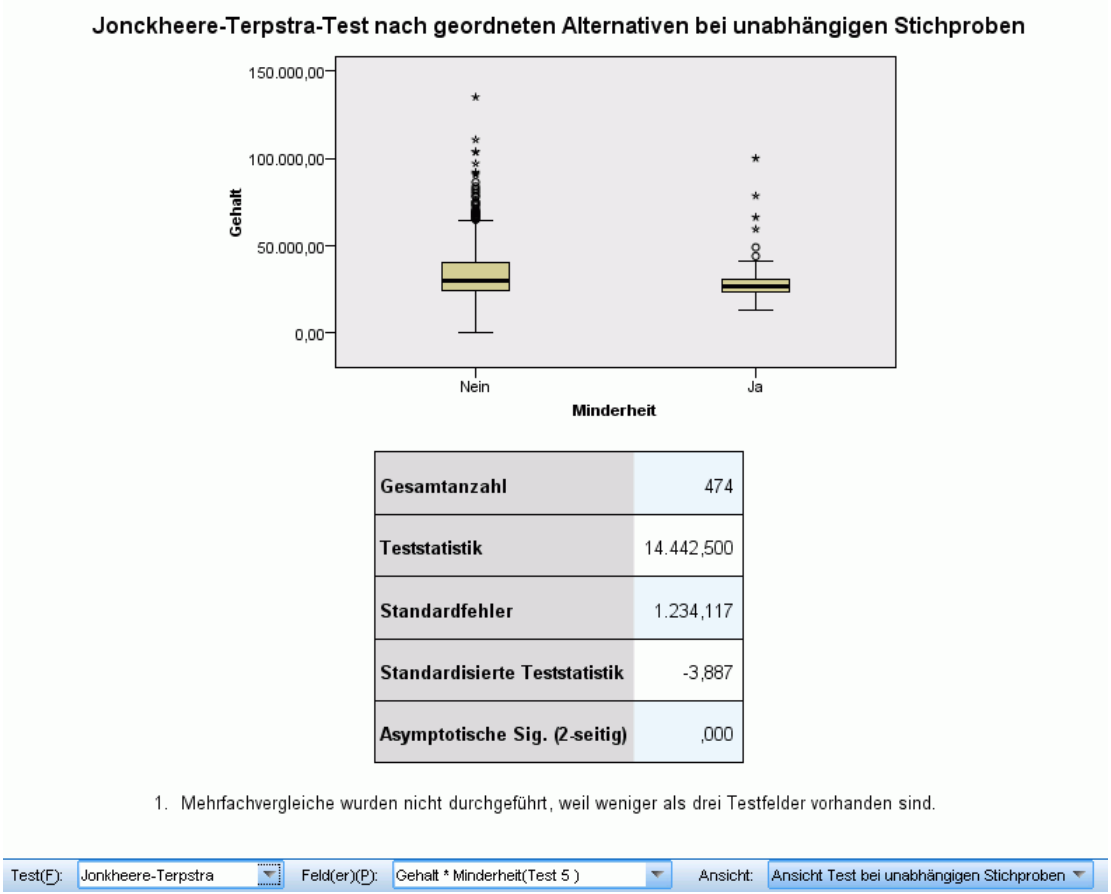


Der Kruskal-Wallis-Test zeigt Boxplots und eine Testtabelle an.

- Für jede Kategorie des Gruppierungsfelds werden separate Boxplots angezeigt. Wenn Sie die Maus über eine Box bewegen, wird in einer QuickInfo der mittlere Rang angezeigt.
- Die Tabelle zeigt Details zum Test an.

Jonckheere-Terpstra-Test

Abbildung 27-42
Ansicht Test bei unabhängigen Stichproben, Jonckheere-Terpstra-Test

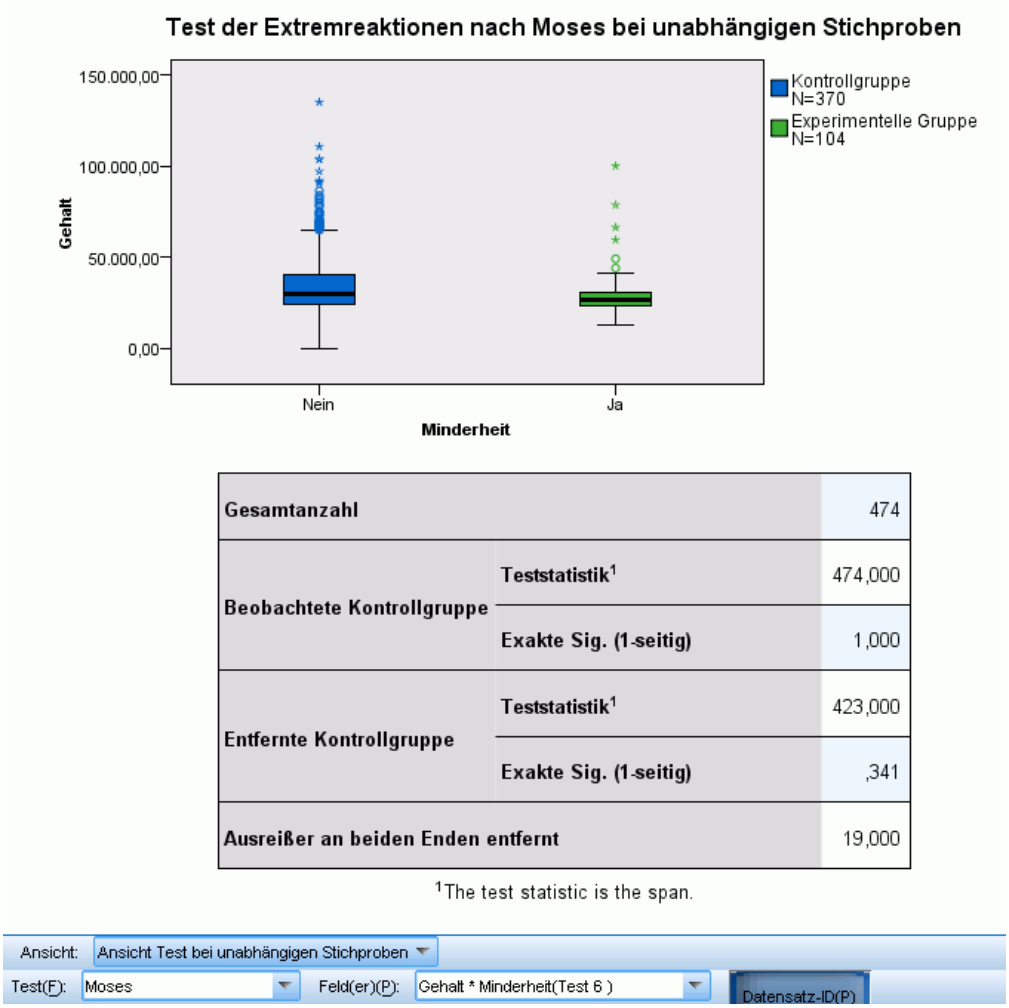


Der Jonckheere-Terpstra-Test zeigt Boxplots und eine Testtabelle an.

- Für jede Kategorie des Gruppierungsfelds werden separate Boxplots angezeigt.
- Die Tabelle zeigt Details zum Test an.

Test auf Extremreaktionen nach Moses

Abbildung 27-43
Ansicht Test bei unabhängigen Stichproben, Test auf Extremreaktionen nach Moses



Gesamtanzahl		474
Beobachtete Kontrollgruppe	**Teststatistik[1]**	474,000
	Exakte Sig. (1-seitig)	1,000
Entfernte Kontrollgruppe	**Teststatistik[1]**	423,000
	Exakte Sig. (1-seitig)	,341
Ausreißer an beiden Enden entfernt		19,000

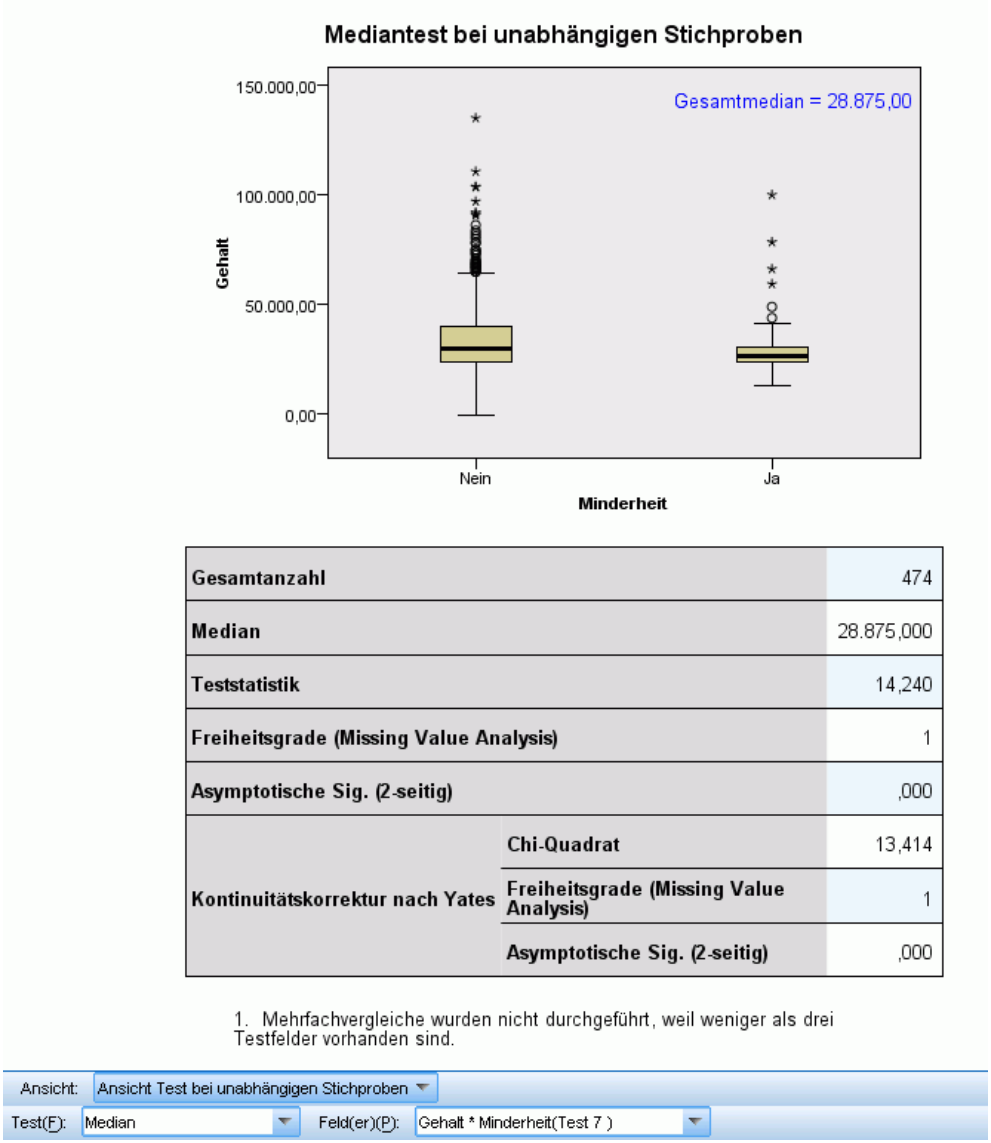
[1]The test statistic is the span.

Der Test auf Extremreaktionen nach Moses zeigt Boxplots und eine Testtabelle an.

- Für jede Kategorie des Gruppierungsfelds werden separate Boxplots angezeigt. Die Punktebeschriftungen können angezeigt oder ausgeblendet werden, indem Sie auf die Schaltfläche Datensatz-ID klicken.
- Die Tabelle zeigt Details zum Test an.

Mediantest

Abbildung 27-44
Ansicht Test bei unabhängigen Stichproben, Mediantest



Gesamtanzahl		474
Median		28.875,000
Teststatistik		14,240
Freiheitsgrade (Missing Value Analysis)		1
Asymptotische Sig. (2-seitig)		,000
Kontinuitätskorrektur nach Yates	**Chi-Quadrat**	13,414
	Freiheitsgrade (Missing Value Analysis)	1
	Asymptotische Sig. (2-seitig)	,000

1. Mehrfachvergleiche wurden nicht durchgeführt, weil weniger als drei Testfelder vorhanden sind.

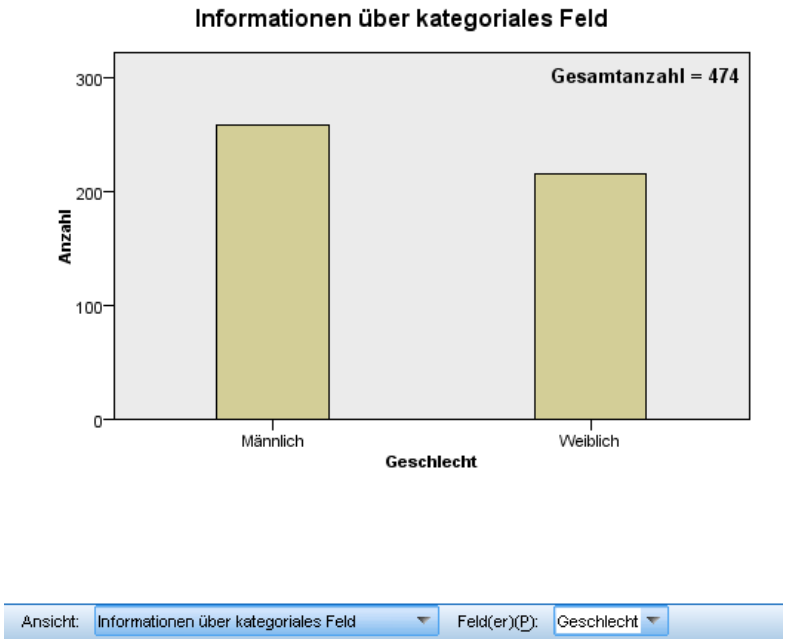
Ansicht: Ansicht Test bei unabhängigen Stichproben
Test(F): Median Feld(er)(P): Gehalt * Minderheit(Test 7)

Der Mediantest zeigt Boxplots und eine Testtabelle an.

- Für jede Kategorie des Gruppierungsfelds werden separate Boxplots angezeigt.
- Die Tabelle zeigt Details zum Test an.

Informationen über kategoriales Feld,

Abbildung 27-45
Informationen über kategoriales Feld

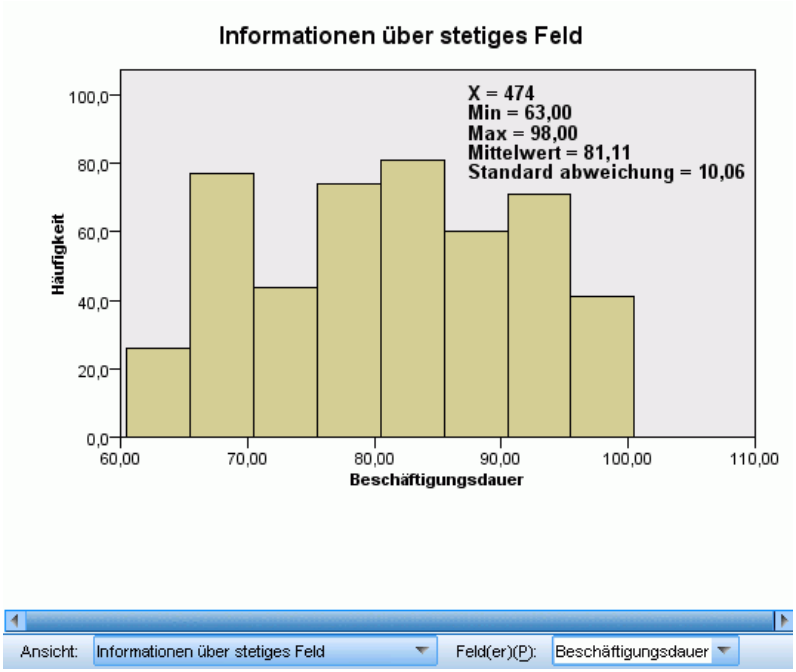


Die Ansicht Informationen über kategoriales Feld zeigt ein Balkendiagramm für das in der Dropdown-Liste Feld(er) ausgewählte kategoriale Feld an. Die Liste der verfügbaren Felder ist auf die kategorialen Felder beschränkt, die im aktuell in der Ansicht Hypothesenübersicht ausgewählten Test verwendet werden.

- Wenn Sie die Maus über einen Balken bewegen, werden in eine QuickInfo die Prozentwerte der Kategorien angezeigt.

Informationen über stetiges Feld,

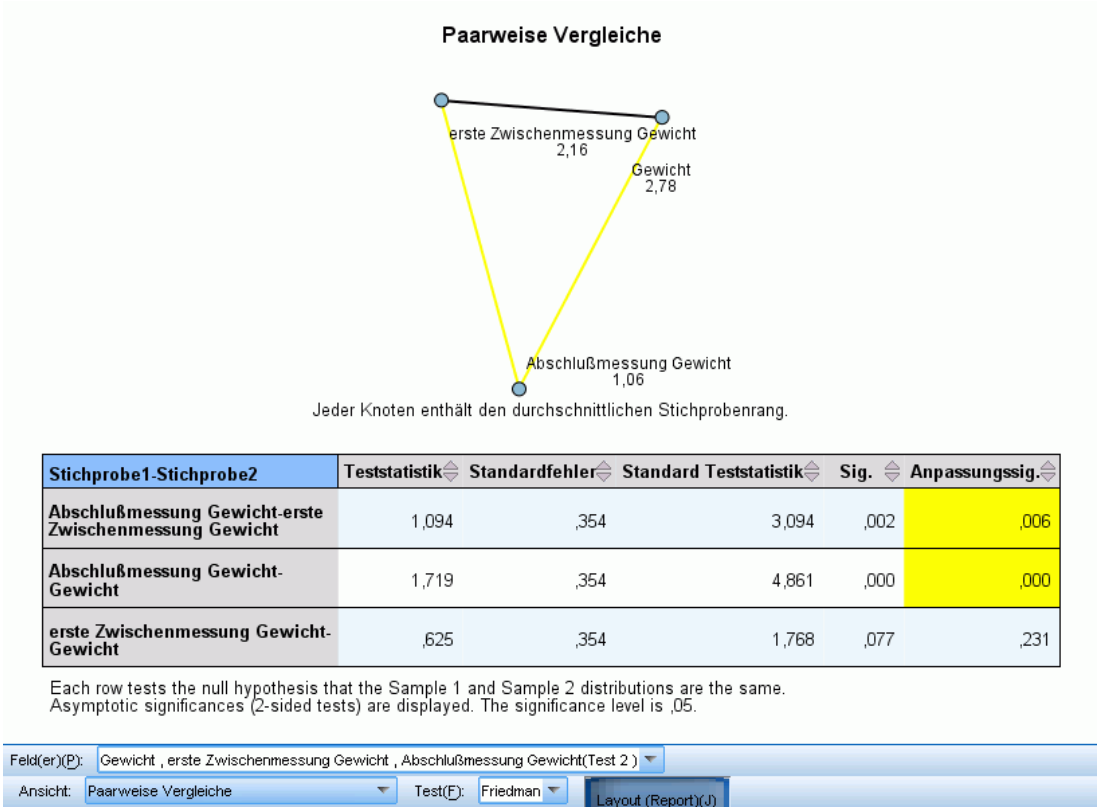
Abbildung 27-46
Informationen über stetiges Feld



Die Ansicht Informationen über stetiges Feld zeigt ein Histogramm für das in der Dropdown-Liste Feld(er) ausgewählte stetige Feld an. Die Liste der verfügbaren Felder ist auf die stetigen Felder beschränkt, die im aktuell in der Ansicht Hypothesenübersicht ausgewählten Test verwendet werden.

Paarweise Vergleiche

Abbildung 27-47
Paarweise Vergleiche



Stichprobe1-Stichprobe2	Teststatistik	Standardfehler	Standard Teststatistik	Sig.	Anpassungssig.
Abschlußmessung Gewicht-erste Zwischenmessung Gewicht	1,094	,354	3,094	,002	,006
Abschlußmessung Gewicht-Gewicht	1,719	,354	4,861	,000	,000
erste Zwischenmessung Gewicht-Gewicht	,625	,354	1,768	,077	,231

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is ,05.

Die Ansicht Paarweise Vergleiche zeigt ein Abstandsnetzwerkdiagramm und eine Vergleichstabelle an, die von nichtparametrischen Tests bei *k* Stichproben erstellt werden, wenn paarweise Mehrfachvergleiche angefordert werden.

- Das Abstandsnetzwerkdiagramm ist eine grafische Darstellung der Vergleichstabelle, in der die Abstände zwischen Knoten im Netzwerk den Unterschieden zwischen Stichproben entsprechen. Gelbe Linien entsprechen statistisch signifikanten Unterschieden, schwarze Linien nichtsignifikanten Unterschieden. Wenn Sie die Maus über eine Linie im Netzwerk bewegen, wird eine QickInfo mit der angepassten Signifikanz des Unterschieds zwischen den durch die Linie verbundenen Knoten angezeigt.
- Die Vergleichstabelle zeigt das numerische Ergebnis aller paarweisen Vergleiche an. Jede Zeile entspricht einem separaten paarweisen Vergleich. Durch Klicken auf eine Spaltenüberschrift werden die Zeilen nach den Werten in dieser Spalte sortiert.

Homogene Untergruppen

Abbildung 27-48
Homogene Untergruppen

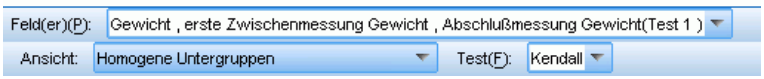
Homogene Untergruppen

		Teilmenge		
		1	2	3
Muster[1]	Abschlußmessung Gewicht	1,063		
	erste Zwischenmessung Gewicht		2,156	
	Gewicht			2,781
Teststatistik		.[2]	.[2]	.[2]
Sig. (2-seitig)		.	.	.
Korrigierte Sig. (2-seitig)		.	.	.

Homogene Untergruppen basieren auf asymptotischen Signifikanzen. Das Signifikanzniveau ist ,05.

[1]Jede Zelle enthält den durchschnittlichen Stichprobenrang.

[2]Unable to compute because the subset contains only one sample.



Die Ansicht Homogene Untergruppen zeigt eine Vergleichstabelle an, die von nichtparametrischen Tests bei *k* Stichproben erstellt wird, wenn schrittweise Step-Down-Mehrfachvergleiche angefordert werden.

- Jede Zeile in der Stichprobengruppe entspricht einer separaten verbundenen Stichprobe (in den Daten als separates Feld dargestellt). Stichproben, die statistisch nicht signifikant unterschiedliche sind, werden in gleichfarbigen Untergruppen gruppiert. Für jede identifizierte Untergruppe ist eine separate Spalte vorhanden. Wenn alle Stichproben statistisch signifikant unterschiedlich sind, ist für jede Stichprobe eine separate Untergruppe vorhanden. Wenn keine der Stichproben statistisch signifikant unterschiedlich ist, ist nur eine Untergruppe vorhanden.
- Für jede Untergruppe mit mehr als einer Stichprobe werden eine Teststatistik, ein Signifikanzwert und ein angepasster Signifikanzwert berechnet.

Zusätzliche Funktionen beim Befehl NPTESTS

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Geben Sie Tests bei einer, bei verbundenen und bei unabhängigen Stichproben in einem einzigen Lauf der Prozedur an.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Veraltete Dialogfelder

Es gibt einige "veraltete" Dialogfelder, die ebenfalls nichtparametrische Tests durchführen. Diese Dialogfelder unterstützen die Funktionen der Option "Exakte Tests".

Chi-Quadrat-Test. Mit diesem Test wird eine Variable nach Kategorien aufgelistet und auf der Grundlage der Differenzen zwischen beobachteten und erwarteten Häufigkeiten eine Chi-Quadrat-Statistik berechnet.

Test auf Binomialverteilung. In diesem Test wird die beobachtete Häufigkeit in jeder Kategorie einer dichotomen Variablen mit den erwarteten Häufigkeiten der binomialen Verteilung verglichen.

Sequenztest. Hiermit können Sie testen, ob zwei Werte einer Variablen in zufälliger Reihenfolge auftreten.

Kolmogorov-Smirnov-Test bei einer Stichprobe. Hierbei wird die beobachtete kumulative Verteilungsfunktion einer Variablen mit einer bestimmten theoretischen Verteilung verglichen. Bei der Verteilung kann es sich um eine Normalverteilung, eine Gleichverteilung, Exponentialverteilung oder um eine Poisson-Verteilung handeln.

Test bei zwei unabhängigen Stichproben. Mit diesem Test können zwei Fallgruppen bei einer Variablen verglichen werden. Dabei stehen die folgenden Tests zur Verfügung: Mann-Whitney-*U*-Test, Kolmogorov-Smirnov-Test bei zwei Stichproben, Test auf Extremreaktionen nach Moses und Sequenzentest nach Wald-Wolfowitz.

Tests bei zwei verbundenen Stichproben. Hiermit können die Verteilungen von zwei Variablen verglichen werden. Dafür stehen der Wilcoxon-Test, der Vorzeichentest und der McNemar-Test zur Verfügung.

Test bei mehreren unabhängigen Stichproben. Hiermit können Sie zwei oder mehrere Fallgruppen bei einer Variablen vergleichen. Dafür stehen der Kruskal-Wallis-H-Test, der Mediantest und der Jonckheere-Terpstra-Test zur Verfügung.

Tests bei mehreren verbundenen Stichproben. Hiermit können Sie die Verteilungen von zwei oder mehr Variablen vergleichen. Dafür stehen der Friedman-Test, Kendall-*W* und Cochrans *Q*-Test zur Verfügung.

Bei allen oben aufgeführten Tests können Quartile, Mittelwert, Standardabweichung, Minimum, Maximum und die Anzahl nichtfehlender Fälle berechnet werden.

Chi-Quadrat-Test

Mit der Prozedur "Chi-Quadrat-Test" können Sie eine Variable nach Kategorien auflisten und eine Chi-Quadrat-Statistik berechnen lassen. Bei diesem Anpassungstest werden die beobachteten und erwarteten Häufigkeiten in allen Kategorien miteinander verglichen. Dadurch wird überprüft, ob entweder alle Kategorien den gleichen Anteil an Werten enthalten oder ob jede Kategorie jeweils einen vom Benutzer festgelegten Anteil an Werten enthält.

Beispiele. Mithilfe des Chi-Quadrat-Tests können Sie bestimmen, ob in einer Tüte mit Gummibärchen die gleiche Anzahl an weißen, grünen, orangefarbenen, roten und gelben Gummibärchen vorhanden sind. Sie können auch prüfen, ob eine Tüte 30% weiße, 17% grüne, 23% orangefarbene, 15% rote und 15% gelbe Gummibärchen enthält.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum und Quartile. Die Anzahl und der Prozentsatz nichtfehlender und fehlender Fälle, die Anzahl der für jede Kategorie beobachteten und erwarteten Fälle, Residuen und die Chi-Quadrat-Statistik.

Daten. Verwenden Sie geordnete oder nichtgeordnete numerische kategoriale Variablen (nominales oder ordinales Niveau der Messwerte). Verwenden Sie zum Umwandeln von String-Variablen in numerische Variablen den Befehl "Automatisch umkodieren" im Menü "Transformieren".

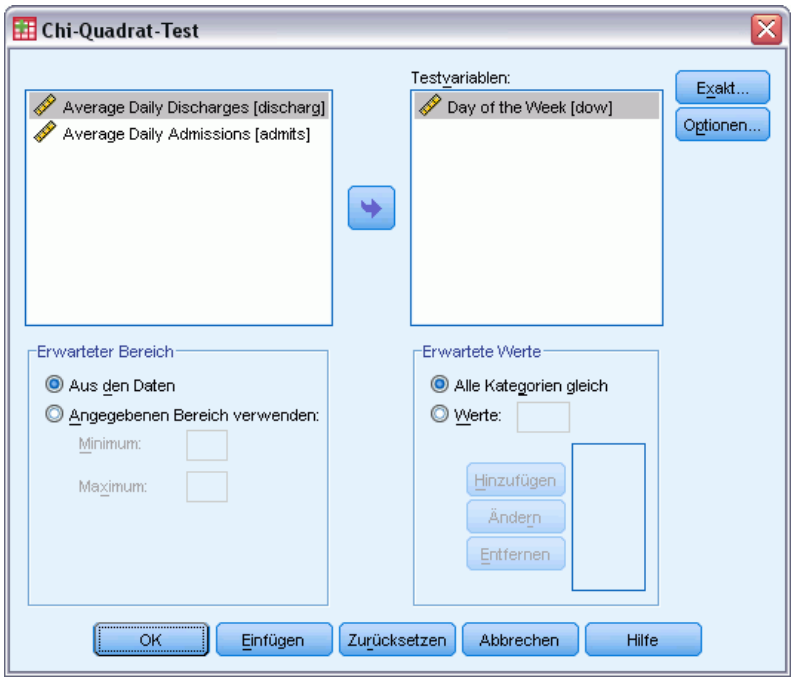
Annahmen. Nichtparametrische Tests erfordern keine Annahmen über die Form der zugrunde liegenden Verteilung. Die Daten werden als zufällige Stichprobe betrachtet. Die erwartete Häufigkeit in jeder Kategorie muss mindestens 1 betragen. Bei höchstens 20% der Kategorien darf die erwartete Häufigkeit unter 5 liegen.

So lassen Sie einen Chi-Quadrat-Test berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:

  Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > Chi-Quadrat...

Abbildung 27-49
Dialogfeld "Chi-Quadrat-Test"



- Wählen Sie mindestens eine Testvariable aus. Mit jeder Variablen wird ein separater Test erzeugt.
- Wenn Sie auf Optionen klicken, können Sie deskriptive Statistiken und Quartile abrufen sowie festlegen, wie fehlende Werte verarbeitet werden.

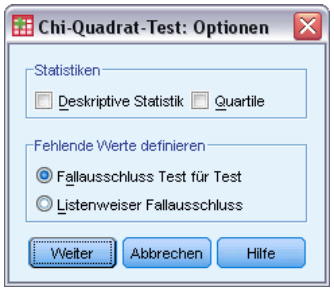
Chi-Quadrat-Test: erwarteter Bereich und erwartete Werte

Erwarteter Bereich. In der Standardeinstellung wird jeder einzelne Wert einer Variablen als eine Kategorie definiert. Zum Aufstellen von Kategorien in einem bestimmten Bereich wählen Sie AngegebenenBereich verwenden und geben Sie für die obere und die untere Grenze jeweils einen ganzzahligen Wert an. Für jeden ganzzahligen Wert in dem eingeschlossenen Bereich wird eine Kategorie aufgestellt, wobei Fälle mit Werten außerhalb der angegebenen Grenzen ausgeschlossen werden. Wenn Sie zum Beispiel für das Minimum den Wert 1 und für das Maximum den Wert 4 angeben, werden für den Chi-Quadrat-Test nur die Werte von 1 bis 4 verwendet.

Erwartete Werte. In der Standardeinstellung sind die erwarteten Werte für alle Kategorien gleich. Die erwarteten Anteile der Kategorien können vom Benutzer festgelegt werden. Wählen Sie Werte aus. Geben Sie für jede Kategorie der Testvariablen einen Wert größer als 0 ein und klicken Sie dann auf Hinzufügen. Jeder neu eingegebene Wert wird am Ende der Werteliste angezeigt. Die Reihenfolge der Werte ist von Bedeutung. Sie entspricht der aufsteigenden Folge der Kategoriewerte für die Testvariable. Der erste Wert in der Liste entspricht dem niedrigsten Gruppenwert der Testvariablen, der letzte Wert entspricht dem höchsten Wert. Die Elemente der Werteliste werden summiert. Anschließend wird jeder Wert durch diese Summe dividiert, um den Anteil der in der entsprechenden Kategorie erwarteten Fälle zu berechnen. So ergibt eine Werteliste mit 3, 4, 5 und 4 beispielsweise die erwarteten Anteile 3/16, 4/16, 5/16 und 4/16.

Chi-Quadrat-Test: Optionen

Abbildung 27-50
Dialogfeld "Chi-Quadrat-Test: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle angezeigt.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (Chi-Quadrat-Test)

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl CHISQUARE können verschiedene Minimal- und Maximalwerte sowie erwartete Häufigkeiten für verschiedene Variablen angegeben werden.
- Mit dem Unterbefehl EXPECTED kann eine Variable bei verschiedenen erwarteten Häufigkeiten getestet werden oder es können verschiedene Bereiche verwendet werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Test auf Binomialverteilung

Mit der Prozedur "Test auf Binomialverteilung" können Sie die beobachteten Häufigkeiten der beiden Kategorien einer dichotomen Variablen mit den Häufigkeiten vergleichen, die unter einer Binomialverteilung mit einem angegebenen Wahrscheinlichkeitsparameter zu erwarten sind. In der Standardeinstellung ist der Wahrscheinlichkeitsparameter für beide Gruppen auf 0,5 gesetzt. Zum Ändern der Wahrscheinlichkeiten können Sie einen Testanteil für die erste Gruppe angeben. Die Wahrscheinlichkeit für die zweite Gruppe beträgt 1 minus der für die erste Gruppe angegebenen Wahrscheinlichkeit.

Beispiel. Wenn Sie eine Münze werfen, ist die Wahrscheinlichkeit, dass diese mit dem Kopf nach oben zu liegen kommt, gleich 1/2. Auf der Grundlage dieser Hypothese wird nun eine Münze 40mal geworfen, wobei die Ergebnisse aufgezeichnet werden (Kopf oder Zahl). Der Test auf Binomialverteilung könnte dann beispielsweise ergeben, dass 3/4 der Würfe "Kopf" waren und das beobachtete Signifikanzniveau gering ist (0,0027). Diese Ergebnisse zeigen an, dass die Wahrscheinlichkeit für "Kopf" nicht 1/2 beträgt und die Münze somit wahrscheinlich manipuliert ist.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile.

Daten. Die getesteten Variablen müssen numerisch und dichotom sein. Verwenden Sie zum Umwandeln von String-Variablen in numerische Variablen den Befehl "Automatisch umkodieren" im Menü "Transformieren". **Dichotome Variablen** sind Variablen, die nur zwei mögliche Werte annehmen können: *ja* oder *nein*, *wahr* oder *falsch*, 0 oder 1 usw. Der erste in dem Daten-Set gefundene Wert definiert die erste Gruppe, der andere Wert definiert die zweite Gruppe. Wenn die Varaiblen nicht dichotom sind, müssen Sie einen Trennwert angeben. Durch den Trennwert werden Fälle mit Werten unter oder gleich dem Trennwert der ersten Gruppe und alle anderen Fälle der zweiten Gruppe zugeordnet.

Annahmen. Nichtparametrische Tests erfordern keine Annahmen über die Form der zugrunde liegenden Verteilung. Die Daten werden als zufällige Stichprobe betrachtet.

So lassen Sie einen Test auf Binomialverteilung berechnen:

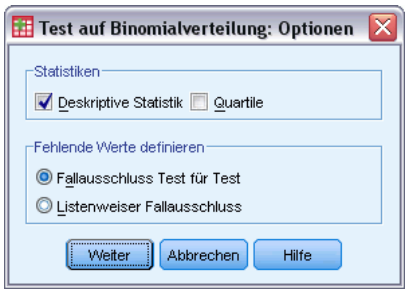
▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > Binomial...

Abbildung 27-51
Dialogfeld "Test auf Binomialverteilung"

- Wählen Sie mindestens eine numerische Testvariable.
- Wenn Sie auf Optionen klicken, können Sie deskriptive Statistiken und Quartile abrufen sowie festlegen, wie fehlende Werte verarbeitet werden.

Optionen für den Test auf Binomialverteilung

Abbildung 27-52
Dialogfeld "Test auf Binomialverteilung: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle angezeigt.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.** Fälle mit fehlenden Werten für eine beliebige getestete Variable werden von allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (Test auf Binomialverteilung)

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl `BINOMIAL` können bestimmte Gruppen ausgewählt und andere Gruppen ausgeschlossen werden, wenn eine Variable über mehr als zwei Kategorien verfügt.
- Mit dem Unterbefehl `BINOMIAL` können verschiedene Trennwerte oder Wahrscheinlichkeiten für verschiedene Variablen angeben werden.
- Mit dem Unterbefehl `EXPECTED` kann dieselbe Variable bei verschiedenen Trennwerten oder Wahrscheinlichkeiten getestet werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Sequenzentest

Mit der Prozedur "Sequenzentest" können Sie testen, ob zwei Werte einer Variablen in zufälliger Reihenfolge auftreten. Eine Sequenz ist eine Folge von gleichen Beobachtungen. Eine Stichprobe mit zu vielen oder zu wenigen Sequenzen legt nahe, dass die Stichprobe nicht zufällig ist.

Beispiele. Es werden 20 Personen befragt, ob sie ein bestimmtes Produkt kaufen würden. Die angenommene zufällige Auswahl der Stichprobe wäre ernsthaft zu bezweifeln, wenn alle 20 Personen demselben Geschlecht angehören würden. Mit dem Sequenzentest kann bestimmt werden, ob die Stichprobe zufällig entnommen wurde.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile.

Daten. Die Variablen müssen numerisch sein. Verwenden Sie zum Umwandeln von String-Variablen in numerische Variablen den Befehl "Automatisch umkodieren" im Menü "Transformieren".

Annahmen. Nichtparametrische Tests erfordern keine Annahmen über die Form der zugrunde liegenden Verteilung. Verwenden Sie Stichproben aus stetigen Wahrscheinlichkeitsverteilungen.

So lassen Sie einen Sequenzentest berechnen:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > Sequenzen...

Abbildung 27-53
Hinzufügen eines benutzerdefinierten Trennwerts

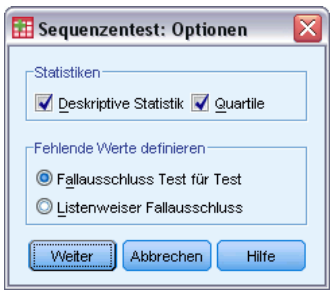
- Wählen Sie mindestens eine numerische Testvariable.
- Wenn Sie auf Optionen klicken, können Sie deskriptive Statistiken und Quartile abrufen sowie festlegen, wie fehlende Werte verarbeitet werden.

Sequenzentest: Trennwert

Trennwert. Hier wird ein Trennwert zum Dichotomisieren der gewählten Variablen angegeben. Sie können den beobachteten Mittelwert, den Median, den Modalwert oder einen angegebenen Wert als Trennwert wählen. Fälle mit Werten kleiner als der Trennwert werden einer Gruppe, Fälle mit Werten größer oder gleich dem Trennwert einer anderen Gruppe zugeordnet. Für jeden gewählten Trennwert wird ein Test ausgeführt.

Sequenzentest: Optionen

Abbildung 27-54
Dialogfeld "Sequenzentest: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle angezeigt.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (Sequenzentest)

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl `RUNS` können verschiedene Trennwerte für verschiedene Variablen angegeben werden.
- Mit dem Unterbefehl `RUNS` kann dieselbe Variable mit verschiedenen benutzerdefinierten Trennwerten getestet werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kolmogorov-Smirnov-Test bei einer Stichprobe

Mit dem Kolmogorov-Smirnov-Test bei einer Stichprobe (Anpassungstest) wird die beobachtete kumulative Verteilungsfunktion für eine Variable mit einer festgelegten theoretischen Verteilung verglichen, die eine Normalverteilung, eine Gleichverteilung, eine Poisson-Verteilung oder Exponentialverteilung sein kann. Das Kolmogorov-Smirnov-*Z* wird aus der größten Differenz (in Absolutwerten) zwischen beobachteten und theoretischen kumulativen Verteilungsfunktionen berechnet. Mit diesem Test für die Güte der Anpassung wird getestet, ob die Beobachtung wahrscheinlich aus der angegebenen Verteilung stammt.

Beispiel.Für viele parametrische Tests sind normalverteilte Variablen erforderlich. Mit dem Kolmogorov-Smirnov-Anpassungstest kann getestet werden, ob eine Variable, zum Beispiel *Einkommen*, normalverteilt ist.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile.

Daten. Die Variablen müssen auf Intervall- oder Verhältnis-Messniveau quantitativ sein.

Annahmen.Für den Kolmogorov-Smirnov-Test wird angenommen, dass die Parameter der zu testenden Verteilung im voraus angegeben wurden. Mit dieser Prozedur werden die Parameter aus der Stichprobe geschätzt. Der Mittelwert und die Standardabweichung der Stichprobe sind die Parameter für eine Normalverteilung. Minimum und Maximum der Stichprobe definieren die Spannweite der Gleichverteilung, und der Mittelwert der Stichprobe ist der Parameter für die Poisson-Verteilung sowie der Parameter für die Exponentialverteilung. Die Stärke des Tests, Abweichungen von der hypothetischen Verteilung zu erkennen, kann dabei deutlich verringert werden. Wenn Sie einen Test gegen eine Normalverteilung mit geschätzten Parametern

durchführen möchten, sollten Sie den Kolmogorov-Smirnov-Test mit der Korrektur nach Lilliefors (in der Prozedur "Explorative Datenanalyse") in Betracht ziehen.

So berechnen Sie einen Kolmogorov-Smirnov-Anpassungstest:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > K-S bei einer Stichprobe...

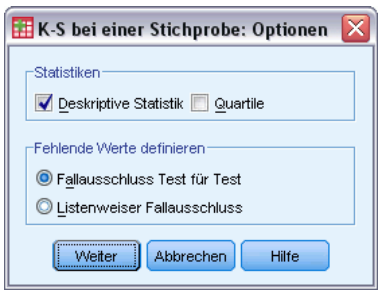
Abbildung 27-55
Dialogfeld "Kolmogorov-Smirnov-Test bei einer Stichprobe"

- Wählen Sie mindestens eine numerische Testvariable. Mit jeder Variablen wird ein separater Test erzeugt.

- Wenn Sie auf Optionen klicken, können Sie deskriptive Statistiken und Quartile abrufen sowie festlegen, wie fehlende Werte verarbeitet werden.

K-S bei einer Stichprobe: Optionen

Abbildung 27-56
Dialogfeld "K-S bei einer Stichprobe: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle angezeigt.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (Kolmogorov-Smirnov-Anpassungstest)

Mit der Befehlssyntax-Sprache können Sie auch die Parameter der zu testenden Verteilung angeben (mit dem Unterbefehl `K-S`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Tests bei zwei unabhängigen Stichproben

Die Prozedur "Test bei zwei unabhängigen Stichproben" vergleicht zwei Gruppen von Fällen von einer Variablen.

Beispiel. Es wurden neue Zahnspangen entwickelt, die bequemer sein sollen, besser aussehen und zu einem schnelleren Erfolg beim Richten der Zähne führen sollen. Um festzustellen, ob die neuen Spangen so lange wie die alten getragen werden müssen, wurden willkürlich 10 Kinder zum Tragen der alten Zahnspangen und weitere 10 Kinder zum Tragen der neuen Spangen ausgewählt. Anhand des Mann-Whitney-*U*-Tests stellen Sie eventuell fest, dass die neuen Spangen im Durchschnitt nicht so lange wie die alten Spangen getragen werden mussten.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile. Tests: Mann-Whitney-*U*-Test, Extremreaktionen nach Moses, Kolmogorov-Smirnov-*Z*-Test, Sequenztest nach Wald-Wolfowitz.

Daten. Verwenden Sie numerische Variablen, die geordnet werden können.

Annahmen. Verwenden Sie unabhängige Zufallsstichproben. Mit dem Mann-Whitney-*U*-Test wird die Gleichheit von zwei Verteilungen getestet. Um damit Unterschiede in der Lage von zwei Verteilungen zu testen, muss davon ausgegangen werden, dass die Verteilungen dieselbe Form haben.

So lassen Sie Tests bei zwei unabhängigen Stichproben berechnen:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > Zwei unabhängige Stichproben...

Abbildung 27-57
Dialogfeld "Tests bei zwei unabhängigen Stichproben"

- Wählen Sie mindestens eine numerische Variable aus.
- Wählen Sie eine Gruppenvariable aus und klicken Sie auf Gruppen definieren, um die Datei in zwei Gruppen oder Stichproben aufzuteilen.

Typen von Tests bei zwei unabhängigen Stichproben

Welche Tests durchführen?Mithilfe von vier Tests können Sie überprüfen, ob zwei unabhängige Stichproben (Gruppen) aus derselben Grundgesamtheit stammen.

Der **Mann-Whitney-U-Test**ist der am häufigsten verwendete Test bei zwei unabhängigen Stichproben. Er ist äquivalent zum Wilcoxon-Rangsummentest und dem Kruskal-Wallis-Test für zwei Gruppen. Mit dem Mann-Whitney-U-Test wird überprüft, ob zwei beprobte Grundgesamtheiten die gleiche Lage besitzen. Die Beobachtungen aus beiden Gruppen werden kombiniert und in eine gemeinsame Reihenfolge gebracht, wobei im Falle von Rangbindungen der durchschnittliche Rang vergeben wird. Die Anzahl der Bindungen sollte im Verhältnis zur Gesamtanzahl der Beobachtungen klein sein. Wenn die Grundgesamtheiten in der Lage identisch sind, sollten die Ränge zufällig zwischen den beiden Stichproben gemischt werden. Im Test wird berechnet, wie oft ein Wert aus Gruppe 1 einem Wert aus Gruppe 2 und wie oft ein Wert aus Gruppe 2 einem Wert aus Gruppe 1 vorangeht. Die Mann-Whithney-*U*-Statistik ist die kleinere dieser beiden Zahlen. Die Statistik der Wilcoxon-Rangsumme *W* wird ebenfalls angezeigt. *W* ist die Summe der Ränge für die Gruppe mit dem kleineren mittleren Rang. Wenn die Gruppen denselben mittleren Rang aufweisen, wird die Rangsumme der Gruppe verwendet, die im Dialogfeld "Zwei unabhängige Stichproben: Gruppen definieren" weiter unten genannt wird.

Der **Kolmogorov-Smirnov-Z-Test** und der **Sequenztest nach Wald-Wolfowitz** stellen eher allgemeine Tests dar, die sowohl Unterschiede in den Lagen als auch in den Formen der Verteilungen erkennen. Der Test nach Kolmogorov-Smirnov arbeitet auf der Grundlage der maximalen absoluten Differenz zwischen den beobachteten kumulativen Verteilungsfunktionen für beide Stichproben. Wenn diese Differenz signifikant groß ist, werden die beiden Verteilungen als verschieden betrachtet. Der Sequenztest nach Wald-Wolfowitz kombiniert die Beobachtungen

aus beiden Gruppen und ordnet ihnen einen Rang zu. Wenn die beiden Stichproben aus derselben Grundgesamtheit stammen, müssen die beiden Gruppen in der Rangverteilung zufällig gestreut sein.

Der **Test "Extremreaktionen nach Moses"** setzt voraus, dass die experimentelle Variable einige Subjekte in der einen Richtung und andere Subjekte in der entgegengesetzten Richtung beeinflußt. In diesem Test wird auf extreme Antworten im Vergleich zu einer Kontrollgruppe geprüft. Dieser Test konzentriert sich auf die Spannweite der Kontrollgruppe und ist ein Maß dafür, wie stark die Spannweite durch die extremen Werte in der experimentellen Gruppe beeinflußt wird, wenn sie mit der Kontrollgruppe verbunden werden. Die Kontrollgruppe wird durch den Wert der Gruppe 1 im Dialogfeld "Zwei unabhängige Stichproben: Gruppen definieren" bestimmt. Die Beobachtungen aus beiden Gruppen werden kombiniert und einem Rang zugeordnet. Die Spanne der Kontrollgruppe wird als die Differenz zwischen den Rängen der größten und kleinsten Werte in der Kontrollgruppe plus 1 berechnet. Da zufällige Ausreißer den Bereich der Spannweite leicht verzerren können, werden 5 % der Kontrollfälle automatisch an jedem Ende weggelassen.

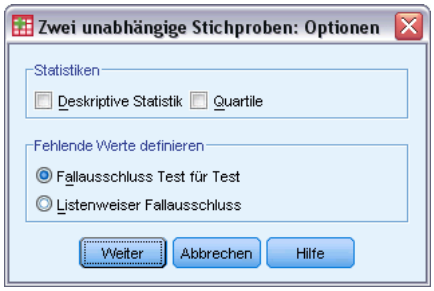
Zwei unabhängige Stichproben: Gruppen definieren

Abbildung 27-58
Dialogfeld "Zwei unabhängige Stichproben: Gruppen definieren"

Um die Datei in zwei Gruppen oder Stichproben aufzuteilen, geben Sie eine ganze Zahl für Gruppe 1 und eine weitere Zahl für Gruppe 2 ein. Fälle mit anderen Werten werden aus der Analyse ausgeschlossen.

Tests bei zwei unabhängigen Stichproben – Optionen

Abbildung 27-59
Dialogfeld "Zwei unabhängige Stichproben: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle an.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.** Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.** Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (Tests bei zwei unabhängigen Stichproben)

Mit dem Unterbefehl `MOSES` der Befehlssyntax-Sprache kann die Anzahl der Fälle angegeben werden, die für den Moses-Test getrimmt werden sollen.

Siehe *Befehlssyntaxreferenz* für die vollständigen Syntaxinformationen.

Tests bei zwei verbundenen Stichproben

Die Prozedur "Tests bei zwei verbundenen Stichproben" vergleicht die Verteilungen von zwei Variablen.

Beispiel. Erhalten Familien, die ihr Haus verkaufen, im allgemeinen den geforderten Preis? Wenn Sie den Wilcoxon-Test auf die Daten von 10 Häusern anwenden, könnten Sie beispielsweise feststellen, dass sieben Familien weniger als den geforderten Preis, eine Familie mehr als den geforderten Preis und zwei Familien den geforderten Preis erhielten.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile. Tests: Wilcoxon-Test, Vorzeichentest, McNemar. Wenn die Option "Exakte Tests" installiert ist (nur unter Windows-Betriebssystemen verfügbar) steht außerdem der Rand-Homogenitätstest zur Verfügung.

Daten. Verwenden Sie numerische Variablen, die geordnet werden können.

Annahmen. Obwohl keine bestimmten Verteilungen für die beiden Variablen vorausgesetzt werden, wird die Verteilung der Grundgesamtheit der gepaarten Differenzen als symmetrisch angenommen.

So lassen Sie Tests bei zwei verbundenen Stichproben berechnen:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > Zwei verbundene Stichproben...

Abbildung 27-60
Dialogfeld "Tests bei zwei verbundenen Stichproben"

▶ Wählen Sie mindestens ein Variablenpaar aus.

Typen von Tests bei zwei verbundenen Stichproben

Die Tests in diesem Abschnitt vergleichen die Verteilungen von zwei verbundenen Variablen. Der geeignete Test hängt vom jeweiligen Datentyp ab.

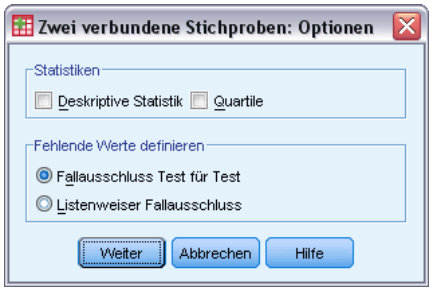
Falls Ihre Daten stetig sind, verwenden Sie den Vorzeichentest oder den Wilcoxon-Test. Der **Vorzeichentest** berechnet für alle Fälle die Differenzen zwischen den beiden Variablen und klassifiziert sie als positiv, negativ oder verbunden. Falls die beiden Variablen ähnlich verteilt sind, unterscheidet sich die Zahl der positiven und negativen Differenzen nicht signifikant. Der **Wilcoxon-Test** berücksichtigt sowohl Informationen über Vorzeichen der Differenzen als auch die Größe der Differenzen zwischen den Paaren. Da der Wilcoxon-Test mehr Informationen über die Daten aufnimmt, kann er mehr leisten als der Vorzeichentest.

Falls Sie mit binären Daten arbeiten, verwenden Sie den **McNemar-Test**. Dieser Test wird üblicherweise bei Messwiederholungen verwendet, wenn jede Antwort eines Subjektes doppelt abgerufen wird, einmal bevor ein festgelegtes Ereignis eintritt und einmal danach. Der McNemar-Test bestimmt, ob die Antwortrate am Anfang (vor dem Ereignis) gleich der Antwortrate am Ende (nach dem Ereignis) ist. Dieser Test ist für das Erkennen von Änderungen bei Antworten nützlich, die durch experimentelle Einflußnahme in sogenannten "Vorher-und-nachher-Designs" entstanden sind.

Falls Sie mit kategorialen Daten arbeiten, verwenden Sie den **Rand-Homogenitätstest**. Dieser Test ist eine Erweiterung des McNemar-Tests von binären Variablen auf multinomiale Variablen. Mithilfe dieses Tests wird unter Verwendung der Chi-Quadrat-Verteilung überprüft, ob Änderungen bei den Antworten vorliegen. Dies ist nützlich, um zu ermitteln, ob die Änderungen in sogenannten "Vorher-und-nachher-Designs" durch experimentelle Einflußnahme verursacht werden. Der Rand-Homogenitätstest ist nur verfügbar, wenn Sie die Option Exact Tests installiert haben.

Optionen für Tests bei zwei verbundenen Stichproben

Abbildung 27-61
Dialogfeld "Zwei verbundene Stichproben: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle an.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (zwei verbundene Stichproben)

Mit der Befehlssyntax-Sprache können Sie außerdem eine Variable mit jeder Variable auf einer Liste überprüfen.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Tests bei mehreren unabhängigen Stichproben

Mit der Prozedur "Tests bei mehreren unabhängigen Stichproben" werden zwei oder mehrere Fallgruppen einer Variablen verglichen.

Beispiel. Unterscheiden sich 100-Watt-Glühlampen dreier Marken in ihrer durchschnittlichen Lebensdauer? Mit der einfaktoriellen Varianzanalyse nach Kruskal-Wallis könnten Sie feststellen, dass die drei Marken sich in ihrer durchschnittlichen Lebensdauer unterscheiden.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile. Tests: Kruskal-Wallis-*H*, Median.

Daten. Verwenden Sie numerische Variablen, die geordnet werden können.

Annahmen. Verwenden Sie unabhängige Zufallsstichproben. Für den Kruskal-Wallis-*H*-Test sind Stichproben erforderlich, die sich in ihrer Form ähneln.

So lassen Sie Tests für mehrere unabhängige Stichproben berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > K unabhängige Stichproben...

Abbildung 27-62
Festlegung des Mediantests

- Wählen Sie mindestens eine numerische Variable aus.
- Wählen Sie eine Gruppenvariable aus und klicken Sie auf Bereich definieren, um die ganzzahligen Minimal- und Maximalwerte der Gruppenvariablen festzulegen.

Tests bei mehreren unabhängigen Stichproben: Welche Tests durchführen?

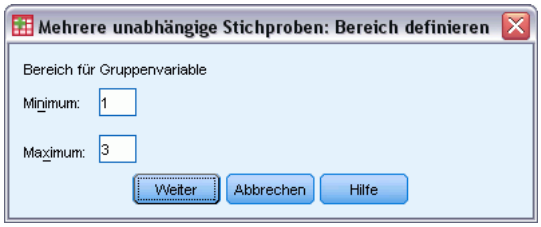
Sie können mit drei Tests bestimmen, ob mehrere unabhängige Stichproben aus derselben Grundgesamtheit stammen. Mit dem Kruskal-Wallis-*H*-Test, dem Mediantest und dem Jonckheere-Terpstra-Test können Sie prüfen, ob mehrere unabhängige Stichproben aus derselben Grundgesamtheit stammen.

Der **Kruskal-Wallis-H-Test**, eine Erweiterung des Mann-Whitney-*U*-Tests, ist die nichtparametrische Entsprechung der einfaktoriellen Varianzanalyse und erkennt Unterschiede in der Lage der Verteilung. Der **Mediantest**, der allgemeiner, aber nicht so leistungsstark ist, erkennt Unterschiede von Verteilungen in Lage und Form. Der Kruskal-Wallis-*H*-Test und der Mediantest setzen voraus, dass keine *a-priori*-Ordnung der *k* Grundgesamtheiten vorliegt, aus denen die Stichproben gezogen werden.

Wenn eine natürliche *a-priori*-Ordnung (aufsteigend oder absteigend) der *k* Grundgesamtheiten *besteht*, ist der **Jonckheere-Terpstra-Test** leistungsfähiger. Die *k* Grundgesamtheiten könnten zum Beispiel *k* ansteigende Temperaturen darstellen. Die Hypothese, dass unterschiedliche Temperaturen die gleiche Verteilung von Antworten erzeugen, wird gegen die Alternative getestet, dass mit Zunahme der Temperatur die Größe der Antwort zunimmt. Hierbei ist die alternative Hypothese geordnet, deshalb ist der Jonckheere-Terpstra-Test für diesen Test am besten geeignet. Der Jonckheere-Terpstra-Test ist nur verfügbar, wenn Sie das Erweiterungsmodul Exact Tests installiert haben.

Tests bei mehreren unabhängigen Stichproben: Bereich definieren

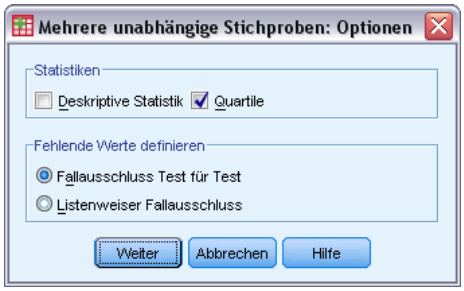
Abbildung 27-63
Dialogfeld "Mehrere unabhängige Stichproben: Bereich definieren"



Um den Bereich zu definieren, geben Sie für Minimum und Maximum ganzzahlige Werte ein, die der niedrigsten und höchsten Kategorie der Gruppenvariablen entsprechen. Der Minimalwert muss kleiner sein als der Maximalwert. Wenn Sie zum Beispiel als Minimum 1 und als Maximum 3 angeben, werden nur die ganzzahligen Werte von 1 bis 3 verwendet. Das Minimum muss kleiner als das Maximum sein. Beide Werte müssen angegeben werden.

Tests bei mehreren unabhängigen Stichproben: Optionen

Abbildung 27-64
Dialogfeld "Mehrere unabhängige Stichproben: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle an.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (K unabhängige Stichproben)

In der Befehlssyntax-Sprache haben Sie außerdem die Möglichkeit, mit dem Unterbefehl `MEDIAN` einen anderen Wert als den beobachteten Median für den Mediantest festzulegen.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Tests bei mehreren verbundenen Stichproben

Bei der Prozedur "Tests bei mehreren verbundenen Stichproben" werden die Verteilungen von zwei oder mehr Variablen verglichen.

Beispiel. Genießen die Berufsgruppen Ärzte, Anwälte, Polizisten oder Lehrer in der Öffentlichkeit ein unterschiedliches Ansehen? Zehn Personen wurden gebeten, diese vier Berufsgruppen in der Reihenfolge ihres Ansehens anzuordnen. Der Test nach Friedman zeigt, dass diese vier Berufsgruppen in der Öffentlichkeit tatsächlich ein unterschiedliches Ansehen genießen.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile. Tests: Friedman, Kendall-*W* und Cochran-*Q*.

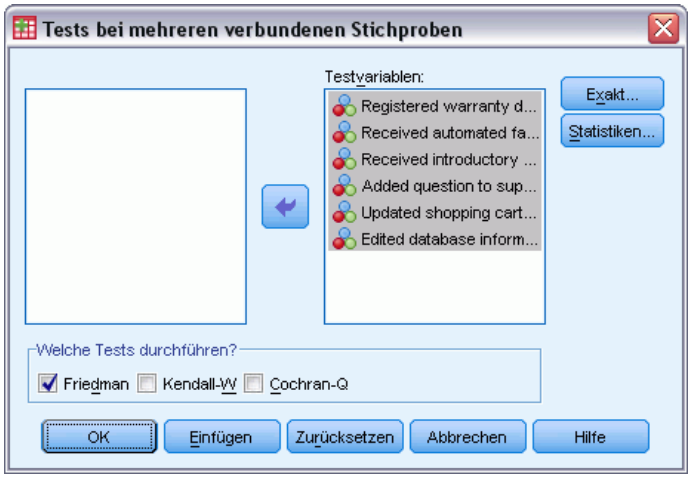
Daten. Verwenden Sie numerische Variablen, die geordnet werden können.

Annahmen. Nichtparametrische Tests erfordern keine Annahmen über die Form der zugrunde liegenden Verteilung. Verwenden Sie abhängige Zufallsstichproben.

So lassen Sie Tests bei mehreren verbundenen Stichproben berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > K verbundene Stichproben...

Abbildung 27-65
Auswahl von "Cochran" als Testtyp



- Wählen Sie zwei oder mehr numerische Testvariablen aus.

Tests bei mehreren verbundenen Stichproben: Welche Tests durchführen?

Sie können die Verteilung von verschiedenen verbundenen Variablen mit drei Tests vergleichen.

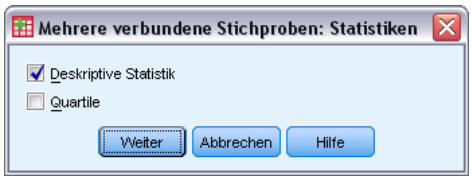
Der **Test** nach **Friedman** stellt das nichtparametrische Äquivalent eines Designs mit Messwiederholungen bei einer Stichprobe bzw. eine Zweifach-Varianzanalyse mit einer Beobachtung pro Zelle dar. Der Friedman-Test überprüft die Nullhypothese, wonach die *k* verbundenen Variablen aus derselben Grundgesamtheit stammen. Für jeden Fall werden den *k* Variablen Rangzahlen von 1 bis *k* zugewiesen. Die Teststatistik wird auf der Grundlage dieser Ränge durchgeführt.

Das **Kendall-W** stellt eine Normalisierung der Statistik nach Friedman dar. Das Kendall-*W* kann als Konkordanzkoeffizient interpretiert werden, der ein Maß für die Übereinstimmung der Prüfer darstellt. Jeder Fall ist ein Richter oder Prüfer, und jede Variable ist ein zu beurteilendes Objekt oder eine zu beurteilende Person. Die Rangsumme jeder Variablen wird berechnet. Das Kendall-*W* liegt im Bereich von 0 (keine Übereinstimmung) bis 1 (vollständige Übereinstimmung).

Das **Cochran-Q** entspricht vollständig dem Friedman-Test. Es wird jedoch angewendet, wenn alle Antworten binär sind. Dieser Test stellt eine Erweiterung des McNemar-Tests auf *k* Stichproben dar. Das Cochran-*Q* überprüft die Hypothese, dass mehrere verbundene dichotome Variablen denselben Mittelwert aufweisen. Die Variablenwerte beziehen sich auf dasselbe Individuum oder auf zusammengehörige Individuen.

Tests bei mehreren verbundenen Stichproben: Statistiken

Abbildung 27-66
Dialogfeld "Mehrere verbundene Stichproben: Statistiken"



Sie können Statistiken auswählen.

- **Deskriptive Statistik.** Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle an.
- **Quartile.** Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (K verbundene Stichproben)

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Test auf Binomialverteilung

Mit der Prozedur "Test auf Binomialverteilung" können Sie die beobachteten Häufigkeiten der beiden Kategorien einer dichotomen Variablen mit den Häufigkeiten vergleichen, die unter einer Binomialverteilung mit einem angegebenen Wahrscheinlichkeitsparameter zu erwarten sind. In der Standardeinstellung ist der Wahrscheinlichkeitsparameter für beide Gruppen auf 0,5 gesetzt. Zum Ändern der Wahrscheinlichkeiten können Sie einen Testanteil für die erste Gruppe angeben. Die Wahrscheinlichkeit für die zweite Gruppe beträgt 1 minus der für die erste Gruppe angegebenen Wahrscheinlichkeit.

Beispiel. Wenn Sie eine Münze werfen, ist die Wahrscheinlichkeit, dass diese mit dem Kopf nach oben zu liegen kommt, gleich 1/2. Auf der Grundlage dieser Hypothese wird nun eine Münze 40mal geworfen, wobei die Ergebnisse aufgezeichnet werden (Kopf oder Zahl). Der Test auf Binomialverteilung könnte dann beispielsweise ergeben, dass 3/4 der Würfe "Kopf" waren und das beobachtete Signifikanzniveau gering ist (0,0027). Diese Ergebnisse zeigen an, dass die Wahrscheinlichkeit für "Kopf" nicht 1/2 beträgt und die Münze somit wahrscheinlich manipuliert ist.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile.

Daten. Die getesteten Variablen müssen numerisch und dichotom sein. Verwenden Sie zum Umwandeln von String-Variablen in numerische Variablen den Befehl "Automatisch umkodieren" im Menü "Transformieren". **Dichotome Variablen** sind Variablen, die nur zwei mögliche Werte annehmen können: *ja* oder *nein*, *wahr* oder *falsch*, 0 oder 1 usw. Der erste in dem Daten-Set gefundene Wert definiert die erste Gruppe, der andere Wert definiert die zweite Gruppe. Wenn die Varaiblen nicht dichotom sind, müssen Sie einen Trennwert angeben. Durch den Trennwert werden Fälle mit Werten unter oder gleich dem Trennwert der ersten Gruppe und alle anderen Fälle der zweiten Gruppe zugeordnet.

Annahmen. Nichtparametrische Tests erfordern keine Annahmen über die Form der zugrunde liegenden Verteilung. Die Daten werden als zufällige Stichprobe betrachtet.

So lassen Sie einen Test auf Binomialverteilung berechnen:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > Binomial...

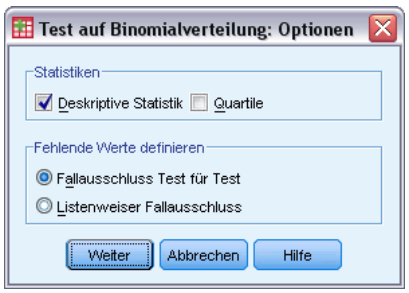
Abbildung 27-67
Dialogfeld "Test auf Binomialverteilung"

▶ Wählen Sie mindestens eine numerische Testvariable.

▶ Wenn Sie auf Optionen klicken, können Sie deskriptive Statistiken und Quartile abrufen sowie festlegen, wie fehlende Werte verarbeitet werden.

Optionen für den Test auf Binomialverteilung

Abbildung 27-68
Dialogfeld "Test auf Binomialverteilung: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle angezeigt.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.** Fälle mit fehlenden Werten für eine beliebige getestete Variable werden von allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (Test auf Binomialverteilung)

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl `BINOMIAL` können bestimmte Gruppen ausgewählt und andere Gruppen ausgeschlossen werden, wenn eine Variable über mehr als zwei Kategorien verfügt.
- Mit dem Unterbefehl `BINOMIAL` können verschiedene Trennwerte oder Wahrscheinlichkeiten für verschiedene Variablen angeben werden.
- Mit dem Unterbefehl `EXPECTED` kann dieselbe Variable bei verschiedenen Trennwerten oder Wahrscheinlichkeiten getestet werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Sequenzentest

Mit der Prozedur "Sequenzentest" können Sie testen, ob zwei Werte einer Variablen in zufälliger Reihenfolge auftreten. Eine Sequenz ist eine Folge von gleichen Beobachtungen. Eine Stichprobe mit zu vielen oder zu wenigen Sequenzen legt nahe, dass die Stichprobe nicht zufällig ist.

Beispiele. Es werden 20 Personen befragt, ob sie ein bestimmtes Produkt kaufen würden. Die angenommene zufällige Auswahl der Stichprobe wäre ernsthaft zu bezweifeln, wenn alle 20 Personen demselben Geschlecht angehören würden. Mit dem Sequenzentest kann bestimmt werden, ob die Stichprobe zufällig entnommen wurde.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile.

Daten. Die Variablen müssen numerisch sein. Verwenden Sie zum Umwandeln von String-Variablen in numerische Variablen den Befehl "Automatisch umkodieren" im Menü "Transformieren".

Annahmen. Nichtparametrische Tests erfordern keine Annahmen über die Form der zugrunde liegenden Verteilung. Verwenden Sie Stichproben aus stetigen Wahrscheinlichkeitsverteilungen.

So lassen Sie einen Sequenzentest berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > Sequenzen...

Abbildung 27-69
Hinzufügen eines benutzerdefinierten Trennwerts

- Wählen Sie mindestens eine numerische Testvariable.

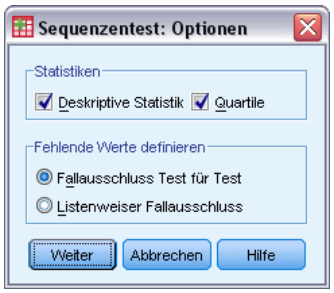
- Wenn Sie auf Optionen klicken, können Sie deskriptive Statistiken und Quartile abrufen sowie festlegen, wie fehlende Werte verarbeitet werden.

Sequenzentest: Trennwert

Trennwert. Hier wird ein Trennwert zum Dichotomisieren der gewählten Variablen angegeben. Sie können den beobachteten Mittelwert, den Median, den Modalwert oder einen angegebenen Wert als Trennwert wählen. Fälle mit Werten kleiner als der Trennwert werden einer Gruppe, Fälle mit Werten größer oder gleich dem Trennwert einer anderen Gruppe zugeordnet. Für jeden gewählten Trennwert wird ein Test ausgeführt.

Sequenzentest: Optionen

Abbildung 27-70
Dialogfeld "Sequenzentest: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle angezeigt.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (Sequenzentest)

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl `RUNS` können verschiedene Trennwerte für verschiedene Variablen angegeben werden.
- Mit dem Unterbefehl `RUNS` kann dieselbe Variable mit verschiedenen benutzerdefinierten Trennwerten getestet werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kolmogorov-Smirnov-Test bei einer Stichprobe

Mit dem Kolmogorov-Smirnov-Test bei einer Stichprobe (Anpassungstest) wird die beobachtete kumulative Verteilungsfunktion für eine Variable mit einer festgelegten theoretischen Verteilung verglichen, die eine Normalverteilung, eine Gleichverteilung, eine Poisson-Verteilung oder Exponentialverteilung sein kann. Das Kolmogorov-Smirnov-Z wird aus der größten Differenz (in Absolutwerten) zwischen beobachteten und theoretischen kumulativen Verteilungsfunktionen berechnet. Mit diesem Test für die Güte der Anpassung wird getestet, ob die Beobachtung wahrscheinlich aus der angegebenen Verteilung stammt.

Beispiel.Für viele parametrische Tests sind normalverteilte Variablen erforderlich. Mit dem Kolmogorov-Smirnov-Anpassungstest kann getestet werden, ob eine Variable, zum Beispiel *Einkommen*, normalverteilt ist.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile.

Daten. Die Variablen müssen auf Intervall- oder Verhältnis-Messniveau quantitativ sein.

Annahmen.Für den Kolmogorov-Smirnov-Test wird angenommen, dass die Parameter der zu testenden Verteilung im voraus angegeben wurden. Mit dieser Prozedur werden die Parameter aus der Stichprobe geschätzt. Der Mittelwert und die Standardabweichung der Stichprobe sind die Parameter für eine Normalverteilung. Minimum und Maximum der Stichprobe definieren die Spannweite der Gleichverteilung, und der Mittelwert der Stichprobe ist der Parameter für die Poisson-Verteilung sowie der Parameter für die Exponentialverteilung. Die Stärke des Tests, Abweichungen von der hypothetischen Verteilung zu erkennen, kann dabei deutlich verringert werden. Wenn Sie einen Test gegen eine Normalverteilung mit geschätzten Parametern durchführen möchten, sollten Sie den Kolmogorov-Smirnov-Test mit der Korrektur nach Lilliefors (in der Prozedur "Explorative Datenanalyse") in Betracht ziehen.

So berechnen Sie einen Kolmogorov-Smirnov-Anpassungstest:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > K-S bei einer Stichprobe...

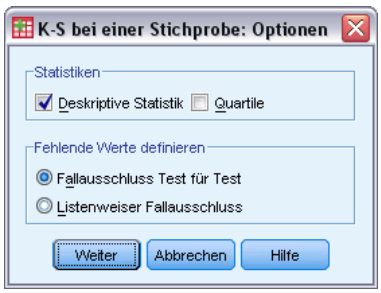
Abbildung 27-71
Dialogfeld "Kolmogorov-Smirnov-Test bei einer Stichprobe"

- Wählen Sie mindestens eine numerische Testvariable. Mit jeder Variablen wird ein separater Test erzeugt.

- Wenn Sie auf Optionen klicken, können Sie deskriptive Statistiken und Quartile abrufen sowie festlegen, wie fehlende Werte verarbeitet werden.

K-S bei einer Stichprobe: Optionen

Abbildung 27-72
Dialogfeld "K-S bei einer Stichprobe: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle angezeigt.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (Kolmogorov-Smirnov-Anpassungstest)

Mit der Befehlssyntax-Sprache können Sie auch die Parameter der zu testenden Verteilung angeben (mit dem Unterbefehl `K-S`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Tests bei zwei unabhängigen Stichproben

Die Prozedur "Test bei zwei unabhängigen Stichproben" vergleicht zwei Gruppen von Fällen von einer Variablen.

Beispiel. Es wurden neue Zahnspangen entwickelt, die bequemer sein sollen, besser aussehen und zu einem schnelleren Erfolg beim Richten der Zähne führen sollen. Um festzustellen, ob die neuen Spangen so lange wie die alten getragen werden müssen, wurden willkürlich 10 Kinder zum Tragen der alten Zahnspangen und weitere 10 Kinder zum Tragen der neuen Spangen ausgewählt. Anhand des Mann-Whitney-*U*-Tests stellen Sie eventuell fest, dass die neuen Spangen im Durchschnitt nicht so lange wie die alten Spangen getragen werden mussten.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile. Tests: Mann-Whitney-*U*-Test, Extremreaktionen nach Moses, Kolmogorov-Smirnov-*Z*-Test, Sequenztest nach Wald-Wolfowitz.

Daten. Verwenden Sie numerische Variablen, die geordnet werden können.

Annahmen. Verwenden Sie unabhängige Zufallsstichproben. Mit dem Mann-Whitney-*U*-Test wird die Gleichheit von zwei Verteilungen getestet. Um damit Unterschiede in der Lage von zwei Verteilungen zu testen, muss davon ausgegangen werden, dass die Verteilungen dieselbe Form haben.

So lassen Sie Tests bei zwei unabhängigen Stichproben berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > Zwei unabhängige Stichproben...

Abbildung 27-73
Dialogfeld "Tests bei zwei unabhängigen Stichproben"

- Wählen Sie mindestens eine numerische Variable aus.
- Wählen Sie eine Gruppenvariable aus und klicken Sie auf Gruppen definieren, um die Datei in zwei Gruppen oder Stichproben aufzuteilen.

Typen von Tests bei zwei unabhängigen Stichproben

Welche Tests durchführen?Mithilfe von vier Tests können Sie überprüfen, ob zwei unabhängige Stichproben (Gruppen) aus derselben Grundgesamtheit stammen.

Der **Mann-Whitney-U-Test**ist der am häufigsten verwendete Test bei zwei unabhängigen Stichproben. Er ist äquivalent zum Wilcoxon-Rangsummentest und dem Kruskal-Wallis-Test für zwei Gruppen. Mit dem Mann-Whitney-U-Test wird überprüft, ob zwei beprobte Grundgesamtheiten die gleiche Lage besitzen. Die Beobachtungen aus beiden Gruppen werden kombiniert und in eine gemeinsame Reihenfolge gebracht, wobei im Falle von Rangbindungen der durchschnittliche Rang vergeben wird. Die Anzahl der Bindungen sollte im Verhältnis zur Gesamtanzahl der Beobachtungen klein sein. Wenn die Grundgesamtheiten in der Lage identisch sind, sollten die Ränge zufällig zwischen den beiden Stichproben gemischt werden. Im Test wird berechnet, wie oft ein Wert aus Gruppe 1 einem Wert aus Gruppe 2 und wie oft ein Wert aus Gruppe 2 einem Wert aus Gruppe 1 vorangeht. Die Mann-Whithney-*U*-Statistik ist die kleinere dieser beiden Zahlen. Die Statistik der Wilcoxon-Rangsumme *W* wird ebenfalls angezeigt. *W* ist die Summe der Ränge für die Gruppe mit dem kleineren mittleren Rang. Wenn die Gruppen

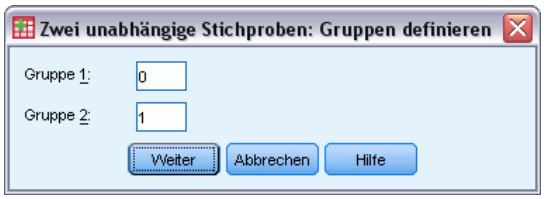
denselben mittleren Rang aufweisen, wird die Rangsumme der Gruppe verwendet, die im Dialogfeld "Zwei unabhängige Stichproben: Gruppen definieren" weiter unten genannt wird.

Der **Kolmogorov-Smirnov-Z-Test** und der **Sequenztest nach Wald-Wolfowitz** stellen eher allgemeine Tests dar, die sowohl Unterschiede in den Lagen als auch in den Formen der Verteilungen erkennen. Der Test nach Kolmogorov-Smirnov arbeitet auf der Grundlage der maximalen absoluten Differenz zwischen den beobachteten kumulativen Verteilungsfunktionen für beide Stichproben. Wenn diese Differenz signifikant groß ist, werden die beiden Verteilungen als verschieden betrachtet. Der Sequenztest nach Wald-Wolfowitz kombiniert die Beobachtungen aus beiden Gruppen und ordnet ihnen einen Rang zu. Wenn die beiden Stichproben aus derselben Grundgesamtheit stammen, müssen die beiden Gruppen in der Rangverteilung zufällig gestreut sein.

Der **Test "Extremreaktionen nach Moses"** setzt voraus, dass die experimentelle Variable einige Subjekte in der einen Richtung und andere Subjekte in der entgegengesetzten Richtung beeinflußt. In diesem Test wird auf extreme Antworten im Vergleich zu einer Kontrollgruppe geprüft. Dieser Test konzentriert sich auf die Spannweite der Kontrollgruppe und ist ein Maß dafür, wie stark die Spannweite durch die extremen Werte in der experimentellen Gruppe beeinflußt wird, wenn sie mit der Kontrollgruppe verbunden werden. Die Kontrollgruppe wird durch den Wert der Gruppe 1 im Dialogfeld "Zwei unabhängige Stichproben: Gruppen definieren" bestimmt. Die Beobachtungen aus beiden Gruppen werden kombiniert und einem Rang zugeordnet. Die Spanne der Kontrollgruppe wird als die Differenz zwischen den Rängen der größten und kleinsten Werte in der Kontrollgruppe plus 1 berechnet. Da zufällige Ausreißer den Bereich der Spannweite leicht verzerren können, werden 5 % der Kontrollfälle automatisch an jedem Ende weggelassen.

Zwei unabhängige Stichproben: Gruppen definieren

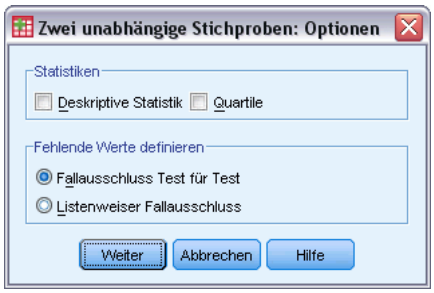
Abbildung 27-74
Dialogfeld "Zwei unabhängige Stichproben: Gruppen definieren"



Um die Datei in zwei Gruppen oder Stichproben aufzuteilen, geben Sie eine ganze Zahl für Gruppe 1 und eine weitere Zahl für Gruppe 2 ein. Fälle mit anderen Werten werden aus der Analyse ausgeschlossen.

Tests bei zwei unabhängigen Stichproben – Optionen

Abbildung 27-75
Dialogfeld "Zwei unabhängige Stichproben: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle an.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (Tests bei zwei unabhängigen Stichproben)

Mit dem Unterbefehl `MOSES` der Befehlssyntax-Sprache kann die Anzahl der Fälle angegeben werden, die für den Moses-Test getrimmt werden sollen.

Siehe *Befehlssyntaxreferenz* für die vollständigen Syntaxinformationen.

Tests bei zwei verbundenen Stichproben

Die Prozedur "Tests bei zwei verbundenen Stichproben" vergleicht die Verteilungen von zwei Variablen.

Beispiel. Erhalten Familien, die ihr Haus verkaufen, im allgemeinen den geforderten Preis? Wenn Sie den Wilcoxon-Test auf die Daten von 10 Häusern anwenden, könnten Sie beispielsweise feststellen, dass sieben Familien weniger als den geforderten Preis, eine Familie mehr als den geforderten Preis und zwei Familien den geforderten Preis erhielten.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile. Tests: Wilcoxon-Test, Vorzeichentest, McNemar. Wenn die Option "Exakte Tests" installiert ist (nur unter Windows-Betriebssystemen verfügbar) steht außerdem der Rand-Homogenitätstest zur Verfügung.

Daten. Verwenden Sie numerische Variablen, die geordnet werden können.

Annahmen. Obwohl keine bestimmten Verteilungen für die beiden Variablen vorausgesetzt werden, wird die Verteilung der Grundgesamtheit der gepaarten Differenzen als symmetrisch angenommen.

So lassen Sie Tests bei zwei verbundenen Stichproben berechnen:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > Zwei verbundene Stichproben...

Abbildung 27-76
Dialogfeld "Tests bei zwei verbundenen Stichproben"

▶ Wählen Sie mindestens ein Variablenpaar aus.

Typen von Tests bei zwei verbundenen Stichproben

Die Tests in diesem Abschnitt vergleichen die Verteilungen von zwei verbundenen Variablen. Der geeignete Test hängt vom jeweiligen Datentyp ab.

Falls Ihre Daten stetig sind, verwenden Sie den Vorzeichentest oder den Wilcoxon-Test. Der **Vorzeichentest** berechnet für alle Fälle die Differenzen zwischen den beiden Variablen und klassifiziert sie als positiv, negativ oder verbunden. Falls die beiden Variablen ähnlich verteilt sind, unterscheidet sich die Zahl der positiven und negativen Differenzen nicht signifikant. Der **Wilcoxon-Test** berücksichtigt sowohl Informationen über Vorzeichen der Differenzen als auch die Größe der Differenzen zwischen den Paaren. Da der Wilcoxon-Test mehr Informationen über die Daten aufnimmt, kann er mehr leisten als der Vorzeichentest.

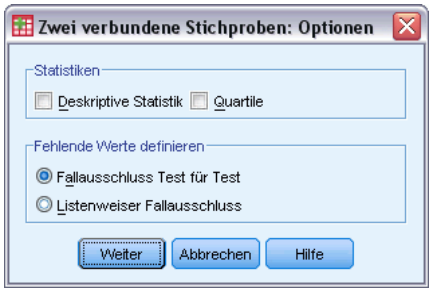
Falls Sie mit binären Daten arbeiten, verwenden Sie den **McNemar-Test**. Dieser Test wird üblicherweise bei Messwiederholungen verwendet, wenn jede Antwort eines Subjektes doppelt abgerufen wird, einmal bevor ein festgelegtes Ereignis eintritt und einmal danach. Der McNemar-Test bestimmt, ob die Antwortrate am Anfang (vor dem Ereignis) gleich der Antwortrate am Ende (nach dem Ereignis) ist. Dieser Test ist für das Erkennen von Änderungen bei Antworten nützlich, die durch experimentelle Einflußnahme in sogenannten "Vorher-und-nachher-Designs" entstanden sind.

Falls Sie mit kategorialen Daten arbeiten, verwenden Sie den **Rand-Homogenitätstest**. Dieser Test ist eine Erweiterung des McNemar-Tests von binären Variablen auf multinomiale Variablen. Mithilfe dieses Tests wird unter Verwendung der Chi-Quadrat-Verteilung überprüft, ob Änderungen bei den Antworten vorliegen. Dies ist nützlich, um zu ermitteln, ob die Änderungen in sogenannten "Vorher-und-nachher-Designs" durch experimentelle Einflußnahme verursacht

werden. Der Rand-Homogenitätstest ist nur verfügbar, wenn Sie die Option Exact Tests installiert haben.

Optionen für Tests bei zwei verbundenen Stichproben

Abbildung 27-77
Dialogfeld "Zwei verbundene Stichproben: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle an.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (zwei verbundene Stichproben)

Mit der Befehlssyntax-Sprache können Sie außerdem eine Variable mit jeder Variable auf einer Liste überprüfen.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Tests bei mehreren unabhängigen Stichproben

Mit der Prozedur "Tests bei mehreren unabhängigen Stichproben" werden zwei oder mehrere Fallgruppen einer Variablen verglichen.

Beispiel. Unterscheiden sich 100-Watt-Glühlampen dreier Marken in ihrer durchschnittlichen Lebensdauer? Mit der einfaktoriellen Varianzanalyse nach Kruskal-Wallis könnten Sie feststellen, dass die drei Marken sich in ihrer durchschnittlichen Lebensdauer unterscheiden.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile. Tests: Kruskal-Wallis-*H*, Median.

Daten. Verwenden Sie numerische Variablen, die geordnet werden können.

Annahmen. Verwenden Sie unabhängige Zufallsstichproben. Für den Kruskal-Wallis-*H*-Test sind Stichproben erforderlich, die sich in ihrer Form ähneln.

So lassen Sie Tests für mehrere unabhängige Stichproben berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > K unabhängige Stichproben...

Abbildung 27-78
Festlegung des Mediantests

- Wählen Sie mindestens eine numerische Variable aus.
- Wählen Sie eine Gruppenvariable aus und klicken Sie auf Bereich definieren, um die ganzzahligen Minimal- und Maximalwerte der Gruppenvariablen festzulegen.

Tests bei mehreren unabhängigen Stichproben: Welche Tests durchführen?

Sie können mit drei Tests bestimmen, ob mehrere unabhängige Stichproben aus derselben Grundgesamtheit stammen. Mit dem Kruskal-Wallis-*H*-Test, dem Mediantest und dem Jonckheere-Terpstra-Test können Sie prüfen, ob mehrere unabhängige Stichproben aus derselben Grundgesamtheit stammen.

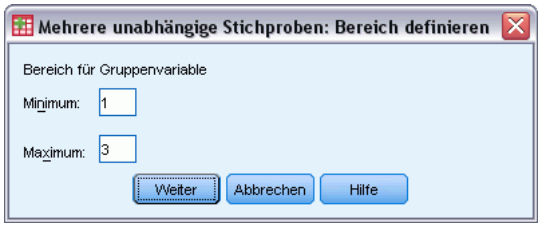
Der **Kruskal-Wallis-H-Test**, eine Erweiterung des Mann-Whitney-*U*-Tests, ist die nichtparametrische Entsprechung der einfaktoriellen Varianzanalyse und erkennt Unterschiede in der Lage der Verteilung. Der **Mediantest**, der allgemeiner, aber nicht so leistungsstark ist, erkennt Unterschiede von Verteilungen in Lage und Form. Der Kruskal-Wallis-*H*-Test und der Mediantest setzen voraus, dass keine *a-priori*-Ordnung der *k* Grundgesamtheiten vorliegt, aus denen die Stichproben gezogen werden.

Wenn eine natürliche *a-priori*-Ordnung (aufsteigend oder absteigend) der *k* Grundgesamtheiten *besteht*, ist der **Jonckheere-Terpstra-Test** leistungsfähiger. Die *k* Grundgesamtheiten könnten zum Beispiel *k* ansteigende Temperaturen darstellen. Die Hypothese, dass unterschiedliche Temperaturen die gleiche Verteilung von Antworten erzeugen, wird gegen die Alternative getestet, dass mit Zunahme der Temperatur die Größe der Antwort zunimmt. Hierbei ist die alternative

Hypothese geordnet, deshalb ist der Jonckheere-Terpstra-Test für diesen Test am besten geeignet. Der Jonckheere-Terpstra-Test ist nur verfügbar, wenn Sie das Erweiterungsmodul Exact Tests installiert haben.

Tests bei mehreren unabhängigen Stichproben: Bereich definieren

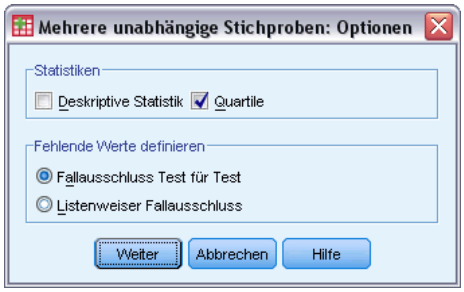
Abbildung 27-79
Dialogfeld "Mehrere unabhängige Stichproben: Bereich definieren"



Um den Bereich zu definieren, geben Sie für Minimum und Maximum ganzzahlige Werte ein, die der niedrigsten und höchsten Kategorie der Gruppenvariablen entsprechen. Der Minimalwert muss kleiner sein als der Maximalwert. Wenn Sie zum Beispiel als Minimum 1 und als Maximum 3 angeben, werden nur die ganzzahligen Werte von 1 bis 3 verwendet. Das Minimum muss kleiner als das Maximum sein. Beide Werte müssen angegeben werden.

Tests bei mehreren unabhängigen Stichproben: Optionen

Abbildung 27-80
Dialogfeld "Mehrere unabhängige Stichproben: Optionen"



Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistik.**Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle an.
- **Quartile.**Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Fehlende Werte. Bestimmt die Verarbeitung fehlender Werte.

- **Fallausschluss Test für Test.**Werden mehrere Tests festgelegt, so wird jeder Test einzeln auf fehlende Werte geprüft.
- **Listenweiser Fallausschluss.**Fälle mit fehlenden Werten für eine Variable werden aus allen Analysen ausgeschlossen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (K unabhängige Stichproben)

In der Befehlssyntax-Sprache haben Sie außerdem die Möglichkeit, mit dem Unterbefehl `MEDIAN` einen anderen Wert als den beobachteten Median für den Mediantest festzulegen.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Tests bei mehreren verbundenen Stichproben

Bei der Prozedur "Tests bei mehreren verbundenen Stichproben" werden die Verteilungen von zwei oder mehr Variablen verglichen.

Beispiel. Genießen die Berufsgruppen Ärzte, Anwälte, Polizisten oder Lehrer in der Öffentlichkeit ein unterschiedliches Ansehen? Zehn Personen wurden gebeten, diese vier Berufsgruppen in der Reihenfolge ihres Ansehens anzuordnen. Der Test nach Friedman zeigt, dass diese vier Berufsgruppen in der Öffentlichkeit tatsächlich ein unterschiedliches Ansehen genießen.

Statistiken. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nichtfehlenden Fälle und Quartile. Tests: Friedman, Kendall-*W* und Cochran-*Q*.

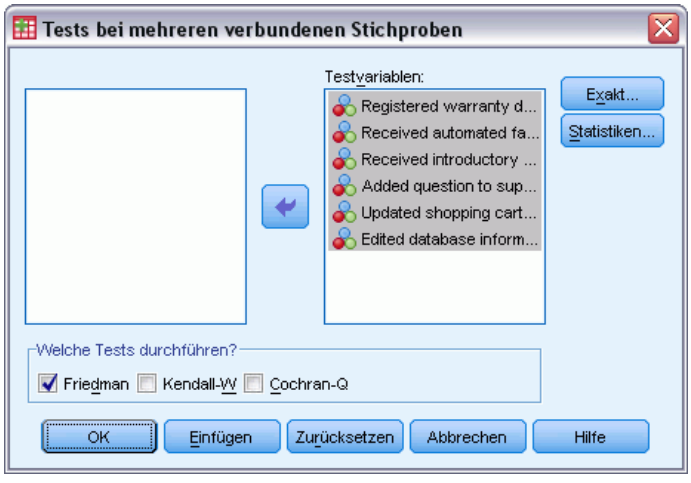
Daten. Verwenden Sie numerische Variablen, die geordnet werden können.

Annahmen. Nichtparametrische Tests erfordern keine Annahmen über die Form der zugrunde liegenden Verteilung. Verwenden Sie abhängige Zufallsstichproben.

So lassen Sie Tests bei mehreren verbundenen Stichproben berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Nichtparametrische Tests > Veraltete Dialogfelder > K verbundene Stichproben...

Abbildung 27-81
Auswahl von "Cochran" als Testtyp



- Wählen Sie zwei oder mehr numerische Testvariablen aus.

Tests bei mehreren verbundenen Stichproben: Welche Tests durchführen?

Sie können die Verteilung von verschiedenen verbundenen Variablen mit drei Tests vergleichen.

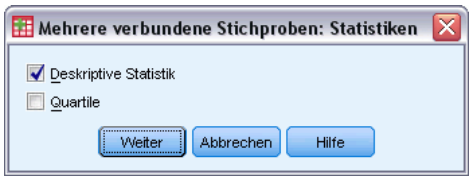
Der **Test** nach **Friedman** stellt das nichtparametrische Äquivalent eines Designs mit Messwiederholungen bei einer Stichprobe bzw. eine Zweifach-Varianzanalyse mit einer Beobachtung pro Zelle dar. Der Friedman-Test überprüft die Nullhypothese, wonach die *k* verbundenen Variablen aus derselben Grundgesamtheit stammen. Für jeden Fall werden den *k* Variablen Rangzahlen von 1 bis *k* zugewiesen. Die Teststatistik wird auf der Grundlage dieser Ränge durchgeführt.

Das **Kendall-W** stellt eine Normalisierung der Statistik nach Friedman dar. Das Kendall-*W* kann als Konkordanzkoeffizient interpretiert werden, der ein Maß für die Übereinstimmung der Prüfer darstellt. Jeder Fall ist ein Richter oder Prüfer, und jede Variable ist ein zu beurteilendes Objekt oder eine zu beurteilende Person. Die Rangsumme jeder Variablen wird berechnet. Das Kendall-*W* liegt im Bereich von 0 (keine Übereinstimmung) bis 1 (vollständige Übereinstimmung).

Das **Cochran-Q** entspricht vollständig dem Friedman-Test. Es wird jedoch angewendet, wenn alle Antworten binär sind. Dieser Test stellt eine Erweiterung des McNemar-Tests auf *k* Stichproben dar. Das Cochran-*Q* überprüft die Hypothese, dass mehrere verbundene dichotome Variablen denselben Mittelwert aufweisen. Die Variablenwerte beziehen sich auf dasselbe Individuum oder auf zusammengehörige Individuen.

Tests bei mehreren verbundenen Stichproben: Statistiken

Abbildung 27-82
Dialogfeld "Mehrere verbundene Stichproben: Statistiken"



Sie können Statistiken auswählen.

- **Deskriptive Statistik.** Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nichtfehlenden Fälle an.
- **Quartile.** Zeigt die Werte an, die den 25., 50. und 75. Perzentilen entsprechen.

Zusätzliche Funktionen beim Befehl NPAR TESTS (K verbundene Stichproben)

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
28

Analyse von Mehrfachantworten

Sie können für die Analyse von Sets aus dichotomen Variablen und von Sets aus kategorialen Variablen zwei Prozeduren verwenden. Mit der Prozedur "Mehrfachantworten: Häufigkeiten" können Sie Häufigkeitstabellen erstellen. Mit der Prozedur "Mehrfachantworten: Kreuztabellen" werden zwei- oder dreidimensionale Kreuztabellen angezeigt. Sie müssen Mehrfachantworten-Sets definieren, ehe Sie mit einer der Prozeduren beginnen.

Beispiel. Dieses Beispiel veranschaulicht den Gebrauch von Mehrfachantworten in einer Marktforschungsanalyse. Die hier verwendeten Daten sind frei erfunden und dürfen nicht als real interpretiert werden. Eine Fluggesellschaft führt eine Umfrage unter den Passagieren einer bestimmten Flugroute durch, um Informationen über konkurrierende Fluggesellschaften zu erhalten. In diesem Beispiel möchte American Airlines in Erfahrung bringen, welche anderen Fluggesellschaften ihre Passagiere auf der Route Chicago-New York nutzen und welche Rolle der Flugplan sowie der Service bei der Auswahl der Fluggesellschaft spielen. Der Flugbegleiter händigt jedem Passagier beim Einsteigen in die Maschine einen kurzen Fragebogen aus. Die erste Frage lautet: "Kreuzen Sie bitte alle der folgenden Fluggesellschaften an, mit denen Sie diese Route in den letzten sechs Monaten geflogen sind: American, United, TWA, USAir und andere." Dies ist eine Frage, die mit Mehrfachantworten beantwortet werden kann, weil jeder Passagier mehr als eine Antwort ankreuzen kann. Sie können diese Frage aber nicht direkt kodieren, weil eine IBM SPSS Statistics-Variable für jeden Fall nur einen Wert annehmen kann. Sie müssen mehrere Variablen verwenden, um die Antworten zu jeder Frage zu erfassen. Dazu haben Sie zwei Möglichkeiten. Eine Möglichkeit besteht darin, zu jeder Antwortmöglichkeit eine entsprechende Variable zu definieren, also zum Beispiel "American", "United", "TWA", "USAir" und "andere". Wenn ein Passagier "United" ankreuzt, wird der Variablen *united* der Code 1 zugewiesen, sonst erhält diese den Code 0. Bei dieser Methode werden Variablen in **mehreren Dichotomien** erfaßt. Eine andere Möglichkeit stellt das Erfassen der Antworten in **mehreren Kategorien** dar, bei der Sie die maximale Anzahl möglicher Antworten auf die Frage schätzen und eine entsprechende Anzahl von Variablen festlegen. Hierbei wird die verwendete Fluggesellschaft mit Hilfe eines Codes angegeben. Beim Durchsehen einer Stichprobe von Fragebögen stellen Sie vielleicht fest, daß in den letzten sechs Monaten kein Passagier mit mehr als drei verschiedenen Fluggesellschaften auf dieser Route geflogen ist. Außerdem bemerken Sie, daß aufgrund der Liberalisierung des Luftverkehrs 10 weitere Fluggesellschaften in der Kategorie "Andere" genannt sind. Mit der Methode für mehrere Kategorien würden Sie drei Variablen definieren. Jede würde wie folgt kodiert sein: 1 = *american*, 2 = *united*, 3 = *twa*, 4 = *usair*, 5 = *delta* usw. Wenn ein Passagier "American" und "TWA" ankreuzt, wird der ersten Variablen der Code 1 zugewiesen, der zweiten der Code 3 und der dritten ein Code für fehlende Werte. Ein anderer Passagier hat vielleicht "American" und "Delta" angekreuzt. Dementsprechend wird der ersten Variablen der Code 1, der zweiten der Code 5 und der dritten ein Code für fehlende Werte zugewiesen. Dagegen führt die Methode für mehrfache Dichotomie zu 14 verschiedenen Variablen. Obwohl

beide Methoden für dieses Umfragebeispiel geeignet sind, hängt die Wahl der Methode von der Verteilung der Antworten ab.

Mehrfachantworten: Sets definieren

Mit der Prozedur "Mehrfachantworten: Sets definieren" können Sie elementare Variablen in Sets aus dichotomen Variablen und Sets aus kategorialen Variablen gruppieren. Für diese Sets können Sie Häufigkeitstabellen und Kreuztabellen erstellen. Sie können bis zu 20 Mehrfachantworten-Sets definieren. Jedes Set muß über einen eigenen eindeutigen Namen verfügen. Sie können ein Set entfernen, indem Sie es in der Liste der Mehrfachantworten-Sets markieren und anschließend auf Entfernen klicken. Sie können ein Set ändern, indem Sie es in der Liste markieren, die Charakteristiken der Set-Definition ändern und anschließend auf Ändern klicken.

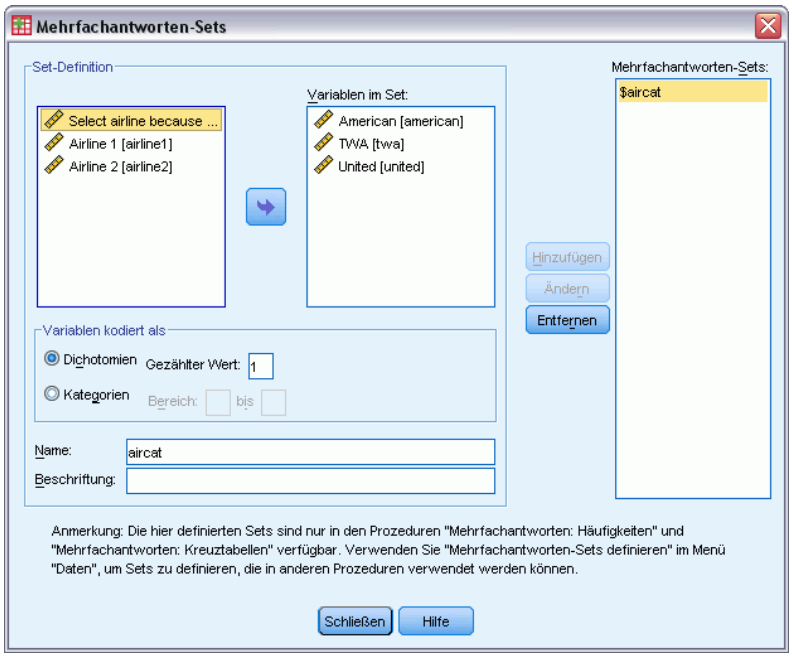
Sie können die elementaren Variablen als Dichotomien oder als Kategorien definieren. Wenn Sie dichotome Variablen verwenden möchten, aktivieren Sie das Optionsfeld Dichotomien, um ein Set aus dichotomen Variablen zu erstellen. Geben Sie für "Gezählter Wert" eine ganze Zahl ein. Jede Variable, bei welcher der gezählte Wert mindestens einmal auftritt, wird zu einer Kategorie des Sets aus dichotomen Variablen. Aktivieren Sie das Optionsfeld Kategorien, um ein Set aus kategorialen Variablen zu erstellen, das den gleichen Wertebereich wie die Komponentenvariablen umfaßt. Geben Sie ganzzahlige Werte für die Minimal- und Maximalwerte des Bereichs für die Kategorien des Sets aus kategorialen Variablen ein. IBM SPSS Statistics bildet die Summe aller unterschiedlichen ganzzahligen Werte im Bereich aller Komponentenvariablen. Leere Kategorien werden nicht in Tabellen übernommen.

Sie müssen jedem Mehrfachantworten-Set einen eindeutigen Namen zuweisen, der aus bis zu sieben Zeichen bestehen darf. IBM SPSS Statistics stellt dem von Ihnen zugewiesenen Namen das Dollarzeichen ($) als Präfix voran. Die folgenden reservierten Namen dürfen Sie nicht verwenden: *casenum*, *sysmis*, *jdate*, *date*, *time*, *length* und *width*. Der Name des Mehrfachantworten-Sets ist nur zur Verwendung in Mehrfachantworten-Prozeduren vorgesehen. In anderen Prozeduren können Sie sich nicht auf Namen von Mehrfachantworten-Sets beziehen. Wahlweise können Sie für das Mehrfachantworten-Set ein aussagekräftiges Variablenlabel eingeben. Das Label kann bis zu 40 Zeichen lang sein.

So definieren Sie Mehrfachantworten-Sets

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Mehrfachantworten > Variablen-Sets definieren...

Abbildung 28-1
Dialogfeld "Mehrfachantworten-Sets"



- Wählen Sie mindestens zwei Variablen aus.
- Wenn Ihre Variablen als Dichotomien kodiert sind, geben Sie an, welcher Wert gezählt werden soll. Wenn Ihre Variablen als Kategorien kodiert sind, legen Sie den Bereich für die Kategorien fest.
- Geben Sie einen eindeutigen Namen für jedes Mehrfachantworten-Set ein.
- Klicken Sie auf Hinzufügen, um das Mehrfachantworten-Set zur Liste der definierten Sets hinzuzufügen.

Mehrfachantworten: Häufigkeiten

Mit der Prozedur "Mehrfachantworten: Häufigkeiten" erstellen Sie Häufigkeitstabellen für Mehrfachantworten-Sets. Zuvor müssen Sie mindestens ein Mehrfachantworten-Set definieren (siehe "Mehrfachantworten: Sets definieren").

Bei Sets aus dichotomen Variablen entsprechen die in der Ausgabe gezeigten Kategorienamen den Variablenlabels, die für die elementaren Variablen in der Gruppe festgelegt wurden. Wenn keine Variablenlabels festgelegt wurden, werden die Variablennamen als Labels verwendet. Bei Sets aus kategorialen Variablen entsprechen die Kategoriebeschriftungen den Wertelabels der ersten Variable in der Gruppe. Wenn Kategorien, die bei der ersten Variable fehlen, bei anderen Variablen in der Gruppe vorhanden sind, müssen Sie ein Wertelabel für die fehlenden Kategorien festlegen.

Fehlende Werte. Fälle mit fehlenden Werten werden jeweils für einzelne Tabellen ausgeschlossen. Sie können aber auch eine oder beide der folgenden Möglichkeiten auswählen:

- **Für dichotome Variablen Fälle listenweise ausschließen.** Fälle, bei denen Werte einer beliebigen Variablen fehlen, werden aus der Tabelle des Sets aus dichotomen Variablen ausgeschlossen. Dies gilt nur für Mehrfachantworten-Sets, die als Sets aus dichotomen Variablen definiert wurden. In der Standardeinstellung gilt ein Fall in einem Set aus dichotomen Variablen als fehlend, wenn keine der Variablen des Falls den gezählten Wert enthält. Fälle mit fehlenden Werten für nur einige, aber nicht alle der Variablen werden in die Tabellen der Gruppe aufgenommen, wenn mindestens eine Variable den gezählten Wert enthält.
- **Für kategoriale Variablen Fälle listenweise ausschließen.** Fälle, bei denen Werte einer beliebigen Variablen fehlen, werden aus der Tabelle des Sets aus kategorialen Variablen ausgeschlossen. Dies gilt nur für Mehrfachantworten-Sets, die als Sets aus kategorialen Variablen definiert wurden. In der Standardeinstellung gilt ein Fall in einem Set aus kategorialen Variablen nur als fehlend, wenn keine der Komponenten des Falls gültige Werte innerhalb des definierten Bereichs enthält.

Beispiel. Jede Variable, die aus einer Frage in einer Umfrage erstellt wurde, stellt eine elementare Variable dar. Zum Analysieren der Mehrfachantworten müssen Sie die Variablen in einem der beiden möglichen Typen von Mehrfachantworten-Sets zusammenfassen: in einem Set aus dichotomen Variablen oder in einem Set aus kategorialen Variablen. Wenn zum Beispiel in einer Umfrage ermittelt wurde, mit welcher von drei verschiedenen Fluggesellschaften (American, United und TWA) die befragten Personen in den letzten sechs Monaten geflogen sind, und Sie haben dichotome Variablen verwendet und ein **Set aus dichotomen Variablen** definiert, dann würde jede der drei Variablen im Set zu einer Kategorie der Gruppenvariablen werden. Die Angaben zu Anzahl und Prozentwert für jede Fluggesellschaft werden zusammen in einer Häufigkeitstabelle angezeigt. Wenn Sie feststellen, dass keiner der Befragten mit mehr als zwei Fluggesellschaften geantwortet hat, können Sie zwei Variablen erstellen, die jeweils einen von drei Codes annehmen können. Dabei stellt jeder Code eine Fluggesellschaft dar. Wenn Sie ein **Set aus kategorialen Variablen** definieren, stellen die Werte in der Tabelle die Anzahl von gleichen Codes in den elementaren Variablen dar. Das resultierende Set von Werten entspricht denen für jede einzelne der elementaren Variablen. So entsprechen beispielsweise 30 Antworten mit "United" der Summe von fünf Antworten mit "United" für "Fluglinie 1" und 25 Antworten mit "United" für "Fluglinie 2". Die Angaben zu Anzahl und Prozentwert für jede Fluggesellschaft werden zusammen in einer Häufigkeitstabelle angezeigt.

Statistiken. Häufigkeitstabellen mit den Häufigkeiten, Prozentsätzen der Antworten, Prozentsätzen der Fälle, der Anzahl gültiger Fälle und der Anzahl fehlender Fälle.

Daten. Verwenden Sie Mehrfachantworten-Sets.

Annahmen. Die Häufigkeiten und Prozentsätze geben nützliche Beschreibungen für Daten mit beliebigen Verteilungen.

Verwandte Prozeduren. Mit der Prozedur "Mehrfachantworten: Sets definieren" können Sie Mehrfachantworten-Sets definieren.

So berechnen Sie Häufigkeiten mit Mehrfachantworten:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Mehrfachantworten > Häufigkeiten...

Abbildung 28-2
Dialogfeld "Mehrfachantworten: Häufigkeiten"

- Wählen Sie mindestens ein Mehrfachantworten-Set aus.

Mehrfachantworten: Kreuztabellen

Mit der Prozedur "Mehrfachantworten: Kreuztabellen" können Kreuztabellen für definierte Mehrfachantworten-Sets, elementare Variablen oder eine Kombination dieser Elemente berechnet werden. Sie können außerdem Prozentwerte für Zellen basierend auf Fällen oder Antworten berechnen lassen, die Verarbeitung von fehlenden Werten ändern oder gepaarte Kreuztabellen erstellen lassen. Zuvor müssen Sie mindestens ein Mehrfachantworten-Set definieren (siehe "So definieren Sie Mehrfachantworten-Sets").

Bei Sets aus dichotomen Variablen entsprechen die in der Ausgabe gezeigten Kategorienamen den Variablenlabels, die für die elementaren Variablen in der Gruppe festgelegt wurden. Wenn keine Variablenlabels festgelegt wurden, werden die Variablennamen als Labels verwendet. Bei Sets aus kategorialen Variablen entsprechen die Kategoriebeschriftungen den Wertelabels der ersten Variable in der Gruppe. Wenn Kategorien, die bei der ersten Variable fehlen, bei anderen Variablen in der Gruppe vorhanden sind, müssen Sie ein Wertelabel für die fehlenden Kategorien festlegen. Die Prozedur zeigt die Kategorienbeschriftungen für Spalten auf drei Zeilen mit bis zu acht Zeichen pro Zeile an. Wenn Sie vermeiden möchten, dass Wörter getrennt werden, können Sie die Anordnung von Zeilen und Spalten umdrehen oder die Labels neu festlegen.

Beispiel. Sowohl Sets aus dichotomen Variablen als auch Sets aus kategorialen Variablen können bei dieser Prozedur mit anderen Variablen in eine Kreuztabelle eingehen. Bei einer Befragung von Passagieren einer Fluglinie werden die Reisenden um die folgenden Informationen gebeten: Kreuzen Sie bitte alle der folgenden Fluggesellschaften an, mit denen Sie in den letzten sechs Monaten geflogen sind (American, United und TWA). Was ist wichtiger, wenn Sie einen Flug buchen: der Flugplan oder der Service? Wählen Sie nur eine Möglichkeit aus. Nachdem Sie die Daten als Dichotomien oder multiple Kategorien eingegeben und diese in einem Set zusammengefaßt haben, können Sie die Auswahl der Fluggesellschaften zusammen mit der Frage nach Service bzw. Flugplan als Kreuztabelle berechnen lassen.

Statistiken. Kreuztabellen mit Häufigkeiten pro Zelle, Zeile, Spalte und Gesamt sowie Prozentsätzen für Zellen, Zeilen, Spalten und Gesamt. Die Prozentwerte für die Zellen können auf Fällen oder auf Antworten basieren.

Daten. Verwenden Sie Mehrfachantworten-Sets oder numerische kategoriale Variablen.

Annahmen. Die Häufigkeiten und Prozentsätze geben nützliche Beschreibungen für Daten mit beliebigen Verteilungen.

Verwandte Prozeduren. Mit der Prozedur "Mehrfachantworten: Sets definieren" können Sie Mehrfachantworten-Sets definieren.

So berechnen Sie Kreuztabellen mit Mehrfachantworten:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Analysieren > Mehrfachantworten > Kreuztabellen...

Abbildung 28-3
Dialogfeld "Mehrfachantworten: Kreuztabellen"

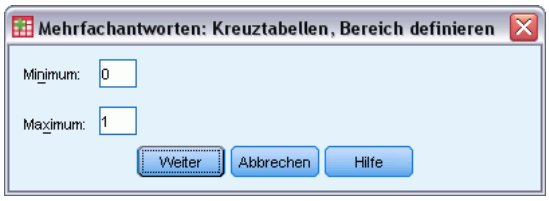
▶ Wählen Sie mindestens eine numerische Variable oder mindestens ein Mehrfachantworten-Set für jede Dimension der Kreuztabelle aus.

▶ Definieren Sie den Bereich jeder elementaren Variablen.

Außerdem können Sie eine Zweifach-Kreuztabelle für jede Kategorie einer Kontroll-Variablen oder eines Mehrfachantworten-Sets berechnen lassen. Wählen Sie mindestens einen Eintrag für die Liste "Schicht(en)" aus.

Mehrfachantworten: Kreuztabellen, Bereich definieren

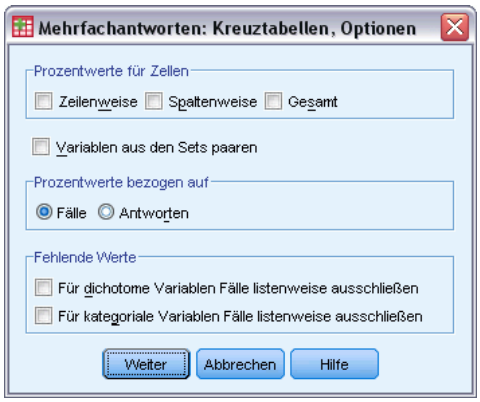
Abbildung 28-4
Dialogfeld "Mehrfachantworten: Kreuztabellen, Bereich definieren"



Für jede elementare Variable in der Kreuztabelle muss ein gültiger Wertebereich festgelegt werden. Geben Sie für die niedrigsten und höchsten Kategoriewerte, die in die Berechnung eingehen sollen, ganze Zahlen ein. Kategorien außerhalb des gültigen Bereichs werden aus der Analyse ausgeschlossen. Bei Werten innerhalb des einschließenden Bereichs wird von ganzen Zahlen ausgegangen, Stellen nach dem Komma werden abgeschnitten.

Mehrfachantworten: Kreuztabellen, Optionen

Abbildung 28-5
Dialogfeld "Mehrfachantworten: Kreuztabellen, Optionen"



Prozentwerte für Zellen. Die Zellenhäufigkeiten werden immer angezeigt. Sie können aber auch Spalten- und Zeilenprozentwerte sowie Prozentwerte für Zweifach-Tabellen (Gesamtwerte) anzeigen lassen.

Prozentwerte bezogen auf. Sie können festlegen, dass die Prozentsätze für die Zellen auf Fällen basieren. Diese Option ist nicht verfügbar, wenn Sie Variablen aus verschiedenen Sets von kategorialen Variablen paaren. Die Prozentsätze für die Zellen können außerdem auf den Antworten basieren. Bei Sets aus dichotomen Variablen entspricht die Anzahl der Antworten der Anzahl von gezählten Werten in allen Fällen. Bei Sets aus kategorialen Variablen entspricht die Anzahl der Antworten der Anzahl von Werten im festgelegten Bereich.

Fehlende Werte. Sie können eine oder beide der folgenden Möglichkeiten auswählen:

- **Für dichotome Variablen Fälle listenweise ausschließen.** Fälle, bei denen Werte einer beliebigen Variablen fehlen, werden aus der Tabelle des Sets aus dichotomen Variablen ausgeschlossen. Dies gilt nur für Mehrfachantworten-Sets, die als Sets aus dichotomen Variablen definiert wurden. In der Standardeinstellung gilt ein Fall in einem Set aus dichotomen Variablen als

fehlend, wenn keine der Variablen des Falls den gezählten Wert enthält. Fälle mit fehlenden Werten für nur einige, aber nicht alle der Variablen werden in die Tabellen der Gruppe aufgenommen, wenn mindestens eine Variable den gezählten Wert enthält.

- **Für kategoriale Variablen Fälle listenweise ausschließen.** Fälle, bei denen Werte einer beliebigen Variablen fehlen, werden aus der Tabelle des Sets aus kategorialen Variablen ausgeschlossen. Dies gilt nur für Mehrfachantworten-Sets, die als Sets aus kategorialen Variablen definiert wurden. In der Standardeinstellung gilt ein Fall in einem Set aus kategorialen Variablen nur als fehlend, wenn keine der Komponenten des Falls gültige Werte innerhalb des definierten Bereichs enthält.

Die Standardeinstellung sieht vor, dass beim Erstellen von Kreuztabellen für Sets aus kategorialen Variablen jede Variable in der ersten Gruppe mit jeder Variablen in der zweiten Gruppe gepaart wird und die Häufigkeiten für jede Zelle addiert werden. Deshalb können manche Antworten mehr als einmal in einer Tabelle vorkommen. Sie können die folgende Option auswählen:

Variablen aus den Sets paaren. Hiermit wird die erste Variable aus der ersten Gruppe mit der ersten Variable aus der zweiten Gruppe gepaart usw. Wenn Sie diese Option auswählen, basieren die relativen Häufigkeiten in den Zellen nicht auf den Fällen, sondern auf den Antworten. Bei Sets aus dichotomen Variablen und elementaren Variablen steht das Paaren nicht zur Verfügung.

Zusätzliche Funktionen beim Befehl MULT RESPONSE

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl `BY` können Kreuztabellen mit bis zu fünf Dimensionen berechnet werden.
- Mit dem Unterbefehl `FORMAT` können die Optionen für die Ausgabeformatierung geändert werden. So können beispielsweise Wertelabels unterdrückt werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
29

Ergebnisberichte

Auflistungen von Fällen und deskriptive Statistiken sind wichtige Hilfsmittel zur Untersuchung und Darstellung von Daten. Mit dem Daten-Editor oder der Prozedur "Berichte" können Sie Fälle auflisten, mit den Prozeduren "Häufigkeiten" Häufigkeitszählungen und deskriptive Statistiken erstellen und mit der Prozedur "Mittelwert" Statistiken für Teilgrundgesamtheiten anfordern. In jeder dieser Prozeduren wird ein zur übersichtlichen Darstellung von Informationen geeignetes Format verwendet. Mit den Funktionen "Bericht in Zeilen" und "Bericht in Spalten" können Sie für Informationen auch ein anderes Format der Datendarstellung wählen.

Bericht in Zeilen

Mit der Funktion "Bericht in Zeilen" werden Berichte erstellt, in denen verschiedene Auswertungsstatistiken in Zeilen angegeben sind. Ebenso sind Listen von Fällen mit oder ohne Auswertungsstatistik verfügbar.

Beispiel. In einem Einzelhandelsunternehmen mit Filialen werden Informationen über Angestellte, Gehälter, Anstellungszeiten sowie Filiale und Abteilung jedes Beschäftigten in Datensätzen gespeichert. Sie können einen Bericht erstellen, der nach Filiale und Abteilung (Break-Variablen) aufgeteilte Informationen (Listen) zu den einzelnen Beschäftigten liefert und eine Auswertungsstatistik (zum Beispiel Durchschnittsgehalt) für jede Filiale, jedes Ressort und jede Abteilung einer Filiale enthält.

Datenspalten. Hier werden die Berichtsvariablen aufgelistet, für die Sie Fälle auflisten oder Auswertungsstatistiken erstellen möchten, und das Anzeigeformat der Datenspalten festgelegt.

Break-Spalten. Hier werden optionale Break-Variablen aufgelistet, die den Bericht in Gruppen aufteilen, und Einstellungen für die Auswertungsstatistik sowie Anzeigeformate für Break-Spalten festgelegt. Bei mehreren Break-Variablen wird für jede Kategorie einer Break-Variablen eine getrennte Gruppe innerhalb der Kategorien der vorhergehenden Break-Variablen in der Liste erzeugt. Die Break-Variablen müssen diskrete kategoriale Variablen sein, welche die Fälle in eine begrenzte Anzahl von sinnvollen Kategorien aufteilen. Die Einzelwerte jeder Break-Variablen werden in einer getrennten Spalte links von allen Datenspalten angezeigt.

Bericht. Hiermit werden alle Merkmale eines Berichts festgelegt, einschließlich zusammenfassender Gesamtstatistiken, Anzeige der fehlenden Werte, Seitennumerierung und Titel.

Fälle anzeigen. Hiermit werden für jeden Fall die aktuellen Werte (oder Wertelabels) von den Variablen der Datenspalten angezeigt. Dadurch wird ein Listenbericht erzeugt, der wesentlich umfangreicher als ein Zusammenfassungsbericht sein kann.

Vorschau. Es wird nur die erste Seite des Berichtes angezeigt. Mit dieser Option erhalten Sie eine Vorschau auf das Format Ihres Berichts, ohne diesen komplett bearbeiten zu müssen.

Daten sind schon sortiert. Bei Berichten mit Break-Variablen muss die Datendatei vor dem Erstellen des Berichts nach den Werten der Break-Variablen sortiert werden. Wenn Ihre Datendatei bereits nach den Werten der Break-Variablen sortiert ist, können Sie durch Auswählen dieser Option Rechenzeit sparen. Diese Option ist besonders hilfreich, wenn Sie bereits einen Bericht für die Vorschau erstellt haben.

So erstellen Sie eine Zusammenfassung: Bericht in Zeilen

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Berichte > Bericht in Zeilen...

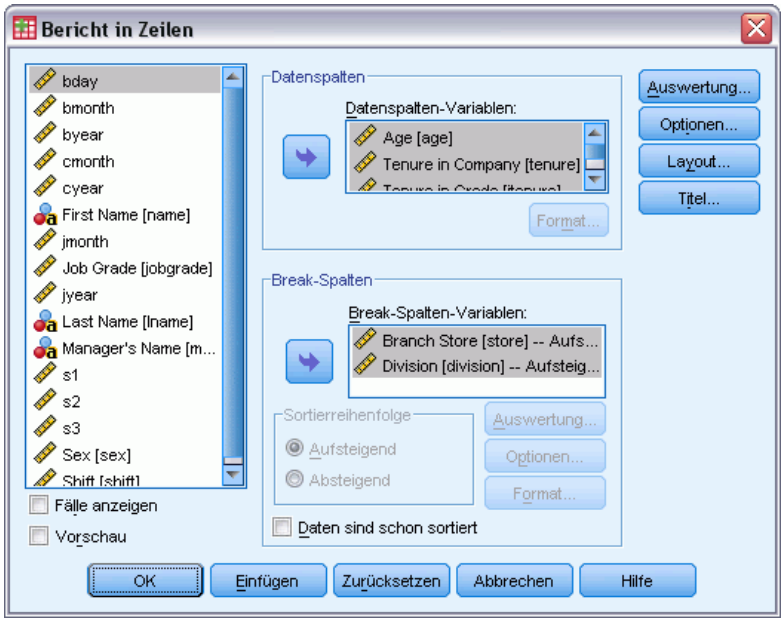
- Wählen Sie mindestens eine Variable für die Datenspalten aus. Für jede ausgewählte Variable wird eine Spalte im Bericht erzeugt.

- Wählen Sie bei sortierten und nach Untergruppen angezeigten Berichten mindestens eine Variable für die Break-Spalten aus.

- Bei Berichten mit Auswertungsstatistiken für Untergruppen, die durch Break-Variablen definiert wurden, wählen Sie in der Liste "Break-Spalten-Variablen" die Break-Variablen aus und klicken Sie im Gruppenfeld "Break-Spalten" auf Auswertung, um das (die) Auswertungsmaß(e) festzulegen.

- Bei Berichten mit zusammenfassenden Auswertungsstatistiken klicken Sie auf Auswertung, um das (die) Auswertungsmaß(e) festzulegen.

Abbildung 29-1
Dialogfeld "Bericht in Zeilen"



Datenspaltenformat/Break-Format in Berichten

In den Format-Dialogfeldern werden Spaltentitel, Spaltenbreite, Textausrichtung sowie Anzeige der Datenwerte oder Wertelabels festgelegt. Mit “Datenspaltenformat” wird das Format der Datenspalten auf der rechten Seite des Berichtes festgelegt. Das Format der Break-Spalten auf der linken Seite wird mit “Break-Format” festgelegt.

Abbildung 29-2
Dialogfeld “Bericht: Datenspaltenformat”

Spaltentitel. Hiermit legen Sie den Spaltentitel für die ausgewählte Variable fest. Lange Titel werden in der Spalte automatisch umgebrochen. Verwenden Sie die Eingabetaste, um Zeilenumbrüche für Titel manuell einzufügen.

Position des Werts in der Spalte. Hiermit wird für die ausgewählte Variable die Ausrichtung des Datenwerts oder Wertelabels in der Spalte festgelegt. Die Ausrichtung der Werte oder Labels hat keinen Einfluß auf die Ausrichtung der Spaltenüberschriften. Der Spalteninhalt kann entweder um eine festgelegte Anzahl von Zeichen eingerückt oder zentriert werden.

Spalteninhalt. Steuert die Anzeige von Datenwerten oder definierten Wertelabels der ausgewählten Variablen. Für Werte ohne definierte Wertelabels werden immer Datenwerte angezeigt. (Nicht verfügbar für Datenspalten in Bericht in Spalten.)

Bericht: Auswertungszeilen für/Endgültige Auswertungszeilen

Die beiden Dialogfelder für Auswertungszeilen legen Einstellungen für die Anzeige der Auswertungsstatistik für Break-Gruppen und für den gesamten Bericht fest. Mit “Auswertung” können Sie Einstellungen bezüglich der Untergruppenstatistik für jede durch die Break-Variablen definierte Kategorie vornehmen. Mit “Endgültige Auswertungszeilen” können Sie Einstellungen für die am Ende des Berichts angezeigte Gesamtstatistik vornehmen.

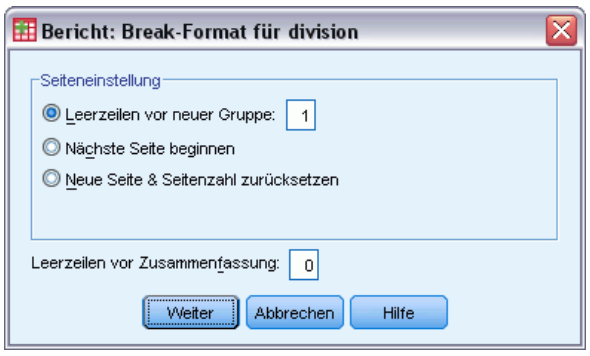
Abbildung 29-3
Dialogfeld "Bericht: Auswertung"

Die verfügbaren Auswertungsstatistiken sind Summe, Mittelwert, Minimum, Maximum, Anzahl der Fälle, Prozent der Fälle über oder unter einem festgelegten Wert, Prozent der Fälle innerhalb eines festgelegten Wertebereichs, Standardabweichung, Kurtosis, Varianz und Schiefe.

Bericht: Break-Optionen

Mit "Break-Optionen" werden Abstand und Seitenaufteilung der Informationen in den Break-Kategorien festgelegt.

Abbildung 29-4
Dialogfeld "Bericht: Break-Optionen"



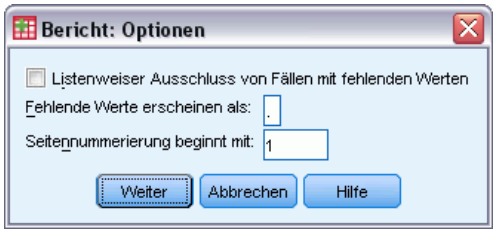
Seiteneinstellung. Hiermit werden Abstand und Seitenaufteilung für Kategorien der ausgewählten Break-Variablen festgelegt. Sie können eine Anzahl von Leerzeilen zwischen den Break-Kategorien festlegen oder eine Break-Kategorie an einen neuen Seitenanfang legen.

Leerzeilen vor Zusammenfassung. Hiermit legen Sie die Anzahl der Leerzeilen zwischen Beschriftungen oder Daten von Break-Kategorien und Auswertungsstatistiken fest. Dies bietet sich besonders für kombinierte Berichte mit Listen von einzelnen Fällen und Auswertungsstatistiken für Break-Kategorien an. In diesen Berichten können Sie Leerraum zwischen Listen von Fällen und Auswertungsstatistiken einfügen.

Bericht: Optionen

Mit "Bericht: Optionen" werden Behandlung und Anzeige der fehlenden Werte sowie Seitenaufteilung des Berichts festgelegt.

Abbildung 29-5
Dialogfeld "Bericht: Optionen"



Fälle mit fehlenden Werten listenweise ausschließen.Für jede der Berichtsvariablen werden sämtliche Fälle mit fehlenden Werten (im Bericht) ausgeschlossen.

Fehlende Werte erscheinen als. Hier legen Sie das Symbol für fehlende Werte in der Datendatei fest. Das Symbol darf nur aus einem Zeichen bestehen und wird sowohl zur Darstellung **systembedingt fehlender** als auch **benutzerdefiniert fehlender** Werte verwendet.

Seitennumerierung beginnen mit. Mit dieser Option können Sie für die erste Seite des Berichts eine Seitennummer festlegen.

Bericht: Layout

Mit "Bericht: Layout" werden Breite und Länge jeder Berichtsseite, Seitenanordnung des Berichts sowie Einfügen von Leerzeilen und Beschriftungen festgelegt.

Abbildung 29-6
Dialogfeld "Bericht: Layout"

Seitenformat. Legt die Seitenränder, ausgedrückt in Zeilen (oben und unten) und Leerzeichen (links und rechts) sowie die Ausrichtung der Berichte innerhalb der Ränder fest.

Titel und Fußzeilen der Seite. Legt die Anzahl von Zeilen fest, welche die Kopf- und Fußzeile jeweils vom Text des Berichts trennen.

Break-Spalten. Hiermit wird die Anzeige der Break-Spalten festgelegt. Wenn mehrere Break-Variablen festgelegt wurden, können sie sich in getrennten Spalten oder in der ersten Spalte befinden. Das Anordnen aller Break-Variablen in der ersten Spalte erzeugt einen schmaleren Bericht.

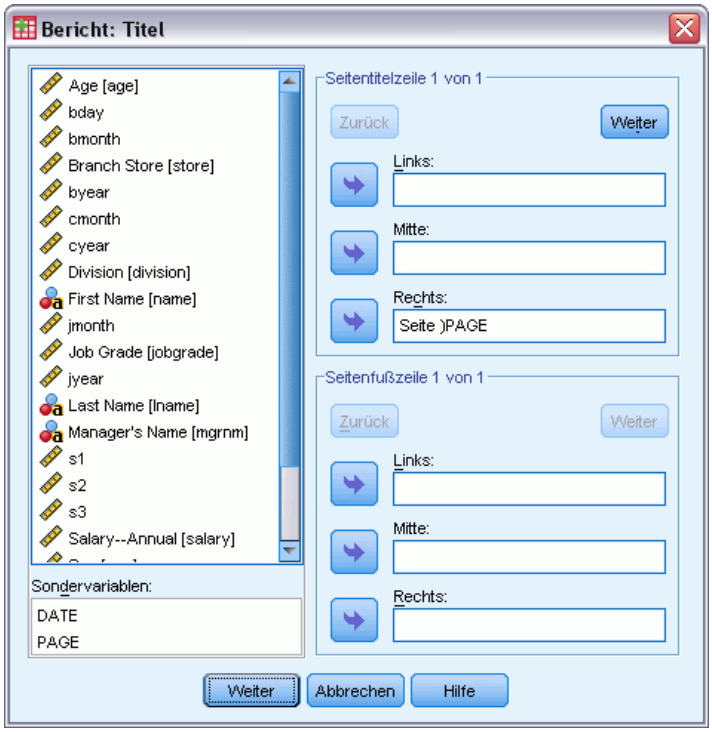
Spaltentitel. Legt die Anzeige von Spaltentiteln fest und umfaßt Unterstreichung des Titels, Anzahl von Leerzeilen zwischen Titel und Text des Berichts sowie die vertikale Ausrichtung.

Beschriftung für Zeilen und Breaks der Datenspalte. Steuert die Anordnung von Informationen in Datenspalten (Datenwerte und/oder Auswertungsstatistiken) bezüglich der Break-Beschriftungen zu Beginn jeder Break-Kategorie. Die erste Informationszeile in der Datenspalte kann entweder in der gleichen Zeile wie die Beschriftung der Break-Kategorie oder nach einer festgelegten Anzahl von Zeilen nach der Beschriftung der Break-Kategorie beginnen. (Nicht für Auswertungsberichte in Spalten verfügbar.)

Bericht: Titel

Im Dialogfeld “Bericht: Titel” werden Inhalt und Anordnung der Titel- und Fußzeilen des Berichts festgelegt. Sie können jeweils bis zu zehn Titel- und Fußzeilen festlegen, wobei in jeder Zeile linksbündige, zentrierte oder rechtsbündige Komponenten enthalten sein können.

Abbildung 29-7
Dialogfeld "Bericht: Titel"



Wenn Sie in Titeln oder Fußzeilen Variablen eingeben, wird das aktuelle Wertelabel oder der Wert der Variablen im Titel oder in der Fußzeile angezeigt. In Titeln wird das Wertelabel angezeigt, das dem Wert der Variablen am Beginn der Seite entspricht. In den Fußzeilen wird das Wertelabel angezeigt, das dem Wert der Variablen am Ende der Seite entspricht. Ist kein Wertelabel vorhanden, wird der aktuelle Wert angezeigt.

Sondervariablen. Mit den Sondervariablen *DATE* und *PAGE* können Sie das aktuelle Datum oder die Seitenzahl in eine beliebige Zeile des Kopf- oder Fußzeilenbereichs des Berichts eingeben. Wenn Ihre Datendatei Variablen wie *DATE* oder *PAGE* enthält, können Sie diese in Titeln oder Fußzeilen des Berichts nicht verwenden.

Bericht in Spalten

Mit "Bericht in Spalten" werden Auswertungsberichte erstellt, die in verschiedenen Spalten unterschiedliche Auswertungsstatistiken enthalten.

Beispiel. In einem Einzelhandelsunternehmen mit Filialen werden Informationen über Angestellte, Gehälter, Anstellungszeiten sowie Filiale und Abteilung jedes Beschäftigten in Datensätzen gespeichert. Sie können einen Bericht erstellen, der eine zusammenfassende Gehaltsstatistik (zum Beispiel Mittelwert, Minimum und Maximum) für jede Abteilung liefert.

Datenspalten. Hier werden die Berichtsvariablen aufgelistet, für die Sie eine Auswertungsstatistik anfordern möchten, und das Anzeigeformat sowie die für jede Variable angezeigte Auswertungsstatistik festgelegt.

Break-Spalten. Hiermit werden optionale Break-Variablen, die den Bericht in Gruppen aufteilen, aufgelistet und das Anzeigeformat der Break-Spalten festgelegt. Bei mehreren Break-Variablen wird für jede Kategorie einer Break-Variablen eine getrennte Gruppe innerhalb der Kategorien der vorhergehenden Break-Variablen in der Liste erzeugt. Die Break-Variablen müssen diskrete kategoriale Variablen sein, welche die Fälle in eine begrenzte Anzahl von sinnvollen Kategorien aufteilen.

Bericht. Hiermit legen Sie alle Merkmale des Berichts fest, beispielsweise die Anzeige der fehlenden Werte, Seitennumerierung und Titel.

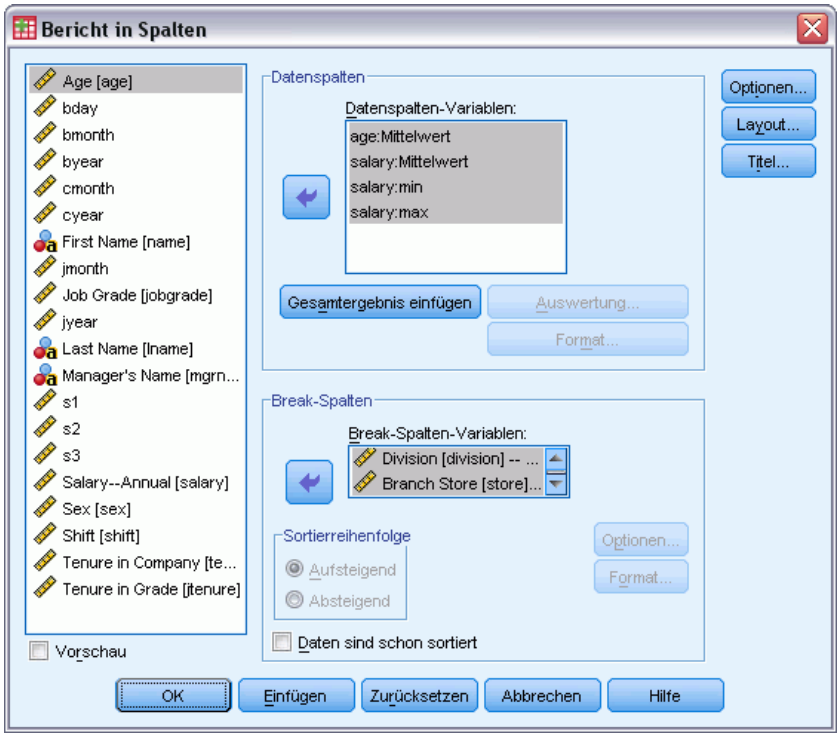
Vorschau. Es wird nur die erste Seite des Berichtes angezeigt. Mit dieser Option erhalten Sie eine Vorschau auf das Format Ihres Berichts, ohne diesen komplett bearbeiten zu müssen.

Daten sind schon sortiert. Bei Berichten mit Break-Variablen muss die Datendatei vor dem Erstellen des Berichts nach den Werten der Break-Variablen sortiert werden. Wenn Ihre Datendatei bereits nach den Werten der Break-Variablen sortiert ist, können Sie durch Auswählen dieser Option Rechenzeit sparen. Diese Option ist besonders hilfreich, wenn Sie bereits einen Bericht für die Vorschau erstellt haben.

So erstellen Sie eine Zusammenfassung: Bericht in Spalten

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Berichte > Bericht in Spalten...
- Wählen Sie mindestens eine Variable für die Datenspalten aus. Für jede ausgewählte Variable wird eine Spalte im Bericht erzeugt.
- Um das Auswertungsmaß für eine Variable zu ändern, wählen Sie die Variable in der Liste “Datenspalten-Variablen” aus und klicken Sie auf Auswertung.
- Um mehr als ein Auswertungsmaß für eine Variable berechnen zu lassen, wählen Sie die Variable in der Quellliste aus und übernehmen diese für jedes gewünschte Auswertungsmaß in die Liste “Datenspalten-Variablen”.
- Um eine Spalte mit Summe, Mittelwert, Verhältnis oder einer anderen Funktion einer vorhandenen Spalte anzuzeigen, klicken Sie auf Gesamtergebnis einfügen. Dadurch wird die Variable *Gesamt* in die Liste “Datenspalten” aufgenommen.
- Wählen Sie bei sortierten und nach Untergruppen angezeigten Berichten mindestens eine Variable für die Break-Spalten aus.

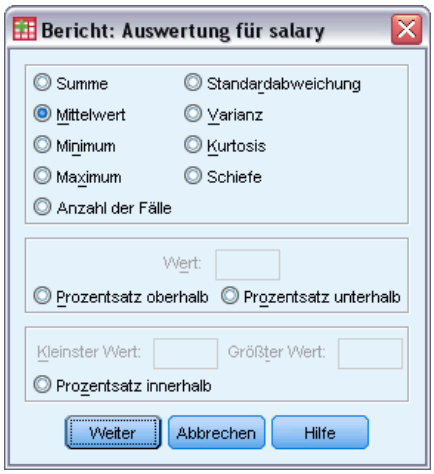
Abbildung 29-8
Dialogfeld "Bericht in Spalten"



Datenspalten: Auswertungsfunktion

Im Dialogfeld "Auswertung" wird die angezeigte Auswertungsstatistik der ausgewählten Datenspalten-Variablen festgelegt.

Abbildung 29-9
Dialogfeld "Bericht: Auswertung"



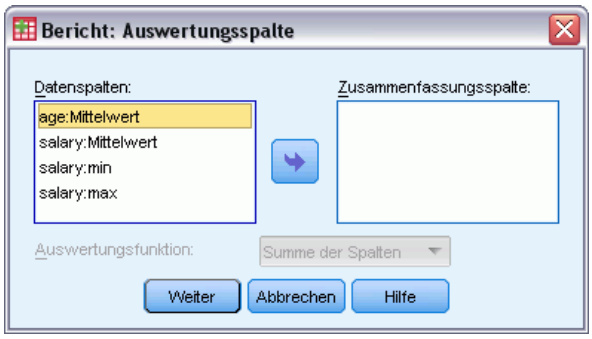
Die verfügbaren Auswertungsstatistiken sind Summe, Mittelwert, Minimum, Maximum, Anzahl der Fälle, Prozent der Fälle über oder unter einem festgelegten Wert, Prozent der Fälle innerhalb eines festgelegten Wertebereichs, Standardabweichung, Varianz, Kurtosis und Schiefe.

Auswertungsspalte für Gesamtergebnis

Im Dialogfeld "Bericht: Auswertungsspalte" werden Einstellungen für die Gesamt-Auswertungsstatistik festgelegt, die zwei oder mehr Datenspalten zusammenfaßt.

Die folgenden Gesamt-Auswertungsstatistiken sind verfügbar: Summe der Spalten, Mittelwert der Spalten, Minimum, Maximum, Differenz zwischen den Werten zweier Spalten, Quotient der Werte in einer Spalte dividiert durch die Werte einer anderen Spalte und das Produkt der miteinander multiplizierten Spaltenwerte.

Abbildung 29-10
Dialogfeld "Bericht: Auswertungsspalte"



Summe der Spalten. Die Spalte *Gesamt* enthält die Summe der Spalten in der Liste "Zusammenfassungsspalte".

Mittelwert der Spalten. Die Spalte *Gesamt* enthält den Durchschnitt der Spalten in der Liste "Zusammenfassungsspalte".

Minimum der Spalten. Die Spalte *Gesamt* enthält den Minimalwert der Spalten in der Liste "Zusammenfassungsspalte".

Maximum der Spalten. Die Spalte *Gesamt* enthält den Maximalwert der Spalten in der Liste "Zusammenfassungsspalte".

1. Spalte – 2. Spalte. Die Spalte *Gesamt* enthält die Differenz zwischen den Spalten in der Liste "Zusammenfassungsspalte". Die Liste "Zusammenfassungsspalte" muss dabei genau zwei Spalten enthalten.

1. Spalte / 2. Spalte. Die Spalte *Gesamt* enthält den Quotienten der Spalten in der Liste "Zusammenfassungsspalte". Die Liste "Zusammenfassungsspalte" muss dabei genau zwei Spalten enthalten.

% 1. Spalte / 2. Spalte. Die Spalte *Gesamt* enthält den prozentualen Anteil der ersten Spalte an der zweiten Spalte in der Liste "Zusammenfassungsspalte". Die Liste "Zusammenfassungsspalte" muss dabei genau zwei Spalten enthalten.

Produkt der Spalten. Die Spalte *Gesamt* enthält das Produkt der Spalten in der Liste "Zusammenfassungsspalte".

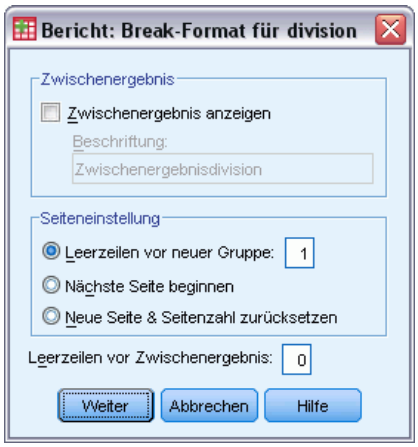
Format der Berichtsspalte

Die Formatoptionen von Daten- und Break-Spalten für “Bericht in Spalten” entsprechen den Optionen für “Bericht in Zeilen”.

Bericht: Break-Optionen für Bericht in Spalten

Mit “Break-Optionen” werden Anzeige der Zwischenergebnisse, Abstand und Seitennumerierung für Break-Kategorien festgelegt.

Abbildung 29-11
Dialogfeld “Bericht: Break-Optionen”



Zwischenergebnis. Hiermit wird die Anzeige der Zwischenergebnisse für Break-Kategorien festgelegt.

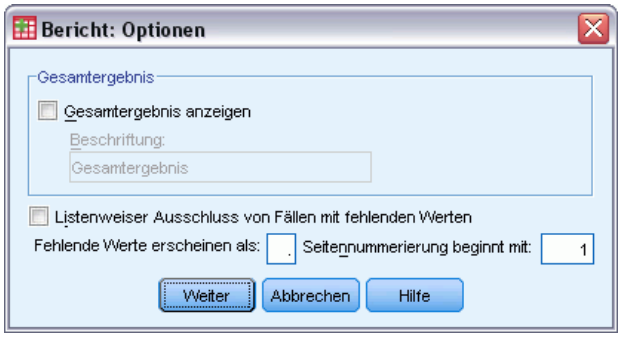
Seiteneinstellung. Hiermit werden Abstand und Seitenaufteilung für Kategorien der ausgewählten Break-Variablen festgelegt. Sie können eine Anzahl von Leerzeilen zwischen den Break-Kategorien festlegen oder eine Break-Kategorie an einen neuen Seitenanfang legen.

Leerzeilen vor Zwischenergebnis. Hiermit legen Sie die Anzahl leerer Zeilen zwischen den Daten der Break-Kategorien und den Zwischenergebnissen fest.

Bericht: Optionen für Bericht in Spalten

Mit “Optionen” werden Anzeige der Gesamtergebnisse, Anzeige der fehlenden Werte und Seitennumerierung in Auswertungsberichten in Spalten festgelegt.

Abbildung 29-12
Dialogfeld "Bericht: Optionen"



Gesamtergebnis. In jeder Spalte wird am unteren Rand ein Gesamtergebnis angezeigt und beschriftet.

Fehlende Werte. Sie können fehlende Werte vom Bericht ausschließen oder fehlende Werte mit einem ausgewählten Zeichen im Bericht kennzeichnen.

Bericht: Layout für Bericht in Spalten

Die Layout-Optionen für "Bericht in Spalten" entsprechen den Optionen für "Bericht in Zeilen".

Zusätzliche Funktionen beim Befehl REPORT

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- In den Spalten einer einzelnen Auswertungszeile lassen sich unterschiedliche Auswertungsfunktionen anzeigen.
- In Datenspalten können Auswertungszeilen für Variablen eingefügt werden, die nicht den Variablen der Datenspalten entsprechen. Außerdem können Zeilen für verschiedene Kombinationen (zusammengesetzte Funktionen) der Auswertungsfunktion eingefügt werden.
- Als Auswertungsfunktionen können Median, Modalwert, Häufigkeit und Prozent verwendet werden.
- Das Anzeigeformat der Auswertungsstatistiken kann genauer festgelegt werden.
- An verschiedenen Stellen des Berichtes können Leerzeilen eingefügt werden.
- In Listenberichten können nach jedem *n*-ten Fall Leerzeilen eingefügt werden.

Wegen der Komplexität der Syntax zum Befehl `REPORT` kann es hilfreich sein, beim Erstellen eines neuen Berichts mit Syntax auf einen vorhandenen Bericht zurückzugreifen. Zum Anpassen eines aus Dialogfeldern erstellten Berichts kopieren Sie die entsprechende Syntax, fügen diese ein und ändern sie so, dass Sie den gewünschten Bericht erstellen können.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
30

Reliabilitätsanalyse

Die Reliabilitätsanalyse ermöglicht es Ihnen, die Eigenschaften von Messniveaus und der Items zu untersuchen, aus denen diese sich zusammensetzen. Mit der Prozedur "Reliabilitätsanalyse" können Sie eine Anzahl von allgemein verwendeten Reliabilitäten des Messniveaus berechnen, und es werden Ihnen Informationen über die Beziehungen zwischen den Items in der Skala zur Verfügung gestellt. Korrelationskoeffizienten in Klassen können verwendet werden, um Reliabilitätsschätzer der Urteiler zu berechnen.

Beispiel. Wird die Kundenzufriedenheit mit Ihrem Fragebogen sinnvoll gemessen? Mit der Reliabilitätsanalyse können Sie das Ausmaß des Zusammenhangs zwischen den Items in Ihrem Fragebogen bestimmen, einen globalen Index der Reproduzierbarkeit bzw. der inneren Konsistenz der vollständigen Skala ermitteln und die kritischen Items herausfinden, welche nicht mehr in der Skala verwendet werden sollten.

Statistiken. Deskriptive Statistiken für jede Variable und für die Skala, Auswertungsstatistik für mehrere Items, Inter-Item-Korrelationen und Inter-Item-Kovarianzen, Reliabilitätsschätzer, ANOVA-Tabelle, Korrelationskoeffizient in Klassen, T^2 nach Hotelling und Tukey-Additivitätstest.

Modelle. Die folgenden Reliabilitätsmodelle sind verfügbar:

- **Alpha (Cronbach).** Dieses Modell ist ein Modell der inneren Konsistenz, welches auf der durchschnittlichen Inter-Item-Korrelation beruht.
- **Split-Half.** Bei diesem Modell wird die Skala in zwei Hälften geteilt und die Korrelation zwischen den Hälften berechnet.
- **Guttman.** Bei diesem Modell werden Guttmans untere Grenzen für die wahre Reliabilität berechnet.
- **Parallel.** Bei diesem Modell wird angenommen, dass alle Items gleiche Varianzen und gleiche Fehlervarianzen für mehrere Wiederholungen aufweisen.
- **Streng parallel.** Bei diesem Modell gelten die Annahmen des parallelen Modells, und es wird zusätzlich die Gleichheit der Mittelwerte der Items angenommen.

Daten. Die Daten können dichotom, ordinal- oder intervallskaliert sein. Sie müssen jedoch numerisch kodiert sein.

Annahmen. Die Beobachtungen sollten unabhängig sein, und Fehler dürfen zwischen den Items nicht korrelieren. Jedes Paar von Items sollte bivariat normalverteilt sein. Die Skalen sollten additiv sein, sodass sich jedes Item linear zum Gesamtwert verhält.

Verwandte Prozeduren. Wenn Sie die Dimensionalität der Skalen-Items untersuchen möchten (um herauszufinden, ob mehr als eine Konstruktion nötig ist, um das Muster der Item-Werte zu erklären), verwenden Sie die Prozedur "Faktorenanalyse" oder "Multidimensionale Skalierung".

Wenn Sie homogene Variablengruppen identifizieren möchten, verwenden Sie die Prozedur “Hierarchische Clusteranalyse”, um Variablen zu clustern.

So lassen Sie eine Reliabilitätsanalyse berechnen:

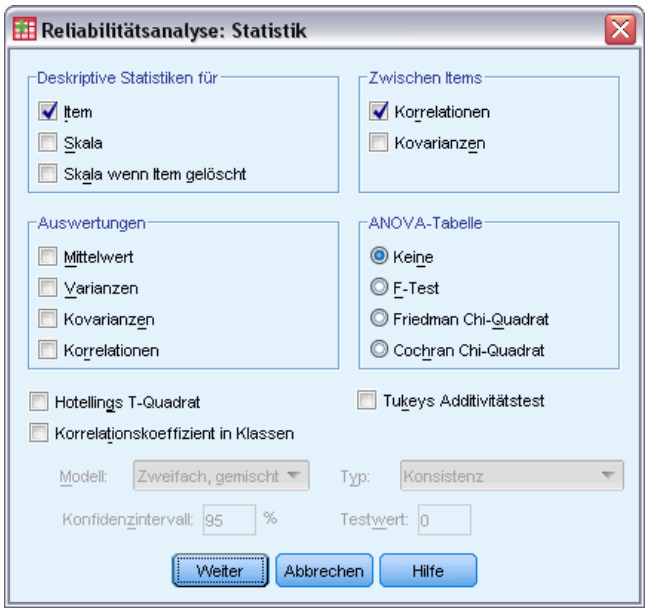
- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Skalierung > Reliabilitätsanalyse...

Abbildung 30-1
Dialogfeld “Reliabilitätsanalyse”

- Wählen Sie mindestens zwei Variablen als potentielle Komponenten einer additiven Skala aus.
- Wählen Sie aus dem Dropdown-Listenfeld “Modell” ein Modell aus.

Reliabilitätsanalyse: Statistik

Abbildung 30-2
Dialogfeld "Reliabilitätsanalyse: Statistik"



Sie können zahlreiche Statistiken auswählen, die sowohl die Skala als auch die Items beschreiben. Die Statistiken, die in der Standardeinstellung angezeigt werden, umfassen die Anzahl der Fälle, die Anzahl der Items und die folgenden Reliabilitätsschätzer:

- **Alpha-Modelle.** Bei dichotomen Daten entspricht dies dem Kuder-Richardson-20-(KR20-)Koeffizienten.
- **Split-Half-Modelle.** Korrelation zwischen den beiden Hälften, Split-Half-Reliabilität nach Guttman, Spearman-Brown-Reliabilität (gleiche und ungleiche Länge) und Alpha-Koeffizienten für jede Hälfte.
- **Guttman-Modelle.** Reliabilitätskoeffizienten Lambda 1 bis Lambda 6.
- **Parallele und streng parallele Modelle.** Anpassungstest für das Modell, Schätzer der Fehlervarianz, der Gesamtvarianz und der wahren Varianz, geschätzte gemeinsame Inter-Item-Korrelation, geschätzte Reliabilität und unverzerrter Schätzer der Reliabilität.

Deskriptive Statistiken für. Erzeugt deskriptive Statistiken für Skalen oder Items über Fälle.

- **Item.** Erzeugt deskriptive Statistiken für Items über Fälle.
- **Skala.** Erzeugt deskriptive Statistiken für Skalen.
- **Skala, wenn Item gelöscht.** Zeigt die Auswertungsstatistik an, bei der jedes Item mit der Skala verglichen wird, die aus den anderen Items aufgebaut wurde. Zu den statistischen Angaben gehören auch Mittelwert und Varianz der Skala, falls das Item aus der Skala gelöscht würde, die Korrelation zwischen dem Element und der Skala aus den anderen Items sowie Cronbachs Alpha, falls das Element aus der Skala gelöscht würde.

Auswertung. Hiermit werden deskriptive Statistiken der Item-Verteilungen für alle Items in der Skala berechnet.

- **Mittelwerte.** Auswertungsstatistik für die Mittelwerte der Items. Auswertungsstatistik für Varianzen der Items. Es werden die kleinsten, größten und mittleren Varianzen der Items, die Spannweite und die Varianz der Item-Varianzen sowie das Verhältnis zwischen der größten und der kleinsten Varianzen angezeigt.
- **Varianzen.** Auswertungsstatistik für Varianzen der Items. Auswertungsstatistik für Varianzen der Items. Es werden die kleinsten, größten und mittleren Varianzen der Items, die Spannweite und die Varianz der Item-Varianzen sowie das Verhältnis zwischen der größten und der kleinsten Varianzen angezeigt.
- **Kovarianzen.** Statistik für die Kovarianzen zwischen den Items. Von den Kovarianzen zwischen den Items werden der kleinste und der größte Wert, der Mittelwert, die Spannweite und die Varianz sowie das Verhältnis vom größten zum kleinsten Wert angezeigt.
- **Korrelationen.** Statistik für die Korrelationen zwischen den Items. Statistik für die Korrelationen zwischen den Items. Von den Korrelationen zwischen den Items werden der kleinste und der größte Wert, der Mittelwert, die Spannweite und die Varianz, sowie das Verhältnis vom größten zum kleinsten Wert angezeigt.

Inter-Item. Hiermit werden Matrizen der Korrelationen oder Kovarianzen zwischen den Items erstellt.

ANOVA-Tabelle. Hiermit werden Tests auf gleiche Mittelwerte berechnet.

- **F-Test.** Zeigt eine Tabelle zur Varianzanalyse mit Messwiederholungen an.
- **Friedman Chi-Quadrat.** Zeigt das Chi-Quadrat nach Friedman und den Konkordanz-Koeffizienten nach Kendall an. Diese Option ist für Daten geeignet, die in Form von Rängen vorliegen. Der Chi-Quadrat-Test ersetzt den üblichen F-Test in der ANOVA-Tabelle.
- **Cochran Chi-Quadrat.** Zeigt Cochrans Q-Test an. Diese Option ist für dichotome Daten geeignet. Die Q-Statistik ersetzt die übliche F-Statistik in der ANOVA-Tabelle.

Hotellings T-Quadrat Erzeugt einen multivariaten Test der Nullhypothese, dass alle Items auf der Skala den gleichen Mittelwert besitzen.

Tukeys Additivitätstest Erzeugt einen Test der Annahme, dass zwischen den Items keine multiplikative Wechselwirkung besteht.

Korrelationskoeffizienten in Klassen. Erzeugt ein Maß der Konsistenz oder Werteübereinstimmung innerhalb von Fällen.

- **Modell.**Wählen Sie das Modell für die Berechnung des Korrelationskoeffizienten in Klassen. Verfügbar sind die Modelle "Zwei-Weg, gemischt", "Zwei-Weg, zufällig" und "Ein-Weg, zufällig". Wählen Sie Zwei-Weg, gemischt aus, wenn die Personeneffekte zufällig und die Item-Effekte fest sind. Wählen Sie Zwei-Weg, zufällig aus, wenn die Personeneffekte und die Item-Effekte zufällig sind. Wählen Sie Ein-Weg, zufällig aus, wenn die Personeneffekte zufällig sind.
- **Typ.**Wählen Sie den Indextyp. "Konsistenz" und "Absolute Übereinstimmung" sind verfügbar.
- **Konfidenzintervall.**Legen Sie das Niveau des Konfidenzintervalls fest. Der Standardwert ist 95%.
- **Testwert.**Legen Sie den hypothetischen Wert des Koeffizienten für den Hypothesentest fest. Dies ist der Wert, mit dem der beobachtete Wert verglichen wird. Der Standardwert ist 0.

Zusätzliche Funktionen beim Befehl RELIABILITY

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Korrelationsmatrizen können gelesen und analysiert werden.
- Korrelationsmatrizen können für spätere Analysen gespeichert werden.
- Für die Split-Half-Methode können Aufteilungen festgelegt werden, die nicht genau Hälften entsprechen.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
31

Multidimensionale Skalierung

Bei der multidimensionalen Skalierung wird versucht, die Struktur in einem Set von Distanzmaßen zwischen Objekten oder Fällen zu erkennen. Diese Aufgabe wird durch das Zuweisen von Beobachtungen zu bestimmten Positionen in einem konzeptuellen Raum (gewöhnlich zwei- oder dreidimensional) erzielt, und zwar so, dass die Distanzen zwischen den Punkten des Raums mit den gegebenen Unähnlichkeiten so gut wie möglich übereinstimmen. In vielen Fällen können die Dimensionen dieses konzeptuellen Raums interpretiert und für ein besseres Verständnis Ihrer Daten verwendet werden.

Wenn Sie über objektiv gemessene Variablen verfügen, können Sie die multidimensionale Skalierung als Technik zur Datenreduktion verwenden (erforderlichenfalls berechnet die Prozedur "Multidimensionale Skalierung" die Distanzen aus multivariaten Daten für Sie). Die multidimensionale Skalierung kann auch auf subjektive Einschätzungen von Unähnlichkeiten zwischen Objekten oder Konzepten angewendet werden. Außerdem kann die Prozedur "Multidimensionale Skalierung" Unähnlichkeitsdaten aus mehreren Quellen verarbeiten, beispielsweise von mehreren Befragern oder Befragten einer Umfrage.

Beispiel. Wie nehmen Personen Ähnlichkeiten zwischen unterschiedlichen Autos wahr? Wenn Sie über Daten verfügen, in denen Befragte ihre Einschätzungen der Ähnlichkeiten von verschiedenen Automarken und -modellen abgegeben haben, kann die multidimensionale Skalierung zur Identifizierung der Dimensionen verwendet werden, welche die Wahrnehmungen von Käufern beschreibt. Sie könnten zum Beispiel feststellen, dass Preis und Größe eines Fahrzeuges einen zweidimensionalen Raum definieren, welcher die von den Befragten geäußerten Ähnlichkeiten erklärt.

Statistiken. Für jedes Modell: Datenmatrix, optimal skalierte Datenmatrix, S-Stress (Young), Stress (Kruskal), RSQ, Stimulus-Koordinaten, durchschnittlicher Stress und RSQ für jeden Stimulus (RMDS-Modelle). Für Modelle der individuellen Differenzen (INDSCAL): Subjektgewichtungen und Seltsamkeits-Index ("weirdness index") für jedes Subjekt. Für jede Matrix in replizierten Modellen für die multidimensionale Skalierung: Stress und RSQ für jeden Stimulus. Diagramme: Stimulus-Koordinaten (zwei- oder dreidimensional), Streudiagramm von Unähnlichkeiten über Distanzen.

Daten. Wenn Sie über Unähnlichkeitsdaten verfügen, sollten alle Unähnlichkeiten quantitativ und mit derselben Maßeinheit gemessen sein. Wenn Sie über multivariate Daten verfügen, können die Variablen quantitativ, binär oder Häufigkeitsdaten sein. Die Skalierung der Variablen ist ein wichtiger Punkt. Unterschiede in der Skalierung können Ihre Lösung beeinflussen. Wenn Ihre Variablen große Differenzen in der Skalierung aufweisen (wenn zum Beispiel eine Variable in Dollar und die andere Variable in Jahren gemessen wird), sollten Sie deren Standardisierung in Betracht ziehen (dies kann mit der Prozedur "Multidimensionale Skalierung" automatisch durchgeführt werden).

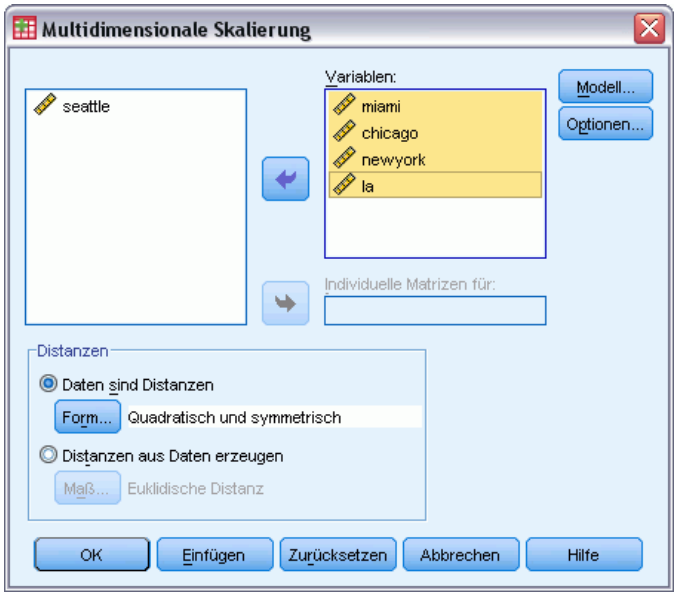
Annahmen. Die Prozedur “Multidimensionale Skalierung” ist relativ frei von Annahmen zur Verteilung. Stellen Sie sicher, dass Sie im Dialogfeld “Multidimensionale Skalierung: Optionen” ein geeignetes Messniveau auswählen (Ordinal-, Intervall- oder Verhältnisdaten), sodass Ihre Ergebnisse richtig berechnet werden können.

Verwandte Prozeduren. Wenn Sie eine Datenreduktion durchführen möchten, können Sie auch eine Faktoranalyse durchführen, insbesondere bei quantitativen Variablen. Wenn Sie Gruppen von ähnlichen Fällen identifizieren möchten, können Sie die multidimensionale Skalierung durch eine hierarchische Clusteranalyse oder eine Clusterzentrenanalyse ergänzen.

So berechnen Sie eine multidimensionale Skalierung:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Skalierung > Multidimensionale Skalierung...

Abbildung 31-1
Dialogfeld “Multidimensionale Skalierung”



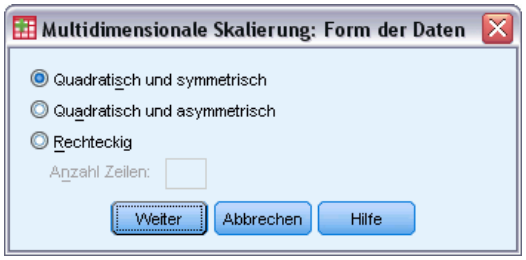
- Wählen Sie für die Analyse mindestens vier numerische Variablen aus.
- Wählen Sie in der Gruppe “Distanzen” entweder Daten sind Distanzen oder Distanzen aus Daten erzeugen aus.
- Wenn Sie Distanzen aus Daten erzeugen auswählen, können Sie für einzelne Matrizen auch eine Gruppierungsvariable auswählen. Die Gruppierungsvariable kann eine numerische Variable oder eine String-Variable sein.

Außerdem sind die folgenden Optionen verfügbar:

- Geben Sie die Form der Distanz-Matrix an, wenn es sich bei den Daten um Distanzen handelt.
- Geben Sie das Distanzmaß an, das beim Erzeigen von Distanzen aus Daten verwendet werden soll.

Multidimensionale Skalierung: Form der Daten

Abbildung 31-2
Dialogfeld "Multidimensionale Skalierung: Form der Daten"



Wenn die Arbeitsdatei Distanzen innerhalb einer Gruppe von Objekten oder zwischen zwei Gruppen von Objekten darstellt, müssen Sie die Form der Datenmatrix angeben, um die richtigen Ergebnisse zu erhalten.

Hinweis: Sie können Quadratisch und symmetrisch nicht auswählen, wenn im Dialogfeld "Modell" eine Konditionalität der Zeilen festgelegt ist.

Multidimensionale Skalierung: Distanzen aus Daten erstellen

Abbildung 31-3
Dialogfeld "Multidimensionale Skalierung: Distanzen aus Daten erstellen"

Die multidimensionale Skalierung verwendet Unähnlichkeitsdaten, um eine Skalierungslösung zu erstellen. Wenn Ihre Daten multivariate Daten darstellen (Werte gemessener Variablen), müssen Sie Unähnlichkeitsdaten erstellen, um eine multidimensionale Skalierungslösung berechnen zu können. Sie können Optionen für das Erstellen von Unähnlichkeitsmaßen aus Ihren Daten festlegen.

Maß. Hier können Sie das Unähnlichkeitsmaß für Ihre Analyse festlegen. Wählen Sie im Gruppenfeld "Maß" die Option aus, die Ihrem Datentyp entspricht. Wählen Sie dann aus dem Dropdown-Listenfeld ein Maß aus, das diesem Messwerttyp entspricht. Die folgenden Optionen sind verfügbar:

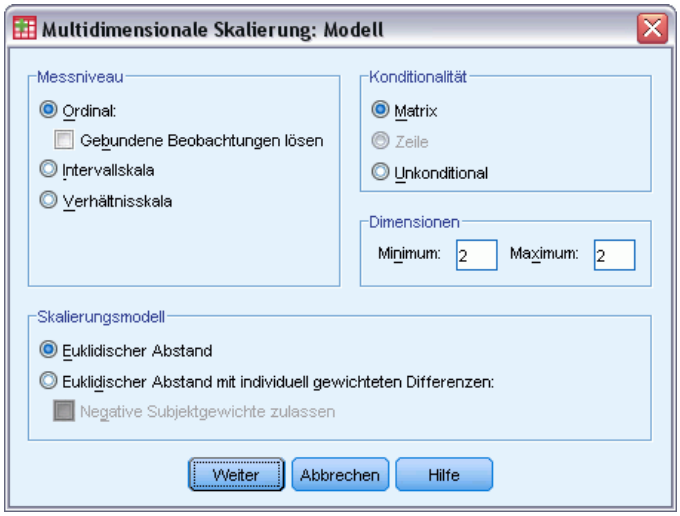
- **Intervall**. Euklidischer Abstand, quadrierter Euklidischer Abstand, Tschebyscheff, Block, Minkowski oder ein benutzerdefiniertes Maß.
- **Häufigkeiten**. Chi-Quadrat-Maß oder Phi-Quadrat-Maß.
- **Binär**. Euklidischer Abstand, quadrierter Euklidischer Abstand, Größendifferenz, Musterdifferenz, Varianz und Distanzmaß nach Lance und Williams.

Distanzmatrix erstellen. Mit dieser Funktion können Sie die Einheit der Analyse wählen. Zur Auswahl stehen "Zwischen den Variablen" oder "Zwischen den Fällen".

Werte transformieren. In bestimmten Fällen, zum Beispiel wenn die Variablen mit sehr unterschiedlichen Skalen gemessen werden, empfiehlt sich das Standardisieren der Werte vor dem Berechnen der Ähnlichkeiten (nicht auf binäre Daten anwendbar). Wählen Sie in der Dropdown-Liste "Standardisieren" eine Standardisierungsmethode aus. Wenn keine Standardisierung erforderlich ist, wählen Sie Keine aus.

Multidimensionale Skalierung: Modell

Abbildung 31-4
Dialogfeld "Multidimensionale Skalierung: Modell"



Die richtige Schätzung eines Modells für die multidimensionale Skalierung hängt von Aspekten der Daten und dem Modell selbst ab.

Messniveau. Mit dieser Funktion können Sie das Niveau Ihrer Daten festlegen. Die Optionen "Ordinalskala", "Intervallskala" und "Verhältnisskala" sind verfügbar. Wenn die Variablen ordinal sind, können Sie Gebundene Beobachtungen lösen auswählen. Die Variablen werden dann wie stetige Variablen behandelt, sodass die Bindungen (gleiche Werte für unterschiedliche Fälle) optimal gelöst werden können.

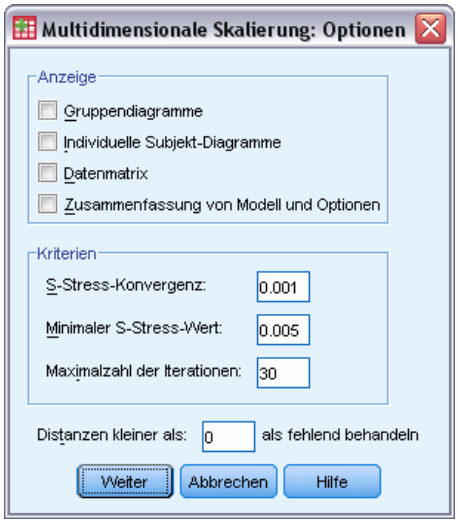
Konditionalität. Hiermit können sie festlegen, welche Vergleiche sinnvoll sind. Als Optionen sind "Matrix", "Zeile" und "Unkonditional" verfügbar.

Dimensionen. Mit dieser Funktion können Sie die Anzahl der Dimensionen für die Skalierungslösung(en) festlegen. Für jede Zahl im Bereich wird eine Lösung berechnet. Legen Sie ganze Zahlen zwischen 1 und 6 fest. Ein Minimum von 1 ist nur möglich, wenn Sie als Skalierungsmodell Euklidischer Abstand auswählen. Legen Sie die gleiche Zahl für das Minimum und das Maximum fest, wenn Sie nur eine Lösung wünschen.

Skalierungsmodell. Hiermit können Sie die Annahmen festlegen, nach denen die Skalierung durchgeführt wird. Als Optionen sind "Euklidischer Abstand" oder "Euklidischer Abstand mit individuell gewichteten Differenzen" (auch als INDSCAL bekannt) verfügbar. Beim Modell "Euklidischer Abstand mit individuell gewichteten Differenzen" können Sie Negative Subjektgewichte zulassen auswählen, wenn dies für Ihre Daten geeignet ist.

Multidimensionale Skalierung: Optionen

Abbildung 31-5
Dialogfeld "Multidimensionale Skalierung: Optionen"



Sie können Optionen für die Analyse der multidimensionalen Skalierung festlegen.

Anzeigen. Mit dieser Funktion können Sie verschiedene Ausgabetypen auswählen. Die Optionen "Gruppendiagramme", "Individuelle Subjekt-Diagramme", "Datenmatrix" und "Zusammenfassung von Modell und Optionen" sind verfügbar.

Kriterien. Hiermit können Sie bestimmen, wann die Iterationen beendet werden sollen. Um die Standardeinstellungen zu ändern, geben Sie Werte für S-Stress-Konvergenz, Minimaler S-Stress-Wert und Iterationen, max. ein.

Distanzen kleiner n als fehlend behandeln. Distanzen, die einen geringeren Wert als diesen Wert aufweisen, werden aus der Analyse ausgeschlossen.

Zusätzliche Funktionen beim Befehl ALSCAL

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Es können drei weitere Modelltypen verwenden werden. Diese sind in der Literatur über die multidimensionale Skalierung als ASCAL, AINDS und GEMSCAL bekannt.
- Es können polynomiale Transformationen von Intervall- und Verhältnisdaten ausgeführt werden.
- Bei ordinalen Daten können statt Distanzen Ähnlichkeiten analysiert werden.
- Es können nominale Daten analysiert werden.
- Verschiedene Koordinatenmatrizen und Gewichtungsmatrizen können in Dateien gespeichert und für eine Analyse erneut eingelesen werden.
- Die multidimensionale Entfaltung kann eingeschränkt werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
32

Verhältnisstatistik

Die Prozedur "Verhältnisstatistik" bietet eine umfassende Liste mit Auswertungsstatistiken zur Beschreibung des Verhältnisses zwischen zwei metrischen Variablen.

Sie können die Ausgabe nach Werten einer Gruppenvariablen in auf- oder absteigender Reihenfolge sortieren. Der Bericht für die Verhältnisstatistik kann in der Ausgabe unterdrückt werden, und die Ergebnisse können in einer externen Datei gespeichert werden.

Beispiel. Ist das Verhältnis zwischen dem Schätzwert und dem Verkaufspreis von Häusern in fünf Verwaltungsbezirken in etwa gleich? Im Ergebnis der Analyse könnte sich herausstellen, dass die Verteilung der Verhältnisse je nach Bezirk erheblich variiert.

Statistiken. Median, Mittel, gewichtetes Mittel, Konfidenzintervalle, Streuungskoeffizient (COD), medianzentrierter Variationskoeffizient, mittelzentrierter Variationskoeffizient, preisbezogenes Differential (PRD), Standardabweichung, durchschnittliche absolute Abweichung (AAD), Bereich, Mindest- und Höchstwerte sowie der Konzentrationsindex, der für einen benutzerdefinierten Bereich oder Prozentsatz innerhalb des Medianverhältnisses berechnet wird.

Daten. Verwenden Sie zum Kodieren von Gruppenvariablen (nominales oder ordinales Messniveau) numerische Codes oder Strings

Annahmen. Die Variablen, durch die Zähler und Nenner des Verhältnisses definiert werden, müssen metrische Variablen sein, die positive Werte akzeptieren.

So lassen Sie Verhältnisstatistiken berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > Deskriptive Statistiken > Verhältnis...

Abbildung 32-1
Dialogfeld "Verhältnisstatistik"

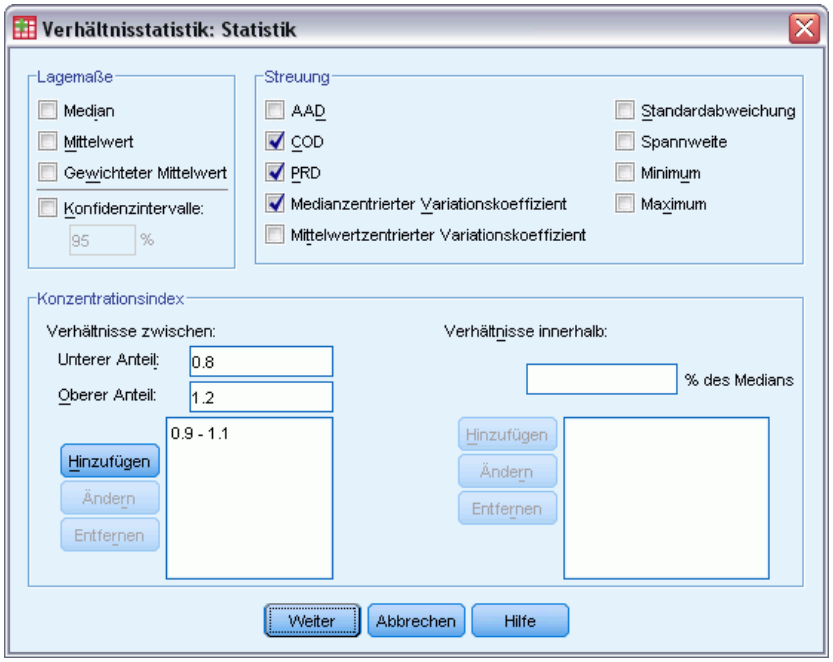
- ▶ Wählen Sie eine Zählervariable.
- ▶ Wählen Sie eine Nennervariable.

Die folgenden Optionen sind verfügbar:

- Wählen Sie eine Gruppenvariable, und legen Sie die Reihenfolge der Gruppen in den Ergebnissen fest.
- Wählen Sie aus, ob die Ergebnisse im Viewer angezeigt werden sollen.
- Legen Sie fest, ob die Ergebnisse zur späteren Verwendung in einer externen Datei gespeichert werden sollen, und geben Sie einen Namen für diese Datei an.

Verhältnisstatistik

Abbildung 32-2
Dialogfeld "Verhältnisstatistik"



Lagemaße. Lagemaße sind Statistiken, mit denen die Verteilung von Verhältnissen beschrieben wird.

- **Median.** Der Wert, der sich ergibt, wenn die Anzahl der Verhältnisse unterhalb dieses Werts gleich der Anzahl der Verhältnisse oberhalb dieses Werts ist.
- **Mittelwert.** Das Ergebnis aus der Summierung aller Verhältnisse und der anschließenden Division des Ergebnisses durch die Gesamtanzahl der Verhältnisse.
- **Gewichteter Mittelwert.** Das Ergebnis aus der Division des Mittelwerts für den Zähler durch den Mittelwert für den Nenner. Der gewichtete Mittelwert ist außerdem der Mittelwert der durch den Nenner gewichteten Verhältnisse.
- **Konfidenzintervalle.** Mit dieser Option werden Konfidenzintervalle für den Mittelwert, den Median und den gewichteten Mittelwert (falls gewünscht) angezeigt. Geben Sie für das Konfidenzniveau einen Wert größer oder gleich 0 und kleiner als 100 ein.

Streuung. Statistiken, mit denen die Variation oder Streubreite in den beobachteten Werten gemessen wird.

- **AAD.**Die durchschnittliche absolute Abweichung ist die Summe aus den absoluten Abweichungen der Verhältnisse des Medians und der Division des Ergebnisses durch die Gesamtanzahl der Verhältnisse.
- **COD.**Der Streuungskoeffizient entspricht der durchschnittlichen absoluten Abweichung in Prozent des Medians.
- **PRD.**Das preisbezogene Differential, auch Index der Regressivität genannt, ist das Ergebnis der Division des Mittelwerts durch den gewichteten Mittelwert.

- **Medianzentrierter Variationskoeffizient.**Der medianzentrierte Variationskoeffizient entspricht der Wurzel der mittleren quadratischen Abweichung vom Median in Prozent des Medians.
- **Mittelwertzentrierter Variationskoeffizient.**Der mittelwertzentrierte Variationskoeffizient entspricht der Standardabweichung in Prozent des Mittelwerts.
- **Standardabweichung.** Die Standardabweichung ist das Ergebnis der Summierung der quadratischen Abweichungen der Verhältnisse zum Mittelwert, der Division des Ergebnisses durch die Gesamtanzahl der Verhältnisse minus eins und der Berechnung der positiven Quadratwurzel.
- **Bereich.** Der Bereich ist das Ergebnis der Subtraktion des minimalen Verhältnisses vom maximalen Verhältnis.
- **Minimum.** Das Minimum ist das kleinste Verhältnis.
- **Maximum.** Das Maximum ist das größte Verhältnis.

Konzentrationsindex. Der Konzentrationskoeffizient mißt den prozentualen Anteil der Verhältnisse, die in einem bestimmten Intervall liegen. Dieser Koeffizient kann auf zwei verschiedene Arten berechnet werden:

- **Verhältnisse zwischen.**Bei dieser Option wird das Intervall explizit durch Angabe der unteren und oberen Intervallwerte definiert. Geben Sie Werte für den unteren Anteil und den oberen Anteil ein und klicken Sie auf Hinzufügen, um ein Intervall auszugeben.
- **Verhältnisse innerhalb.**Bei dieser Option wird das Intervall implizit durch Angabe des prozentualen Medians definiert. Geben Sie einen Wert zwischen 0 und 100 ein und klicken Sie auf Hinzufügen. Die untere Grenze des Intervalls ist gleich (1 – 0,01 × Wert) × Median. Die obere Grenze ist gleich (1 + 0,01 × Wert) × Median.

Kapitel
33

ROC-Kurven

Diese Prozedur stellt einen sinnvollen Weg zur Beurteilung von Klassifikationsschemata dar, bei denen eine Variable mit zwei Kategorien verwendet wird, um Subjekte zu klassifizieren.

Beispiel. Es liegt im Interesse von Banken, Kunden ordnungsgemäß danach zu klassifizieren, ob diese Kunden mit ihren Darlehen in Verzug geraten werden oder nicht. Daher werden spezielle Verfahren für diese Entscheidungen entwickelt. Mithilfe von ROC-Kurven kann beurteilt werden, wie gut diese Verfahren funktionieren.

Statistiken. Fläche unter der ROC-Kurve mit Konfidenzintervall und Koordinaten-Punkten der ROC-Kurve. Diagramme: ROC-Kurve.

Methoden. Die Schätzung der Fläche unter der ROC-Kurve kann parameterunabhängig oder parameterabhängig unter Verwendung eines binegativ exponentiellen Modells erfolgen.

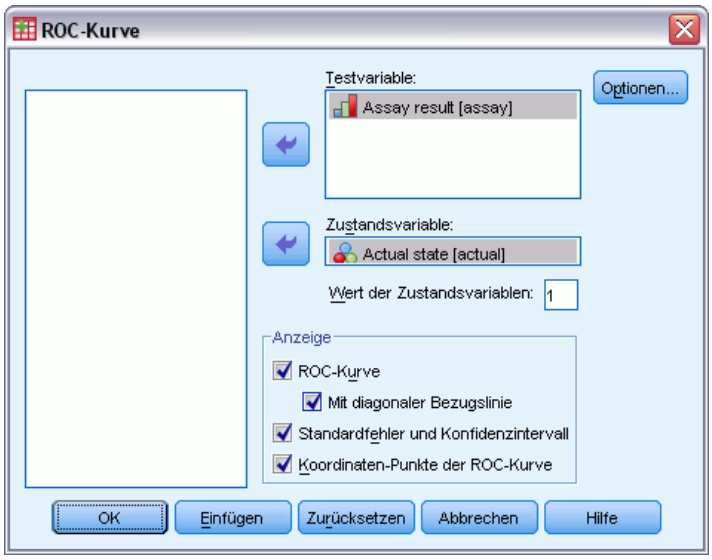
Daten. Die Testvariablen sind quantitativ. Die Testvariablen setzen sich oft aus Wahrscheinlichkeiten aus der Diskriminanzanalyse bzw. logistischen Regression zusammen oder sie werden aus Werten auf einer willkürlichen Skala zusammengesetzt, die anzeigen, wie sehr ein Beurteiler davon "überzeugt" ist, dass ein Subjekt in die eine oder die andere Kategorie fällt. Der Typ der Zustandsvariablen ist nicht vorgegeben. Diese Variable zeigt die tatsächliche Kategorie an, zu der ein Subjekt gehört. Der Wert der Zustandsvariablen zeigt an, welche Kategorie als *positiv* zu betrachten ist.

Annahmen. Es wird angenommen, dass ansteigende Werte auf der Skala des Beurteilers ein Ansteigen der Überzeugung darstellen, dass das Subjekt in die eine Kategorie fällt. Abfallende Werte auf der Skala stellen hingegen eine ansteigende Überzeugung dar, dass das Subjekts der anderen Kategorie angehört. Der Anwender wählt aus, welche Richtung als *positiv* anzusehen ist. Es wird außerdem angenommen, dass die *tatsächliche* Kategorie bekannt ist, zu der jedes Subjekt gehört.

So Erstellen Sie eine ROC-Kurve:

- Wählen Sie die folgenden Befehle aus den Menüs aus:
  Analysieren > ROC-Kurve...

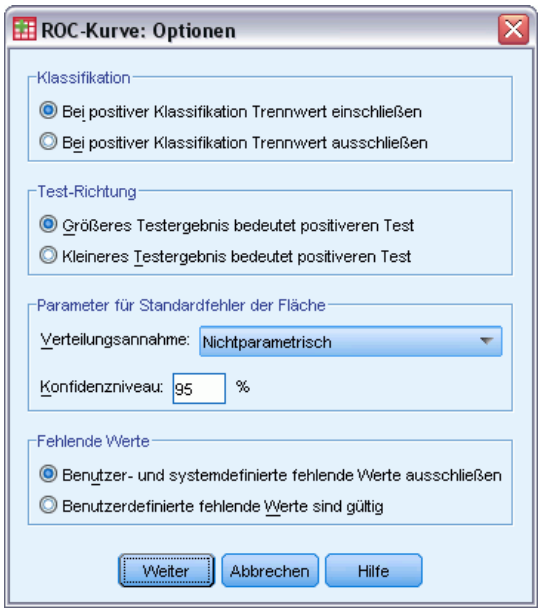
Abbildung 33-1
Dialogfeld "ROC-Kurve"



- Wählen Sie mindestens eine Wahrscheinlichkeitsvariable für den Test aus.
- Wählen Sie eine Zustandsvariable aus.
- Legen Sie den *positiven* Wert für die Zustandsvariable fest.

ROC-Kurve: Optionen

Abbildung 33-2
Dialogfeld "ROC-Kurve: Optionen"



Sie können eine der folgenden Optionen für die ROC-Analyse auswählen:

Klassifikation. Hiermit können Sie festlegen, ob der Trennwert bei einer *positiven* Klassifikation einbezogen oder ausgeschlossen werden soll. Diese Einstellung hat gegenwärtig keine Auswirkungen auf die Ausgabe.

Test-Richtung. Hiermit geben Sie die Richtung der Skala bezogen auf die *positive* Kategorie an.

Parameter für Standardfehler der Fläche. Hiermit geben Sie die Methode an, mit welcher der Standardfehler der Fläche unter der Kurve geschätzt wird. Es stehen eine nichtparametrische und eine binegative exponentielle Methode zur Verfügung. Sie können hier außerdem das Niveau des Konfidenzintervalls festlegen. Es sind Werte zwischen 50,1% und 99,9% möglich.

Fehlende Werte. Hier können Sie festlegen, wie fehlende Werte behandelt werden.

Anhang

A

Notices

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing,

using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Index