IBM SPSS Statistics Base 23

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen" auf Seite 211 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 23, Release 0, Modifikation 0 von IBM SPSS Statistics und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Diese Veröffentlichung ist eine Übersetzung des Handbuchs
IBM SPSS Statistics Base 23,
herausgegeben von International Business Machines Corporation, USA

Informationen, die nur für bestimmte Länder Gültigkeit haben und für Deutschland, Österreich und die Schweiz nicht zutreffen, wurden in dieser Veröffentlichung im Originaltext übernommen.

Möglicherweise sind nicht alle in dieser Übersetzung aufgeführten Produkte in Deutschland angekündigt und verfügbar; vor Entscheidungen empfiehlt sich der Kontakt mit der zuständigen IBM Geschäftsstelle.

Änderung des Textes bleibt vorbehalten.

Herausgegeben von:
TSC Germany
Kst. 2877
Dezember 2014

Inhaltsverzeichnis

Kapitel 1. Codebook

Codebook meldet die Datenwörterbuchinformationen – wie Variablennamen, Variablenbeschriftungen, Wertbeschriftungen, fehlende Werte – und Auswertungsstatistiken für alle oder bestimmte Variablen und Mehrfachantwortsets im aktiven Dataset. Für nominale und ordinale Variablen und Mehrfachantwortsets enthalten die Auswertungsstatistiken Häufigkeiten und Prozentangaben. Für metrische Variablen enthalten die Auswertungsstatistiken Mittelwert, Standardabweichung und Quartile.

Hinweis: Codebook ignoriert den Aufteilungsdateistatus. Hierzu gehören Aufteilungsdateigruppen, die für die multiple Imputation von fehlenden Werten erstellt wurden (verfügbar in der Erweiterungsoption "Missing Values").

Abrufen eines Codebooks

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Berichte** > **Codebook**
2. Klicken Sie auf die Registerkarte "Variablen".
3. Wählen Sie eine(s) oder mehrere Variablen und/oder Mehrfachantwortsets aus.

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 Mehrfachantwortsets angezeigt werden.
- Ändern Sie das Messniveau für Variablen in der Liste der Quellenvariablen, um die angezeigten Auswertungsstatistiken zu ändern. Weitere Informationen finden Sie im Thema „Registerkarte "Codebook-Statistiken"" auf Seite 3.

Ändern des Messniveaus

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

1. Klicken Sie mit der rechten Maustaste auf eine Variable in der Liste der Quellenvariablen.
2. Wählen Sie ein Messniveau im Popup-Menü aus.

Dadurch wird das Messniveau temporär geändert. In der Praxis ist das nur für numerische Variablen sinnvoll. Das Messniveau für Zeichenfolgevariablen ist auf nominal und ordinal beschränkt. Beide werden von derselben Codebook-Prozedur behandelt.

Registerkarte "Codebook-Ausgabe"

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

Variablenbeschreibung

Dies steuert die für jede Variable angezeigten Datenwörterbuchinformationen.

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

Beschriftung. Die deskriptive Beschriftung für die Variable oder das Mehrfachantwortset.

Typ. Grundlegender Datentyp. Entweder *Numerisch*, *Zeichenfolge* oder *Mehrfachantwortset*.

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

Messniveau. Die möglichen Werte sind *Nominal*, *Ordinal*, *Metrisch* und *Unbekannt*. Der angezeigte Wert ist das im Datenwörterbuch 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 Mehrfachantwortsets nicht verfügbar.

Hinweis: 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.

Wertbeschriftungen. Deskriptive Beschriftungen zu spezifischen Datenwerten.

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

Fehlende Werte. Benutzerdefiniert fehlende Werte. Wenn "Häufigkeit" oder "Prozent" auf der Registerkarte "Statistik" ausgewählt ist, werden definierte Wertbeschriftungen in die Ausgabe aufgenommen, selbst wenn Sie hier "Fehlende Werte" nicht auswählen. Dies ist für Mehrfachantwortsets nicht verfügbar.

Benutzerdefinierte Attribute. Benutzerdefinierte Variablenattribute. Die Ausgabe enthält sowohl die Namen als auch die Werte für Attribute von benutzerdefinierten Variablen, die den einzelnen Variablen zugeordnet sind. Dies ist für Mehrfachantwortsets nicht verfügbar.

Reservierte Attribute. Reservierte Systemvariablenattribute. Sie können die Systemattribute anzeigen, Sie sollten sie aber nicht ändern. Systemattributnamen beginnen mit einem Dollarzeichen ($) . Nicht für die Anzeige bestimmte Attribute mit Namen, die mit "@" oder "$@" beginnen, sind nicht enthalten. Die Ausgabe enthält sowohl die Namen als auch die Werte für Systemattribute, die den einzelnen Variablen zugeordnet sind. Dies ist für Mehrfachantwortsets nicht verfügbar.

Dateiinformationen

Die optionale Dateiinformationstabelle kann beliebige der folgenden Dateiattribute enthalten:

Dateiname. Name der IBM® SPSS Statistics-Datendatei. Wenn das Dataset nie in IBM SPSS Statistics-Format gespeichert wurde, gibt es keinen Datendateinamen. (Wenn in der Titelleiste des Fensters "Dateneditor" kein Dateiname angezeigt wird, hat das aktive Dataset keinen Dateinamen.)

Lage. Verzeichnis (Ordner) der IBM SPSS Statistics-Datendatei. Wenn das Dataset nie in IBM SPSS Statistics-Format gespeichert wurde, gibt es keinen Speicherort.

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

Beschriftung. Dies ist die Dateibeschriftung (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 ($) . Nicht für die Anzeige bestimmte Attribute mit Namen, die mit "@" oder "$@" 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 Mehrfachantwortsets angezeigt werden, zu steuern.

Alphabetisch. Alphabetische Reihenfolge nach Variablenname.

Datei. Die Reihenfolge, in der die Variablen im Dataset erscheinen (die Reihenfolge, in der sie im Dateneditor angezeigt werden). In aufsteigender Reihenfolge werden Mehrfachantwortsets zuletzt nach allen ausgewählten Variablen angezeigt.

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

Hinweis: 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.

Liste "Variablen". Die Reihenfolge, in der Variablen und Mehrfachantwortsets 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.

Höchstzahl der Kategorien

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

Registerkarte "Codebook-Statistiken"

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

Häufigkeiten und Prozente

Für nominale und ordinale Variablen, Mehrfachantwortsets und Werte von metrischen Variablen mit Beschriftungen 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 (zentrale Tendenz). 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 an, die den 25., 50. und 75. Perzentilen entsprechen.

Hinweis: 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. Häufigkeitszähler, Prozentsätze, kumulative 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.

Erläuterung der Daten für Häufigkeiten

Daten. Verwenden Sie zum Codieren kategorialer Variablen (nominales oder ordinales Messniveau) numerische Codes oder Zeichenfolgen.

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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Deskriptive Statistiken** > **Häufigkeiten...**
2. Wählen Sie mindestens eine kategoriale oder quantitative Variable aus.

Die folgenden Optionen sind verfügbar:

- Klicken Sie auf **Statistik**, deskriptive Statistiken für quantitative Variablen zu erhalten.
- Klicken Sie auf **Diagramme**, Balkendiagramme, Kreisdiagramme oder Histogramme erhalten.
- Klicken Sie auf **Format**, um die Reihenfolge der angezeigten Ergebnisse zu ändern.

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 (das 25., 50. und 75. Perzentil) unterteilen die Beobachtungen in vier gleich große Gruppen. Falls Sie eine gleiche Anzahl von Gruppen wünschen, die von vier abweicht, klicken Sie auf **Trennwerte für n**

gleiche Gruppen 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 (zentrale Tendenz). 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 nicht fehlenden 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, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.
- *Bereich*. 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 dafür, wie stark der Mittelwert von Stichprobe zu Stichprobe in derselben Verteilung variieren kann. 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 der Asymmetrie der 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 Schiefewert, der mehr als doppelt so groß ist wie sein Standardfehler, als Abweichung von der Symmetrie gilt.
- *Kurtosis*. Ein Maß dafür, wie sehr die Beobachtungen um einen zentralen Punkt gruppiert sind. 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 Extremwerten 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 codiert ist), wählen Sie diese Option, um den Median und die Perzentile für die ursprünglichen, nicht gruppierten Daten berechnen zu lassen.

Häufigkeiten: Diagramme

Diagrammtyp. In einem Kreisdiagramm wird der Anteil der Teile an einem Ganzen angezeigt. Jeder Ausschnitt eines Kreisdiagramms entspricht einer durch eine einzelne Gruppierungsvariable 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 Häufigkeitszählern oder Prozentsätzen beschriftet werden.

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 mehrere 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.

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

Kapitel 3. Deskriptive Statistiken

Mit der Prozedur "Deskriptive Statistiken" werden in einer einzelnen Tabelle univariate Auswertungsstatistiken für verschiedene Variablen angezeigt und standardisierte Werte (*Z*-Scores) 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*-Scores gespeichert werden, werden sie den Daten im Dateneditor hinzugefügt und sind für Diagramme, Datenlisten und Analysen verfügbar. Wenn Variablen in verschiedenen Einheiten aufgezeichnet werden (zum Beispiel Bruttoinlandsprodukt pro Kopf der Bevölkerung und Prozentsatz der Alphabetisierung), werden die Variablen durch eine *Z*-Score-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 Statistiken" wird für Sie jetzt der durchschnittliche Tagesumsatz der einzelnen Angestellten berechnet und das Ergebnis vom höchsten durchschnittlichen Umsatz zum niedrigsten durchschnittlichen Umsatz geordnet.

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

Erläuterungen der Daten für deskriptive Statistiken

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

So lassen Sie deskriptive Statistiken berechnen:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Deskriptive Statistiken** > **Deskriptive Statistiken...**
2. Wählen Sie mindestens eine Variable aus.

Die folgenden Optionen sind verfügbar:

- Wählen Sie **Standardisierte Werte als Variable speichern**, um *Z*-Scores als neue Variablen zu speichern.
- Klicken Sie auf **Optionen**, um optionale Statistiken und die Reihenfolge der Anzeige zu steuern.

Deskriptive Statistiken: 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, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.
- *Bereich*. 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 Mittelwert*. Ein Maß dafür, wie stark der Mittelwert von Stichprobe zu Stichprobe in derselben Verteilung variieren kann. 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 sehr die Beobachtungen um einen zentralen Punkt gruppiert sind. 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 Extremwerten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.
- *Schiefe*. Ein Maß der Asymmetrie der 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 Schiefewert, der mehr als doppelt so groß ist wie sein Standardfehler, als Abweichung von der Symmetrie gilt.

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

Die Befehlssyntax ermöglicht außerdem Folgendes:

- Sie können die standardisierten Werte (*Z*-Scores) 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 Befehlssyntaxreferenz.

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 Teilgesamtheiten (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 nicht parametrischer 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 Stamm-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, Stamm-Blatt-Diagramme, Histogramme, Normalverteilungsdiagramme und Diagramme der Streubreite gegen das mittlere Niveau mit Levene-Test und Transformationen.

Erläuterungen der Daten für die explorative Datenanalyse

Daten. Die Prozedur "Explorative Datenanalyse" kann für quantitative Variablen (mit Intervall- oder Verhältnismessniveau) verwendet werden. Eine Faktorvariable (zum Aufteilen der Daten in Fallgruppen) muss eine sinnvolle Anzahl von unterschiedlichen Werten (Kategorien) enthalten. Diese Werte können kurze Zeichenfolgen oder numerische Werte sein. Die Fallbeschriftungsvariable, die für die Beschriftung von Ausreißern in Boxplots verwendet wird, kann eine kurze Zeichenfolge, eine lange Zeichenfolge (die ersten 15 Byte) oder numerisch sein.

Annahmen. Ihre Daten müssen nicht symmetrisch oder normalverteilt sein.

So führen Sie eine explorative Datenanalyse aus:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Deskriptive Statistiken** > **Explorative Datenanalyse...**
2. Wählen Sie eine oder mehrere abhängige Variablen aus.

Die folgenden Optionen sind verfügbar:

- Wählen Sie mindestens eine Faktorvariable aus, mit deren Werten Fallgruppen definiert werden.
- Wählen Sie eine Identifizierungsvariable für die Beschriftung von Fällen aus.
- Klicken Sie auf **Statistik**, um Zugriff auf robuste Schätzer, Ausreißer, Perzentile und Häufigkeitstabellen erhalten.
- Klicken Sie auf **Diagramme**, um Zugriff auf Histogramme, Normalverteilungsdiagramme und Tests sowie Diagramme der Streubreite gegen das mittlere Niveau mit Levene-Statistik zu erhalten.
- Klicken Sie auf **Optionen**, um die Behandlung fehlender Werte festzulegen.

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 Streuungsmaßen 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

Boxplots. Mit diesen Optionen legen Sie fest, wie Boxplots bei mehreren abhängigen Variablen angezeigt werden. Mit **Faktorstufen zusammen** wird eine getrennte Anzeige für jede abhängige Variable generiert. 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 generiert. 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 Statistiken. Im Gruppenfeld "Deskriptive Statistiken" können Sie Stamm-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 nicht ganzzahligen 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 der Interquartilbereiche 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 auswählen:

- **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.** Die dritte Potenz jedes Datenwerts wird berechnet.

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. Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Befehlssyntaxreferenz.

Kapitel 5. Kreuztabellen

Mit der Prozedur "Kreuztabellen" erzeugen Sie Zwei- und Mehrwegetabellen. Es stehen eine Vielzahl von Tests und Zusammenhangsmaßen für Zweiwegetabellen 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 Zweiwegetabellen berechnet. Wenn Sie eine Zeile, eine Spalte und einen Schichtfaktor (Kontrollvariable) 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 Kontrollvariablen) 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 Zweiwegetabelle 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, Somers-*d*, Kendall-Tau-*b*, Kendall-Tau-*c*, Eta-Koeffizient, Cohen-Kappa, relative Risikoschätzung, Odds-Verhältnis, McNemar-Test, Cochran- und Mantel-Haenszel-Statistik sowie Spaltenanteilestatistik.

Erläuterungen der Daten für Kreuztabellen

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

Annahmen. Einige Statistiken und Maße setzen geordnete Kategorien (Ordinaldaten) oder quantitative Werte (Intervall- oder Verhältnisdaten) voraus, wie bereits im Thema über Statistiken erläutert wurde. Andere sind zulässig, wenn die Tabellenvariablen über ungeordnete Kategorien verfügen (Nominaldaten). 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 Zeichenfolgewerte handeln. Die alphabetische Ordnung der Zeichenfolgewerte gibt dabei die Reihenfolge der Kategorien vor. Bei einer Zeichenfolgevariablen 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Deskriptive Statistiken** > **Kreuztabellen...**
2. Wählen Sie eine oder mehrere Zeilenvariablen und eine oder mehrere Spaltenvariablen aus.

Die folgenden Optionen sind verfügbar:

- Wählen Sie eine oder mehrere Kontrollvariablen aus.
- Klicken Sie auf **Statistik**, um Tests und Zusammenhangsmaße der Zweiwegetabellen oder Untertabellen zu erhalten.
- Klicken Sie **Zellen**, um Informationen zu beobachteten und erwarteten Werten, Prozentsätzen und Residuen zu erhalten.
- Klicken Sie auf **Format**, um die Reihenfolge der Kategorien festzulegen.

Kreuztabellenschichten

Wenn Sie eine oder mehrere Schichtvariablen auswählen, wird für jede Kategorie jeder Schichtvariablen (Kontrollvariablen) jeweils eine Kreuztabelle erzeugt. Wenn Sie zum Beispiel über eine Zeilenvariable, eine Spaltenvariable und eine Schichtvariable mit zwei Kategorien verfügen, erhalten Sie eine Zweiwegetabelle für jede Kategorie der Schichtvariablen. Um eine weitere Schicht von Kontrollvariablen 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, gibt es eine Gruppe von Balken. 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 mehrere Variablen 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.

Im nachfolgenden Beispiel wird die Datendatei *demo.sav* (verfügbar im Verzeichnis "Samples" des Installationsverzeichnisses) verwendet:

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

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

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 Chi-Quadrat 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 Quadratwurzel aus dem Quotienten aus der Chi-Quadrat-Statistik 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 eine starke Beziehung zwischen den beiden Variablen an, Werte nahe 0 eine schwache oder fehlende Beziehung zwischen den Variablen. 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 nicht parametrisches 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 nicht parametrisches 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 codiert 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 Zeilenvariable und der andere die Spaltenvariable als Intervallvariable.

Kappa. Der Cohen-Kappa-Koeffizient misst die Übereinstimmung zwischen den Evaluierungen 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 basiert auf einer quadratischen Tabelle, in der die Zeilen- und Spaltenwerte dieselbe Skala darstellen. Jeder Zelle, in der Werte für eine, jedoch nicht die andere Variable beobachtet wurden, wird die Anzahl 0 zugewiesen. Kappa wird nicht berechnet, wenn der Datenspeichertyp (Zeichenfolge oder numerisch) der beiden Variablen nicht übereinstimmt. Bei Zeichenfolgevariablen müssen beide Variablen dieselbe definierte Länge 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 Odds-Verhältnis (Odds Ratio) kann als Schätzung für das relative Risiko verwendet werden, wenn der Faktor selten auftritt.

McNemar. Ein nicht parametrischer 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 sogenannten "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 Antwortvariablen zu testen, und zwar in Abhängigkeit von Kovariatenmustern, die durch mindestens eine Schichtvariable (Kontrollvariable) definiert werden. Beachten Sie, dass andere Statistiken schichtenweise berechnet werden, die Cochran- und die Mantel-Haenszel-Statistik dagegen einmal für alle Schichten berechnet werden.

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. Sie können festlegen, dass Häufigkeiten ausgeblendet werden, wenn sie einen bestimmten ganzzahligen Wert unterschreiten. Ausgeblendete Werte werden als **<N** angezeigt. Dabei ist **N** die angegebene Ganzzahl. Die angegebene Ganzzahl muss größer oder gleich 2 sein. Allerdings ist der Wert 0 zulässig und gibt an, dass keine Häufigkeiten (Anzahlwerte) ausgeblendet werden.

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. *Hinweis*: Wenn diese Option festgelegt wird, ohne die beobachtete Anzahl oder die Spaltenprozente auszuwählen, werden die Werte für die beobachtete Anzahl mit in die Kreuztabelle aufgenommen, wobei die tiefgestellten Buchstaben im APA-Stil das Ergebnis der Tests für die Spaltenanteile angeben.

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

Prozentsätze. Die Prozentsätze 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. *Hinweis*: Wenn in der Gruppe "Häufigkeiten" die Option **Kleine Werte für Häufigkeiten ausblenden** ausgewählt ist, werden die den ausgeblendeten Häufigkeiten zugeordneten Prozentsätze ebenfalls ausgeblendet.

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 einer Schätzung 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.

Nicht ganzzahlige Gewichtungen. Bei den Zellenhäufigkeiten 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 Zellenhäufigkeiten ebenfalls Bruchwerte sein. Sie können die Werte vor oder nach der Berechnung der Zellenhäufigkeiten abschneiden oder runden oder sowohl für die Tabellenanzeige als auch für statistische Berechnungen gebrochene Zellenhäufigkeiten verwenden.

- *Anzahl in den Zellen runden*. Fallgewichtungen werden verwendet wie gegeben, aber die addierten Gewichtungen für die Zellen werden gerundet, bevor Statistiken berechnet werden.
- *Anzahl in den Zellen kürzen*. Fallgewichtungen werden unverändert verwendet, aber die addierten Gewichtungen für die Zellen werden gekürzt, bevor Statistiken berechnet werden.
- *Fallgewichtungen runden*. Fallgewichtungen werden gerundet, bevor sie verwendet werden.
- *Fallgewichtungen kürzen*. Fallgewichtungen werden gekürzt, bevor sie verwendet werden.
- *Keine Korrekturen*. Fallgewichtungen werden wie vorgegeben verwendet und auch nicht ganzzahlige Zellenanzahlen werden verwendet. Wenn jedoch exakte Statistiken (verfügbar mit dem Modul "Exakte Tests") angefordert werden, dann werden die akkumulierten Gewichtungen in den Zellen entweder auf den ganzzahligen Anteil gekürzt oder gerundet, bevor die Statistiken für exakte Tests berechnet werden.

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 Gruppierungsvariablen berechnet. Alle Ebenen der Gruppierungsvariablen 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 Datasets 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 der größte Umsatz mit gewerblichen Kunden in der westlichen Region erzielt wird.

Statistik. Summe, Anzahl der Fälle, Mittelwert, Median, gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppierungsvariablen, Variablenwert der letzten Kategorie der Gruppierungsvariablen, 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.

Erläuterungen der Daten für das Zusammenfassen

Daten. Die Gruppierungsvariablen stellen kategoriale Variablen dar, deren Werte numerisch oder Zeichenfolgen 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Berichte** > **Fallzusammenfassungen**
2. Wählen Sie mindestens eine Variable aus.

Die folgenden Optionen sind verfügbar:

- Wählen Sie eine oder mehrere Gruppierungsvariablen aus, um die Daten in Untergruppen aufzuteilen.
- Klicken Sie auf **Optionen**, wenn Sie den Ausgabetitel ändern, eine Titelzeile unter der Ausgabe hinzufügen oder Fälle mit fehlenden Werten ausschließen möchten.
- Klicken Sie auf **Statistik**, um optionale Statistiken anzuzeigen.
- Wählen Sie **Fälle anzeigen** aus, um die Fälle in jeder Untergruppe aufzulisten. 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 inaktivieren, um alle Fälle auflisten zu lassen.

Zusammenfassen: Optionen

Sie können den Titel der Ausgabe ändern oder eine Titelzeile hinzufügen, die unter der Ausgabetabelle angezeigt wird. Sie können den Zeilenumbruch in Titeln und Titelzeilen steuern, indem Sie an den 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.

Zusammenfassen: Statistik

Sie können mindestens eine der folgenden Untergruppenstatistiken für die Variablen innerhalb der einzelnen Kategorien jeder Gruppierungsvariablen auswählen: Summe, Anzahl der Fälle, Mittelwert, Median, gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppierungsvariablen, Variablenwert der letzten Kategorie der Gruppierungsvariablen, 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, 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 codiert wurden. Wenn z. B. für das Alter jeder Wert in den Dreißigern als 35 codiert ist, jeder Wert in den Vierzigern als 45 usw., dann wird der gruppierte Median aus den codierten Daten berechnet.

Harmonisches Mittel. Wird verwendet, um die durchschnittliche 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 reziproken Werte der Stichprobengrößen.

Kurtosis. Ein Maß dafür, wie sehr die Beobachtungen um einen zentralen Punkt gruppiert sind. 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 Extremwerten 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 (zentrale Tendenz). 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.

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

Schiefe. Ein Maß der Asymmetrie der 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 Schiefewert, der mehr als doppelt so groß ist wie sein Standardfehler, als Abweichung von der Symmetrie gilt.

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.

Standardfehler der Kurtosis. Der Quotient aus der Kurtosis und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) 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 des Mittelwerts. Ein Maß dafür, wie stark der Mittelwert von Stichprobe zu Stichprobe in derselben Verteilung variieren kann. 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.)

Standardfehler der Schiefe. Der Quotient aus der Schiefe und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) 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 nicht fehlenden Werten.

Varianz. Ein Maß der Streuung um den Mittelwert, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. 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.

Statistik. Summe, Anzahl der Fälle, Mittelwert, Median, gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppierungsvariablen, Variablenwert der letzten Kategorie der Gruppierungsvariablen, 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 und die Linearitätstests *R* und R^2 zur Verfügung.

Erläuterungen der Daten für Mittelwerte

Daten. Die abhängigen Variablen sind quantitativ, die unabhängigen Variablen kategorial. Die Werte der kategorialen Variablen können numerische Variablen oder Zeichenfolgevariablen 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Mittelwerte vergleichen** > **Mittelwerte...**
2. Wählen Sie eine oder mehrere abhängige Variablen aus.
3. 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.
4. 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

Sie können mindestens eine der folgenden Untergruppenstatistiken für die Variablen innerhalb der einzelnen Kategorien jeder Gruppierungsvariablen auswählen: Summe, Anzahl der Fälle, Mittelwert, Median, gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppierungsvariablen, Variablenwert der letzten Kategorie der Gruppierungsvariablen, 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. 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 codiert wurden. Wenn z. B. für das Alter jeder Wert in den Dreißigern als 35 codiert ist, jeder Wert in den Vierzigern als 45 usw., dann wird der gruppierte Median aus den codierten Daten berechnet.

Harmonisches Mittel. Wird verwendet, um die durchschnittliche 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 reziproken Werte der Stichprobengrößen.

Kurtosis. Ein Maß dafür, wie sehr die Beobachtungen um einen zentralen Punkt gruppiert sind. 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 Extremwerten 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 (zentrale Tendenz). 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.

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

Schiefe. Ein Maß der Asymmetrie der 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 Schiefewert, der mehr als doppelt so groß ist wie sein Standardfehler, als Abweichung von der Symmetrie gilt.

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.

Standardfehler der Kurtosis. Der Quotient aus der Kurtosis und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) 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 des Mittelwerts. Ein Maß dafür, wie stark der Mittelwert von Stichprobe zu Stichprobe in derselben Verteilung variieren kann. 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.)

Standardfehler der Schiefe. Der Quotient aus der Schiefe und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) 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 nicht fehlenden Werten.

Varianz. Ein Maß der Streuung um den Mittelwert, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. 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 nicht lineare Komponenten die Quadratsummen, die Freiheitsgrade und das Mittel der Quadrate sowie den F-Quotienten, R und R-Quadrat. Die Berechnungen für Linearität werden nicht durchgeführt, wenn die unabhängige Variable eine kurze Zeichenfolgevariable 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 Gruppierungsvariablen berechnet. Für jede Kategorie der Gruppierungsvariablen wird eine separate Schicht erstellt.

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

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

Erläuterungen der Daten für OLAP-Würfel

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

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Berichte** > **OLAP-Würfel...**
2. Wählen Sie mindestens eine stetige Auswertungsvariable aus.
3. Wählen Sie mindestens eine kategoriale Gruppierungsvariable aus.

Die folgenden Optionen sind verfügbar:

- Wählen Sie verschiedene Auswertungsstatistiken aus, indem Sie auf **Statistiken** klicken. Sie müssen mindestens eine Gruppierungsvariable auswählen, bevor Sie die Auswertungsstatistiken auswählen können.
- Berechnen Sie die Differenzen zwischen Variablenpaaren und Gruppenpaaren, die durch eine Gruppierungsvariable definiert sind, indem Sie auf **Differenzen** klicken.
- Erstellen Sie Titel für benutzerdefinierte Tabellen, indem Sie auf **Titel** klicken.
- Blenden Sie Häufigkeiten (Anzahlwerte) aus, die einen bestimmten ganzzahligen Wert unterschreiten. Ausgeblendete Werte werden als **<N** angezeigt. Dabei ist **N** die angegebene Ganzzahl. Die angegebene Ganzzahl muss größer oder gleich 2 sein.

OLAP-Würfel: Statistiken

Sie können mindestens eine der folgenden Untergruppenstatistiken für die Auswertungsvariablen innerhalb der einzelnen Kategorien jeder Gruppierungsvariablen auswählen: Summe, Anzahl der Fälle, Mittelwert, Median, gruppierter Median, Standardfehler des Mittelwerts, Minimum, Maximum, Spannweite, Variablenwert der ersten Kategorie der Gruppierungsvariablen, Variablenwert der letzten Kategorie der Gruppierungsvariablen, Standardabweichung, Varianz, Kurtosis, Standardfehler der Kurtosis, Schiefe, Standardfehler der Schiefe, Prozentsatz der gesamten Fälle, Prozentsatz der Gesamtsumme, Prozentsatz der gesamten Fälle innerhalb der Gruppierungsvariablen, Prozentsatz der Gesamtsumme innerhalb der Gruppierungsvariablen, 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 codiert wurden. Wenn z. B. für das Alter jeder Wert in den Dreißigern als 35 codiert ist, jeder Wert in den Vierzigern als 45 usw., dann wird der gruppierte Median aus den codierten Daten berechnet.

Harmonisches Mittel. Wird verwendet, um die durchschnittliche 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 reziproken Werte der Stichprobengrößen.

Kurtosis. Ein Maß dafür, wie sehr die Beobachtungen um einen zentralen Punkt gruppiert sind. 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 Extremwerten 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 (zentrale Tendenz). 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 N in. Prozentsatz der Anzahl der Fälle für die angegebene Gruppierungsvariable in den Kategorien der anderen Gruppierungsvariablen. Wenn nur eine Gruppierungsvariable vorhanden ist, ist dieser Wert gleich dem Prozentsatz der Gesamtanzahl von Fällen.

Prozent der Summe in. Prozentsatz der Summe für die angegebene Gruppierungsvariable in den Kategorien der anderen Gruppierungsvariablen. Wenn nur eine Gruppierungsvariable 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.

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

Schiefe. Ein Maß der Asymmetrie der 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 Schiefewert, der mehr als doppelt so groß ist wie sein Standardfehler, als Abweichung von der Symmetrie gilt.

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.

Standardfehler der Kurtosis. Der Quotient aus der Kurtosis und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) 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 des Mittelwerts. Ein Maß dafür, wie stark der Mittelwert von Stichprobe zu Stichprobe in derselben Verteilung variieren kann. 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.)

Standardfehler der Schiefe. Der Quotient aus der Schiefe und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) 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 nicht fehlenden Werten.

Varianz. Ein Maß der Streuung um den Mittelwert, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.

OLAP-Würfel: Differenzen

In diesem Dialogfeld können Sie prozentuale und arithmetische Differenzen zwischen Auswertungsvariablen oder zwischen Gruppen berechnen lassen, die durch eine Gruppierungsvariable 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 Dif-

ferenzen 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 Gruppierungsvariable 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 Gruppierungsvariable im Hauptdialogfeld auswählen, bevor Sie die Differenzen zwischen den Gruppen angeben können.

OLAP-Würfel: Titel

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

Kapitel 9. t-Tests

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 Stichproben mit paarigen Werten (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 paarigen Differenzen, den *T*-Test und ein 95-%-Konfidenzintervall.

t-Test bei einer Stichprobe. Vergleicht den Mittelwert einer Variablen mit einem bekannten oder hypothetischen 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 Behandlungsgruppe verteilt. Die Patienten in der Kontrollgruppe erhalten ein Plazebo. Die Patienten der Behandlungsgruppe 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 Behandlungsgruppe zu vergleichen. Bei jedem Patienten wird eine Messung vorgenommen, und er gehört zu jeweils einer (1) Gruppe.

Statistik. 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.

Erläuterungen der Daten für t-Tests bei unabhängigen Stichproben

Daten. Die Werte der untersuchten quantitativen Variablen müssen in einer einzelnen Spalte in der Datendatei vorliegen. Die Prozedur verwendet eine Gruppierungsvariable mit zwei Werten zur Aufteilung der Fälle in zwei Gruppen. Die Gruppierungsvariable kann numerische Werte (wie zum Beispiel 1 und 2 oder 6,25 und 12,5) oder kurze Zeichenfolgen (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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Mittelwerte vergleichen** > **t-Test bei unabhängigen Stichproben...**
2. Wählen Sie mindestens eine quantitative Testvariable. Für jede Variable wird ein separater *T*-Test berechnet.
3. Wählen Sie eine einzelne Gruppierungsvariable aus und klicken Sie dann auf **Gruppen def.**, um zwei Codes für die zu vergleichenden Gruppen anzugeben.
4. Zusätzlich können Sie auf **Optionen** klicken, um die Behandlung fehlender Daten und das Niveau des Konfidenzintervalls festzulegen.

t-Test bei unabhängigen Stichproben: Gruppen definieren

Definieren Sie bei numerischen Gruppierungsvariablen 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 nicht ganzzahlig sein (so sind beispielsweise 6,25 und 12,5 gültige Werte).
- **Trennwert.** Geben Sie eine Zahl ein, welche die Werte der Gruppierungsvariablen 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.

Bei Zeichenfolge-Gruppierungsvariablen geben Sie eine Zeichenfolge für Gruppe 1 und einen anderen Wert für Gruppe 2 ein, beispielsweise *ja* und *nein*. Fälle mit anderen Zeichenfolgen werden von der Analyse ausgeschlossen.

t-Tests 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 Stichproben mit paarigen Werten

Mit der Prozedur "t-Test bei Stichproben mit paarigen Werten" 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 Kontrolltestperson. In einer Blutdruckstudie könnten den Patienten die Kontrollpersonen nach Alter zugeordnet werden (einem 75-jährigen Patienten ein 75-jähriges Mitglied der Kontrollgruppe).

Statistik. 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.

Erläuterungen der Daten für t-Tests bei Stichproben mit paarigen Werten

Daten. Legen Sie für jeden paarigen Test zwei Variablen fest, die auf Intervallmessniveau oder Verhältnismessniveau 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 Stichproben mit paarigen Werten berechnen:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Mittelwerte vergleichen** > **t-Test bei Stichproben mit paarigen Werten...**
2. Wählen Sie mindestens ein Variablenpaar aus
3. Zusätzlich können Sie auf **Optionen** klicken, um die Behandlung fehlender Daten und das Niveau des Konfidenzintervalls festzulegen.

t-Test bei Stichproben mit paarigen Werten: 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.

Zusätzliche Funktionen beim Befehl T-TEST

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 paarigen t-Test (mit dem Unterbefehl `PAIRS`).

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

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-Score 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.

Statistik. Für jede Testvariable: Mittelwert, Standardabweichung und Standardfehler des Mittelwerts. 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.)

Erläuterungen der Daten für t-Tests bei einer Stichprobe

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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Mittelwerte vergleichen** > **t-Test bei einer Stichprobe...**
2. Wählen Sie eine oder mehrere Variablen aus, die mit demselben hypothetischen Wert verglichen werden sollen.
3. Geben Sie einen numerischen Testwert ein, mit dem jeder Stichprobenmittelwert verglichen werden soll.
4. Zusätzlich können Sie auf **Optionen** klicken, um die Behandlung fehlender Daten und das Niveau des Konfidenzintervalls festzulegen.

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

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 paarigen t-Test (mit dem Unterbefehl `PAIRS`).

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

Zusätzliche Funktionen beim Befehl T-TEST

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 paarigen t-Test (mit dem Unterbefehl `PAIRS`).

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

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.

Statistik. 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, Varianzanalysetabellen und zuverlässige Tests auf Gleichheit der Mittelwerte für jede abhängige Variable, benutzerspezifische A-priori-Kontraste, Post-hoc-Spannweitentests und Mehrfachvergleiche: Bonferroni, Sidak, Tukey-HSD-Test, 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*, multipler Spannweitentest nach Duncan, Student-Newman-Keuls (S-N-K), Tukey-*B*, Waller-Duncan, Scheffé und geringste signifikante Differenz.

Erläuterungen der Daten für Einfaktorielle ANOVA

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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Mittelwerte vergleichen** > **Einfaktorielle ANOVA...**
2. Wählen Sie eine oder mehrere abhängige Variablen aus.
3. Wählen Sie eine unabhängige Faktorvariable aus.

Einfaktorielle ANOVA: Kontraste

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

Polynomial. Damit zerlegen Sie die Quadratsummen zwischen den Gruppen in Trendkomponenten. 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 Kontrastsets festzulegen, klicken Sie auf **Weiter**. Verwenden Sie **Weiter** und **Zurück**, um zwischen den Kontrastsets 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

Sobald Sie festgestellt haben, dass es Abweichungen zwischen den Mittelwerten gibt, können Sie mit Post-hoc-Spannweitentests und paarweisen multiplen Vergleichen untersuchen, welche Mittelwerte sich unterscheiden. Spannweitentests ermitteln homogene Subsets von Mittelwerten, die nicht voneinander abweichen. Mit paarweisen Mehrfachvergleichen testen Sie die Differenz zwischen paarigen 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.

Varianzgleichheit angenommen

Der Tukey-HSD-Test, 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, Tukey-HSD-Test, 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.
- *Scheffe*. 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.
- *R-E-G-W F*. Mehrfaches Rückschrittverfahren nach Ryan-Einot-Gabriel-Welsh, basierend auf einem F-Test.
- *R-E-G-W Q*. Mehrfaches Rückschrittverfahren nach Ryan-Einot-Gabriel-Welsh, das auf der studentisierten Spannweite beruht.

- *S-N-K*. Führt alle paarweisen Vergleiche zwischen Mittelwerten unter Verwendung der studentisierten Bereichsverteilung aus. Bei gleich großen Stichproben werden auch die Mittelwertpaare innerhalb homogener Subsets 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 studentisierte Spannweitenstatistik 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*. Verwendet die studentisierte Bereichsverteilung für paarweise Vergleiche zwischen Gruppen. Der kritische Wert ist der Durchschnitt des entsprechenden Werts für den Tukey-HSD-Test und für Student-Newman-Keuls.
- *Duncan*. Bei diesem Test werden paarweise Vergleiche angestellt, deren schrittweise Reihenfolge mit der Reihenfolge identisch ist, 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 Test für Mehrfachvergleiche und ein Spannweitentest, der auf dem studentisierten Maximalmodulus beruht. Ähnelt dem Tukey-HSD-Test.
- *Gabriel*. Ein paarweiser Vergleichstest, der den studentisierten Maximalmodulus 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 Behandlungen 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** aus, um zu überprüfen, ob der Mittelwert bei allen Stufen des Faktors kleiner als der Mittelwert der Kontrollkategorie ist. Wählen Sie **> Kontrolle** aus, um zu überprüfen, ob der Mittelwert bei allen Stufen des Faktors größer als der Mittelwert der Kontrollkategorie ist.

Keine Varianzgleichheit angenommen

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

- *Tamhane-T2*. Konservativer, paarweiser Vergleichstest auf der Grundlage eines t-Tests. Dieser Test ist für ungleiche Varianzen geeignet.
- *Dunnett-T3*. Ein paarweiser Vergleichstest, der auf dem studentisierten Maximalmodulus beruht. Dieser Test ist für ungleiche Varianzen geeignet.
- *Games-Howell*. Ein manchmal ungenauer, 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** inaktivieren. (In einer aktivierten Pivot-Tabelle: **Tabelleneigenschaften** im Menü "Format".)

Einfaktorielle ANOVA: Optionen

Statistik. Wählen Sie mindestens eine der folgenden Optionen aus:

- **Deskriptive Statistiken.** 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 die Schätzung 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

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Test 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 Befehlssyntaxreferenz.

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 Interaktionen zwischen Faktoren und die Effekte einzelner Faktoren untersuchen, von denen einige zufällig sein können. Außerdem können Sie die Auswirkungen von Kovariaten und Interaktionen zwischen Kovariaten und Faktoren berücksichtigen. Bei der Regressionsanalyse werden die unabhängigen Variablen (Prädiktorvariablen) als Kovariaten angegeben.

Es können sowohl ausgewogene als auch unausgewogene Modelle getestet werden. Ein Design ist ausgewogen, 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 Schätzungen für die vorhergesagten Mittelwerte der Zellen im Modell, und mit Profilplots (Interaktionsdiagrammen) 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 - 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 Interaktion 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, Tukey-HSD-Test, 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 für alle abhängigen Variablen in allen Zellen. Levene-Test auf Homogenität der Varianzen.

Diagramme. Streubreite gegen mittleres Niveau, Residuendiagramme, Profilplots (Interaktion).

Erläuterungen der Daten für "GLM - Univariat"

Daten. Die abhängige Variable ist quantitativ. Faktoren sind kategorial. Sie können numerische Werte oder Zeichenfolgewerte 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 Residuendiagramme anzeigen lassen.

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

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Allgemeines lineares Modell** > **Univariat...**
2. Wählen Sie eine abhängige Variable aus.
3. Wählen Sie in Abhängigkeit von den Daten Variablen als feste Faktoren, Zufallsfaktoren und Kovariaten aus.
4. 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

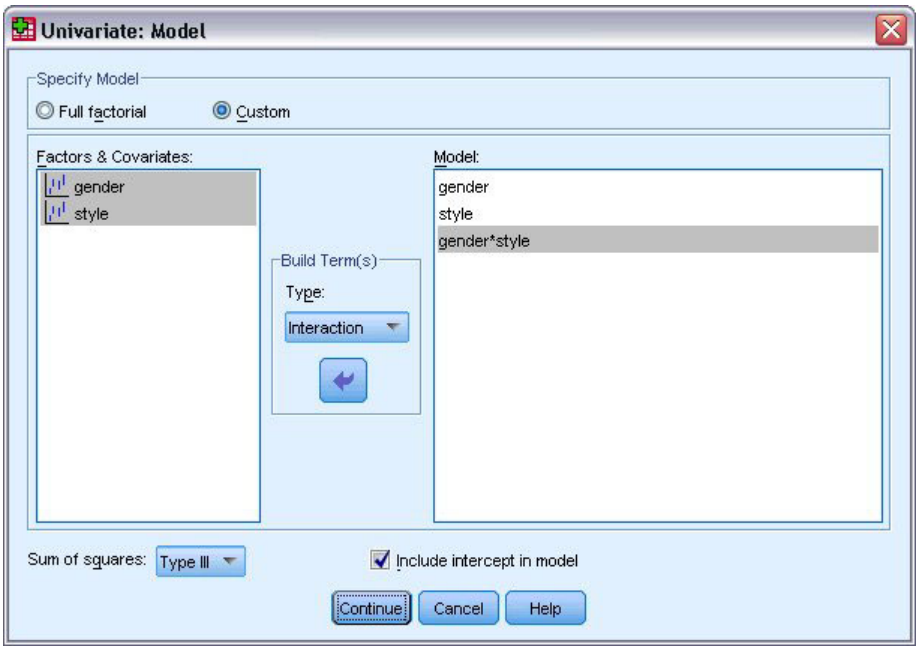


Abbildung 1. Dialogfeld "Univariat: Modell"

Modell angeben. Ein gesättigtes Modell enthält alle Haupteffekte der Faktoren, alle Kovariatenhaupteffekte und alle Interaktionen zwischen Faktoren. Es enthält keine Kovariateninteraktionen. Wählen Sie **Anpassen** aus, um nur ein Subset von Interaktionen oder Interaktionen 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 Interaktionen auswählen, die für Ihre Analyse von Interesse sind.

Quadratsumme. Hier wird die Methode zum Berechnen der Quadratsumme festgelegt. Für ausgewogene und unausgewogene 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.

Erstellen von Termen

Für die ausgewählten Faktoren und Kovariaten:

Interaktion Hiermit wird der Interaktionsterm mit der höchsten Ordnung von allen ausgewählten Variablen erstellt. Dies ist die Standardeinstellung.

Haupteffekte. Erstellt einen Haupteffektterm für jede ausgewählte Variable.

Alle 2-Wege. Hiermit werden alle möglichen Zweiwegeinteraktionen der ausgewählten Variablen erstellt.

Alle 3-Wege. Hiermit werden alle möglichen Dreiwegeinteraktionen der ausgewählten Variablen erstellt.

Alle 4-Wege. Hiermit werden alle möglichen Vierwegeinteraktionen der ausgewählten Variablen erstellt.

Alle 5-Wege. Hiermit werden alle möglichen Fünfwegeinteraktionen der ausgewählten Variablen erstellt.

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 korrigiert. Quadratsummen vom Typ I werden gewöhnlich in den folgenden Situationen verwendet:

- Ein ausgewogenes ANOVA-Modell, in dem alle Haupteffekte vor den Interaktionseffekten 1. Ordnung festgelegt werden, alle Interaktionseffekte 1. Ordnung wiederum vor den Interaktionseffekten 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 ausgewogenen 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 allen Effekten (sofern vorhanden), die den Effekt enthalten, und mit Bereinigung um alle anderen Effekte berechnet, die den Effekt nicht enthalten. Der große Vorteil der Quadrat-

summen 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 unausgewogene 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 ausgewogenen oder unausgewogenen 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 ausgewogenen oder unausgewogenen Modelle mit leeren Zellen.

GLM: 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.

GLM - Univariat. Das Testen der Hypothesen basiert auf der Nullhypothese **LB** = 0. Dabei ist **L** die Kontrastkoeffizientenmatrix 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

Profilplots (Interaktionsdiagramme) 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 erstellen. 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 Interaktion zwischen den Faktoren gibt. Das heißt, dass Sie die Faktorstufen eines einzelnen Faktors untersuchen können. Nicht parallele Linien deuten auf eine Interaktion hin.

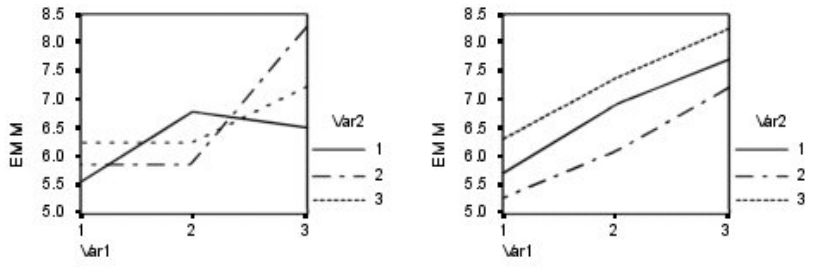


Abbildung 2. Nicht paralleles Diagramm (links) und paralleles Diagramm (rechts)

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-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 Interaktionen aus, für die Sie Schätzungen 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 Statistiken** lassen Sie beobachtete Mittelwerte, Standardabweichungen und Häufigkeiten für alle abhängigen Variablen in allen Zellen berechnen. Die Option **Schätzungen 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 Trennschärfe** liefert die Testschärfe, wenn die alternative Hypothese auf der Grundlage der beobachteten Werte aufgestellt wurde. Mit **Parameterschätzungen** werden Parameterschätzungen, Standardfehler, *T*-Tests, Konfidenzintervalle und die beobachtete Trennschärfe für jeden Test berechnet. Mit der Option **Matrixkontrastkoeffizienten** 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 Residuendiagramme sind beim Überprüfen von Annahmen über die Daten nützlich. Diese Option ist nur verfügbar, wenn Faktoren vorhanden sind. Wählen Sie **Residuendiagramm**, wenn Sie für jede abhängige Variable ein Residuendiagramm (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 Trennschärfe für die Tests verwendet. Wenn Sie ein Signifikanzniveau festlegen, wird das entsprechende Konfidenzniveau im Dialogfeld angezeigt.

Zusätzliche Funktionen beim Befehl UNIANOVA

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 benutzerdefiniert 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 Interaktionen 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 Matrixdatendatei erstellen, die Statistiken aus der Zwischensubjekt-ANOVA-Tabelle enthält.
- Mit dem Unterbefehl `OUTFILE` können Sie die Designmatrix in einer neuen Datendatei speichern.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

GLM: Post-hoc-Vergleiche

Tests für Post-hoc-Mehrfachvergleiche. Sobald Sie festgestellt haben, dass es Abweichungen zwischen den Mittelwerten gibt, können Sie mit Post-hoc-Spannweitentests 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 der Tukey-HSD-Test. 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. Der **Tukey-HSD-Test** 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 Tukey-HSD-Test leistungsfähiger als der Bonferroni-Test. Bei einer kleinen Anzahl von Paaren ist der Bonferroni-Test leistungsfähiger.

GT2 nach Hochberg ähnelt dem Tukey-HSD-Test, es wird jedoch der studentisierte Maximalmodulus verwendet. Meistens ist der Tukey-HSD-Test leistungsfähiger. Der **paarweise Vergleichstest nach Gabriel** verwendet ebenfalls den studentisierten Maximalmodulus und zeigt meistens eine größere Schärfe als das GT2 nach Hochberg, wenn die Zellengrößen ungleich sind. Der Gabriel-Test kann ungenau werden, wenn die Zellengrößen stark variieren.

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** aus, um zu überprüfen, ob der Mittelwert bei allen Stufen des Faktors kleiner als der Mittelwert der Kontrollkategorie ist. Wählen Sie **> Kontrolle** aus, 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 Subsets 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 **Tamhane-T2** (konservativer paarweiser Vergleichstest auf der Grundlage eines *T*-Tests), **Dunnett-T3** (paarweiser Vergleichstest auf der Grundlage des studentisierten Maximalmodulus), 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 Subsets 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. Der Tukey-HSD-Test, das GT2 nach Hochberg, der Gabriel-Test und der Scheffé-Test sind multiple Vergleiche, zugleich aber auch Spannweitentests.

GLM-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 Interaktionen aus, für die Sie Schätzungen 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 Statistiken** lassen Sie beobachtete Mittelwerte, Standardabweichungen und Häufigkeiten für alle abhängigen Variablen in allen Zellen berechnen. Die Option **Schätzungen 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 Trennschärfe** liefert die Testschärfe, wenn die alternative Hypothese auf der Grundlage der beobachteten Werte aufgestellt wurde. Mit **Parameterschätzungen** werden Parameterschätzungen, Standardfehler, *T*-Tests, Konfidenzintervalle und die beobachtete Trennschärfe für jeden Test berechnet. Mit der Option **Matrixkontrastkoeffizienten** 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 Residuendiagramme sind beim Überprüfen von Annahmen über die Daten nützlich. Diese Option ist nur verfügbar, wenn Faktoren vorhanden sind. Wählen Sie **Residuendiagramm**, wenn Sie für jede abhängige Variable ein Residuendiagramm (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 Trennschärfe für die Tests verwendet. Wenn Sie ein Signifikanzniveau festlegen, wird das entsprechende Konfidenzniveau im Dialogfeld angezeigt.

Zusätzliche Funktionen beim Befehl UNIANOVA

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 benutzerdefiniert 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 Interaktionen 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 Matrixdatendatei erstellen, die Statistiken aus der Zwischensubjekt-ANOVA-Tabelle enthält.
- Mit dem Unterbefehl `OUTFILE` können Sie die Designmatrix in einer neuen Datendatei speichern.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

GLM: Speichern

Vom Modell vorhergesagte Werte, Residuen und verwandte Maße können als neue Variablen im Dateneditor 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 die einzelnen Fälle vorhersagt.

- *Nicht standardisiert*. Der Wert, den das Modell für die abhängige Variable vorhersagt.
- *Gewichtet*. Gewichtete nicht standardisierte vorhergesagte Werte. Nur verfügbar, wenn zuvor eine WLS-Variable ausgewählt wurde.
- *Standardfehler*. Ein Schätzwert der Standardabweichung des Durchschnittswertes 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 den 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*. Die Differenz zwischen einem beobachteten Wert und dem durch das Modell vorhergesagten Wert.
- *Gewichtet*. Gewichtete nicht standardisierte Residuen. Nur verfügbar, wenn zuvor eine WLS-Variable ausgewählt wurde.
- *Standardisiert*. Der Quotient aus dem Residuum und einer Schätzung 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. Dies 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 Dataset in der aktuellen Sitzung oder in eine externe Datei im IBM 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

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 Interaktionen aus, für die Sie Schätzungen 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 Statistiken** lassen Sie beobachtete Mittelwerte, Standardabweichungen und Häufigkeiten für alle abhängigen Variablen in allen Zellen berechnen. Die Option **Schätzungen 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 Trennschärfe** liefert die Testschärfe, wenn die alter-

native Hypothese auf der Grundlage der beobachteten Werte aufgestellt wurde. Mit **Parameterschätzungen** werden Parameterschätzungen, Standardfehler, *T*-Tests, Konfidenzintervalle und die beobachtete Trennschärfe für jeden Test berechnet. Mit der Option **Matrixkontrastkoeffizienten** 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 Residuendiagramme sind beim Überprüfen von Annahmen über die Daten nützlich. Diese Option ist nur verfügbar, wenn Faktoren vorhanden sind. Wählen Sie **Residuendiagramm**, wenn Sie für jede abhängige Variable ein Residuendiagramm (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 Trennschärfe für die Tests verwendet. Wenn Sie ein Signifikanzniveau festlegen, wird das entsprechende Konfidenzniveau im Dialogfeld angezeigt.

Zusätzliche Funktionen beim Befehl UNIANOVA

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 benutzerdefiniert 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 Interaktionen 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 Matrixdatendatei erstellen, die Statistiken aus der Zwischensubjekt-ANOVA-Tabelle enthält.
- Mit dem Unterbefehl `OUTFILE` können Sie die Designmatrix in einer neuen Datendatei speichern.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

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.

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

Erläuterungen der Daten für bivariate Korrelationen

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...**

1. 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 es bei der Interpretation Ihrer Ergebnisse, Schlüsse über Ursache und Wirkung aufgrund signifikanter Korrelationen zu ziehen.
- **Test auf Signifikanz.** Sie können einseitige oder zweiseitige Wahrscheinlichkeiten auswählen. Wenn Ihnen die Richtung des Zusammenhangs im Voraus bekannt ist, wählen Sie **Einseitig** aus. Wählen Sie andernfalls **Zweiseitig** aus.

- **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

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 nicht fehlenden 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 einem Set 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

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Befehlssyntaxreferenz.

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.

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

Erläuterungen der Daten für partielle Korrelationen

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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Korrelation** > **Partiell...**
2. Wählen Sie mindestens zwei numerische Variablen aus, für die partielle Korrelationen berechnet werden sollen.
3. Wählen Sie mindestens eine numerische Kontrollvariable aus.

Außerdem sind folgende Optionen verfügbar:

- **Test auf Signifikanz.** Sie können einseitige oder zweiseitige Wahrscheinlichkeiten auswählen. Wenn Ihnen die Richtung des Zusammenhangs im voraus bekannt ist, wählen Sie **Einseitig**. Wählen Sie andernfalls **Zweiseitig**.
- **Tatsächliches Signifikanzniveau anzeigen.** In der Standardeinstellung werden die Wahrscheinlichkeit sowie die Freiheitsgrade für jeden Korrelationskoeffizienten angezeigt. Wenn Sie diese Option inaktivieren, 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

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 nicht fehlenden Werten angezeigt.
- **Korrelationen nullter Ordnung.** Hiermit wird eine einfache Matrix für Korrelationen zwischen allen Variablen (einschließlich Kontrollvariablen) angezeigt.

Fehlende Werte. Sie können eine der folgenden Möglichkeiten auswählen:

- **Listenweiser Fallausschluss.** Fälle mit fehlenden Werten für Variablen (einschließlich Kontrollvariablen) 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 Ausschluss wird der größtmögliche Teil der Daten verwendet. Die Anzahl der Fälle kann jedoch von Koeffizient zu Koeffizient variieren. Wenn der paarweise Ausschluss 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

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 mehreren 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 Befehlssyntaxreferenz.

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 Clusteranalyse oder der multidimensionalen Skalierung zur Analyse komplexer Datasets 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 Clusteranalyse oder einer multidimensionalen Skalierung auf die Ähnlichkeiten können Sie eine formale Analyse durchführen, um die zugrunde liegende Struktur zu untersuchen.

Statistik. Unähnlichkeitsmaße (Distanzmaße) für Intervalldaten: Euklidische Distanz, quadrierte euklidische Distanz, Tschebyscheff, Block, Minkowski oder ein benutzerdefiniertes Maß; für Häufigkeiten: Chi-Quadrat-Maß oder Phi-Quadrat-Maß; für Binärdaten: Euklidische Distanz, quadrierte euklidische Distanz, 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, Würfelähnlichkeitsmaß, 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Korrelation** > **Distanzen...**
2. 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.
3. Wählen Sie im Gruppenfeld "Distanzen berechnen" eine andere Option aus, um Ähnlichkeiten zwischen Fällen oder Variablen zu berechnen.

Unähnlichkeitsmaße für Distanzen

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.** Euklidische Distanz, quadrierte euklidische Distanz, Tschebyscheff, Block, Minkowski oder ein benutzerdefiniertes Maß.
- **Häufigkeiten.** Chi-Quadrat-Maß oder Phi-Quadrat-Maß.
- **Binär.** Euklidische Distanz, quadrierte euklidische Distanz, 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" oder "Standardabweichung 1".

Im Gruppenfeld "Maße transformieren" können Sie festlegen, ob die durch das Distanzmaß generierten 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

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, Würfelähnlichkeitsmaß, 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 generierten 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. Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Befehlssyntaxreferenz.

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 Dataset 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. Weitere Informationen finden Sie im Thema .

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). Kategoriale Felder (nominal und ordinal) werden als Faktoren im Modell verwendet und stetige Felder werden als Kovariaten verwendet.

Hinweis: Wenn ein kategoriales Feld mehr als 1.000 Kategorien enthält, wird diese Prozedur nicht ausgeführt und es wird kein Modell erstellt.

Erstellen eines lineares Modells

Für diese Funktion ist die Option "Statistics Base" erforderlich.

Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren > **Regression** > **Automatische lineare Modelle...**

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

Ziele

Wie lautet Ihr Hauptziel? Wählen Sie das entsprechende Ziel aus.

- **Standardmodell erstellen.** Bei dieser Methode wird ein einzelnes Modell erstellt, um das Ziel mithilfe der Prädiktoren vorherzusagen. Allgemein gesagt, sind Standardmodelle einfacher zu interpretieren und schneller zu bewerten als Boosting-Dataset-Ensembles, Bagging-Dataset-Ensembles oder große Dataset-Ensembles.
- **Modellgenauigkeit verbessern (Boosting).** Bei dieser Methode wird ein Ensemble mithilfe von 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.

  Durch Boosting wird eine Reihe von "Komponentenmodellen" erstellt, wobei jede einzelne Komponente auf dem gesamten Dataset beruht. Vor dem Erstellen der einzelnen aufeinander folgenden Komponen-

tenmodells werden die Datensätze jeweils basierend auf den Residuen des vorangegangenen Komponentenmodells gewichtet. Fälle mit großen Residuen erhalten eine höhere Analysegewichtung, sodass beim nächsten Komponentenmodell das Augenmerk auf einer hochwertigen Vorhersage dieser Datensätze liegt. Zusammen bilden diese Komponentenmodelle ein Ensemblemodell. Das Ensemblemodell bewertet neue Datensätze mithilfe einer Kombinierungsregel. Die verfügbaren Regeln hängen vom Messniveau des Ziels ab.

- **Modellstabilität verbessern (Bagging).** Bei dieser Methode wird ein Ensemblemodell mithilfe von Bagging (Bootstrap-Aggregation) erstellt. Dabei werden mehrere Modelle generiert, um zuverlässigere Vorhersagen zu erhalten. Bei Ensembles können Erstellung und Bewertung länger dauern als bei Standardmodellen.

  Bei der Bootstrap-Aggregation (Bagging) werden Reproduktionen des Trainings-Datasets erstellt, indem aus dem ursprünglichen Dataset Stichproben mit Zurücklegen gezogen werden. Dadurch ergeben sich Bootstrap-Stichproben mit der gleichen Größe wie beim ursprünglichen Dataset. Anschließend wird von jeder Reproduktion ein "Komponentenmodell" erstellt. Zusammen bilden diese Komponentenmodelle ein Ensemblemodell. Das Ensemblemodell bewertet neue Datensätze mithilfe einer Kombinierungsregel. Die verfügbaren Regeln hängen vom Messniveau des Ziels ab.
- **Modell für extrem große Datasets erstellen (IBM SPSS Statistics Server erforderlich).** Bei dieser Methode wird ein Ensemblemodell durch Aufteilung des Datasets in separate Datenblöcke erstellt. Verwenden Sie diese Option, wenn Ihr Dataset zu groß ist, um eines der oben genannten Modelle zu erstellen, oder um inkrementelle Modellerstellung durchzuführen. Bei dieser Option kann die Erstellung weniger zeitaufwändig sein, die Bewertung kann jedoch länger dauern als bei Standardmodellen. Für diese Option ist IBM SPSS Statistics Server-Konnektivität erforderlich.

Informationen zu Boosting, Bagging und sehr umfangreichen Datasets finden Sie in „Ensembles” auf Seite 64.

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 für das Scoring 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 transformiert, 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.
- **Ausreißerbehandlung.** Werte stetiger Prädiktoren, die über einem Trennwert liegen (drei Standardabweichungen vom Mittelwert), werden auf den Trennwert 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 ursprünglichen und abgeleiteten Versionen des Felds aus dem Modell ausgeschlossen, da sie als Prädiktoren keinen Wert haben.

Konfidenzniveau. Das 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

Modellauswahlmethode. Wählen Sie eine der Modellauswahlmethoden (Details unten) oder **Alle Prädiktoren einschließen** aus, wodurch einfach alle verfügbaren Prädiktoren als Haupteffektmodellterme 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 Trainingssets 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 Trainingssets und wird zur Penalisierung übermäßig komplexer Modelle angepasst. **Das Kriterium zur Verhinderung übermäßiger Anpassung (ASE)** basiert auf der Anpassungsgüte (durchschnittlicher quadratischer Fehler, 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 Datasets, 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 von 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 von Effekten beendet, stoppt der Algorithmus beim aktuellen Effektset.
- **Anpassen der maximalen Anzahl Schritte.** Der schrittweise Algorithmus stoppt nach einer bestimmten Anzahl von Schritten. Standardmäßig ist das dreimal die Anzahl der verfügbaren Effekte. Alternativ kann eine positive Ganzzahl als maximale Anzahl von Schritten angegeben werden.

Auswahl "Beste Subsets". 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 Subsets" auszuwählen. Das **Informationskriterium (AICC)** basiert auf der Wahrscheinlichkeit des Trainingssets für das Modell und wird zur Penalisierung übermäßig komplexer Modelle angepasst. **Korrigiertes R-Quadrat** beruht auf der Anpassungsgüte des Trainingssets und wird zur Penalisierung übermäßig komplexer Modelle angepasst. **Das Kriterium zur Verhinderung übermäßiger Anpassung (ASE)** basiert auf der Anpassungsgüte (durchschnittlicher quadratischer Fehler, 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 Datasets, die nicht zum Trainieren des Modells verwendet wird.

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

Hinweis: Die Auswahl "Beste Subsets" ist rechenintensiver als die Auswahl "Schrittweise vorwärts". Wenn "Beste Subsets" zusammen mit "Boosting", "Bagging" oder "Sehr große Datasets" verwendet wird, kann das Erstellen deutlich länger dauern als das Erstellen eines Standardmodells mithilfe der Auswahl "Schrittweise vorwärts".

Ensembles

Diese Einstellungen legen das Verhalten der Ensemble-Bildung fest, die erfolgt, wenn auf der Registerkarte "Ziele" die Option "Boosting", "Bagging" oder "Sehr große Datasets" ausgewählt ist. Optionen, die für das ausgewählte Ziel nicht gelten, werden ignoriert.

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

- **Standardkombinierungsregel 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. Beim Boosting wird für das Scoring der kategorialen Ziele stets eine gewichtete Mehrheit verwendet und für das Scoring stetiger Ziele ein gewichteter Median.

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

Erweitert

Ergebnisse reproduzieren. Durch Einstellen eines Startwerts für Zufallszahlen 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) erstellt. Der Standardwert ist 54752075.

Modelloptionen

Speichert vorhergesagte Werte im Dataset. Der Standardvariablenname lautet *PredictedValue*.

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

Modellübersicht

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, der auf der Registerkarte Felder festgelegt ist,
- 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 im Format "größer ist besser" dargestellt wird. Der Wert ist 100 × der eingestellten R^2 für das endgültige Modell.

Automatische Datenaufbereitung

Diese Ansicht zeigt Informationen darüber an, welche Felder ausgeschlossen wurden und wie transformierte Felder im Schritt "automatische Datenaufbereitung" (ADP - Automatic Data Preparation) 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:

- **Dauer ableiten: Monate** berechnet die verstrichene Zeit in Monaten zwischen den Werten in einem Feld mit Datumsangaben und dem aktuellen Systemdatum.
- **Dauer ableiten: Stunden**berechnet die verstrichene Zeit in Stunden zwischen den Werten in einem Feld mit Zeitangaben und der aktuellen Systemzeit.
- **Messniveau von stetig auf ordinal ändern** wandelt stetige Felder mit weniger als fünf eindeutigen Werten in ordinale Felder um.
- **Messniveau von ordinal auf stetig ändern** wandelt ordinale Felder mit mehr als zehn eindeutigen Werten in stetige Felder um.
- **Ausreißer trimmen** Werte stetiger Prädiktoren, die über einem Trennwert liegen (drei Standardabweichungen vom Mittelwert), werden auf den Trennwert 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 Ausreißerbehandlung/nach der Zusammenführung von Kategorien** entfernt Prädiktoren, die einen einzelnen Wert aufweisen, möglicherweise nachdem andere ADP-Aktionen ausgeführt wurden.

Prädiktoreinfluss

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

Vorhersage nach Beobachtung

Diese Ansicht zeigt ein klassiertes 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

Diese Ansicht zeigt ein Diagnosediagramm der Modellresiduen.

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

- **Histogramm.** Diese Ansicht zeigt ein klassiertes 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 Wahrscheinlichkeit-Wahrscheinlichkeit-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

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 eine niedrigere Gewichtung gegeben werden kann, ob die extremen Werte auf einen akzeptablen Schwellenwert verringert werden können oder ob die einflussreichen Datensätze vollständig entfernt werden sollen.

Effekte

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.

Prädiktoreinfluss. Für den Prädiktoreinfluss 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 Signifikanzschieberegler 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, sodass keine Effekte basierend auf der Signifikanz herausgefiltert werden.

Koeffizienten

Diese Ansicht zeigt den Wert der einzelnen Koeffizienten im Modell. Beachten Sie, dass Faktoren (kategoriale Prädiktoren) innerhalb des Modells indikatorcodiert sind, 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 Wertbeschriftungen 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.

Prädiktoreinfluss. Für den Prädiktoreinfluss 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 Signifikanzschieberegler 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, sodass keine Koeffizienten basierend auf der Signifikanz herausgefiltert werden.

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.

Hinweis: Wenn keine Prädiktoren signifikant sind, werden keine geschätzten Mittel produziert.

Modellerstellungsübersicht

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 Subsets. Wenn der Auswahlalgorithmus "Beste Subsets" 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 Basketballteams und der pro Spiel erzielten mittleren Punktzahl des Teams? Einem Streudiagramm lässt sich entnehmen, dass zwischen diesen Variablen eine lineare Beziehung besteht. Die Anzahl gewonnener Spiele und die erzielte Punktzahl des Gegners stehen gleichfalls in linearer Beziehung zueinander. Diese Variablen enthalten eine negative Beziehung. Einer steigenden Anzahl gewonnener Spiele steht eine fallende mittlere Punktzahl des Gegners gegenüber. Mit der linearen Regression können Sie die Beziehung dieser Variablen modellieren. Mit einem geeigneten Modell lassen sich Spielgewinne von Teams vorhersagen.

Statistik. Für jede Variable: Anzahl gültiger Fälle, Mittelwert und Standardabweichung. 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 Diagnoseinformationen. Diagramme: Streudiagramme, partielle Diagramme, Histogramme und Normalverteilungsdiagramme.

Erläuterungen der Daten für die lineare Regression

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 Kontrastvariablen umcodiert 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Regression** > **Linear...**
2. Wählen Sie im Dialogfeld "Lineare Regression" eine numerische abhängige Variable aus.
3. Wählen Sie eine oder mehrere numerische unabhängige Variablen aus.

Die folgenden Optionen sind verfügbar:

- Fassen Sie unabhängige Variablen in Blöcken zusammen und geben Sie verschiedene Einschlussmethoden für unterschiedliche Subsets von Variablen an.
- Wählen Sie eine Auswahlvariable aus, um die Analyse auf ein Subset von Fällen mit einem bestimmten Wert oder bestimmten Werten für diese Variable zu begrenzen.
- Wählen Sie eine Variable zur Fallunterscheidung aus 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 Set 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.
- *Vorwärtsauswahl*. 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 das Toleranzniveau 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 Subsets 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: Regel definieren

Die durch die Auswahlregel 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 Zeichenfolgewert ist ebenfalls möglich.

Lineare Regression: Diagramme

Diagramme können beim Validieren der Annahmen von Normalverteilung, Linearität und Varianzgleichheit hilfreich sein. Diagramme dienen auch zum Auffinden von Ausreißern, ungewöhnlichen Beobachtungen und Einflussfällen. Nachdem sie als neue Variablen gespeichert wurden, stehen im Dateneditor vorhergesagte Werte, Residuen und andere Diagnoseinformationen zum Erstellen von Diagrammen mit den unabhängigen Variablen zur Verfügung. Folgende Diagramme sind verfügbar:

Streudiagramme. Sie können zwei beliebige der folgenden Elemente darstellen: 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 Varianzgleichheit zu überprüfen.

Liste der Quellenvariablen.. Listet die abhängige Variable (DEPENDNT) und die folgenden vorhergesagten Variablen und Residuenvariablen auf: standardisierte vorhergesagte Werte (*ZPRED), standardisierte Residuen (*ZRESID), gelöschte Residuen (*DRESID), angepasste vorhergesagte Werte (*ADJPRED), studentisierte Residuen (*SRESID) und studentisierte gelöschte Residuen (*SDRESID).

Alle partiellen Diagramme erzeugen. Zeugt Streudiagramme der Residuen aller unabhängigen Variablen und der Residuen der abhängigen Variablen an, 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: Neue Variablen 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 die einzelnen Fälle 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 dieselben Werte für die unabhängigen Variablen haben.

Distanzen. Dies sind Maße zum Auffinden von Fällen mit ungewöhnlichen Wertekombinationen bei den 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 Falls vom Mittelwert aller Fälle abweichen. Eine große Mahalanobis-Distanz charakterisiert einen Fall, der bei einer oder mehreren unabhängigen Variablen Extremwerte besitzt.

- *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 Obergrenzen (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 einer Schätzung 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. Dies ist die Differenz zwischen dem Wert der abhängigen Variablen und dem korrigierten Schätzwert.
- *Studentisiert, ausgeschlossen*. 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*. 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.
- *DfFit*. Die Differenz im Anpassungswert ist 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 Falls von der Berechnung der 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 Dataset oder in einer Datendatei. Datasets 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 Datasets müssen den Regeln zum Benennen von Variablen entsprechen.

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

Lineare Regression: Statistiken

Folgende Statistiken sind verfügbar:

Regressionskoeffizienten. Schätzungen zeigt den Regressionskoeffizienten *B*, den Standardfehler von *B*, das Beta des standardisierten Koeffizienten, den *t*-Wert für *B* und das zweiseitige Signifikanzniveau von *t* an. **Konfidenzintervalle** zeigt 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 Schätzung und eine Tabelle für die 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 ein aussagekräftiger Prädiktor 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, die vorliegt, wenn 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 fallweisen Diagnoseinformationen für die Fälle angezeigt, die das Auswahlkriterium (Ausreißer über *n* Standardabweichungen) erfüllen.

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 als der Aufnahmewert ist. Sie wird ausgeschlossen, wenn das Signifikanzniveau größer als der Ausschlusswert ist. 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 als der Aufnahmewert ist. Sie wird ausgeschlossen, wenn der F-Wert kleiner als der Ausschlusswert ist. 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 inaktiviert 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

Die Befehlssyntax ermöglicht außerdem 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`)

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

Kapitel 17. Ordinale Regression

Die ordinale Regression ermöglicht es, die Abhängigkeit einer polytomen ordinalen Antwortvariablen von einem Set von Prädiktoren 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 Prädiktorvariablen (unabhängig). Die geschätzten Koeffizienten geben die Auswirkung einer Änderung in den Prädiktoren 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 Kovariatenstruktur, asymptotische Korrelations- und Kovarianzmatrizen der Parameterschätzungen, Pearson-Chi-Quadrat und Likelihood-Quotienten-Chi-Quadrat, Statistik der Anpassungsgüte, Iterationsverlauf, Test der Annahme von parallelen Linien, Parameterschätzungen, Standardfehler, Konfidenzintervalle sowie R^2 nach Cox und Snell, Nagelkerke und McFadden.

Erläuterungen der Daten für die ordinale Regression

Daten. Es wird angenommen, dass die abhängige Variable ordinal ist. Sie kann eine numerische oder eine Zeichenfolgevariable 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 Kovariatenvariablen müssen numerisch sein. Beachten Sie, dass die Verwendung von mehreren stetigen Kovariaten leicht zu einer sehr umfangreichen Tabelle mit Zellenwahrscheinlichkeiten führen kann.

Annahmen. Es darf nur eine Antwortvariable 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

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren > **Regression** > **Ordinal...**

2. Wählen Sie eine abhängige Variable aus.
3. 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 Parameterschätzungen bestimmen und eine Verknüpfungsfunktion auswählen.

Iterationen. Sie können den Iterationsprozess anpassen.

- **Maximalzahl der Iterationen.** Geben Sie eine nicht negative Ganzzahl an. Beim Wert 0 gibt die Prozedur die anfänglichen Schätzwerte zurück.
- **Maximalzahl für Schritthalbierung.** 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.
- **Parameterkonvergenz.** Der Prozess wird beendet, wenn die absolute oder relative Änderung in jedem der Parameterschätzungen 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.

Delta. Der 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 Prädiktoren 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. Die folgenden fünf Verknüpfungsfunktionen sind verfügbar:

- **Logit.** $f(x) = \log(x/(1-x))$. Wird üblicherweise für gleichmäßig verteilte Kategorien verwendet.
- **Log-Log komplementär.** $f(x) = \log(-\log(1-x))$. Wird üblicherweise verwendet, wenn höhere Kategorien wahrscheinlicher sind.
- **Log-Log negativ.** $f(x) = -\log(-\log(x))$. Wird üblicherweise verwendet, wenn niedrigere Kategorien wahrscheinlicher sind.
- **Probit.** $f(x) = \Phi^{-1}(x)$. Wird üblicherweise verwendet, wenn die latente Variable normalverteilt ist.
- **Cauchit (Inverse von Cauchy).** $f(x) = \tan(\pi(x-0.5))$. Wird üblicherweise verwendet, wenn die latente Variable viele Extremwerte aufweist.

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.

Anzeigen. Es werden die folgenden Tabellen erstellt:

- **Iterationsverlauf ausgeben.** Die Log-Likelihood und die Parameterschätzungen 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.
- **Parameterschätzungen.** Parameterschätzungen, Standardfehler und Konfidenzintervalle.
- **Asymptotische Korrelation der Parameterschätzungen.** Matrix der Parameterschätzungskorrelationen.
- **Asymptotische Kovarianz der Parameterschätzungen.** Matrix der Parameterschätzungskovarianzen.
- **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 Kovariatenstruktur. Bedenken Sie, dass diese Option bei Modellen mit vielen Kovariatenstrukturen (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 Antwortkategorien.
- **Vorhergesagte Kategorie.** Die Antwortkategorie mit der größten geschätzten Wahrscheinlichkeit für eine Faktor-/Kovariaten-Struktur.
- **Vorhergesagte Kategoriewahrscheinlichkeit.** 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 Kategoriewahrscheinlichkeit.** 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.

Modell bestimmen. Ein Modell mit Haupteffekten enthält die Haupteffekte der Faktoren und Kovariaten, aber keine Interaktionseffekte. Sie können ein benutzerdefiniertes Modell erstellen, um Subsets von Interaktionen 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 Interaktionseffekten.

Erstellen von Termen

Für die ausgewählten Faktoren und Kovariaten:

Interaktion Hiermit wird der Interaktionsterm mit der höchsten Ordnung von allen ausgewählten Variablen erstellt. Dies ist die Standardeinstellung.

Haupteffekte. Erstellt einen Haupteffektterm für jede ausgewählte Variable.

Alle 2-Wege. Hiermit werden alle möglichen Zweiwegeinteraktionen der ausgewählten Variablen erstellt.

Alle 3-Wege. Hiermit werden alle möglichen Dreiwegeinteraktionen der ausgewählten Variablen erstellt.

Alle 4-Wege. Hiermit werden alle möglichen Vierwegeinteraktionen der ausgewählten Variablen erstellt.

Alle 5-Wege. Hiermit werden alle möglichen Fünfwegeinteraktionen der ausgewählten Variablen erstellt.

Ordinale Regression: Skala

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

Faktoren/Kovariaten. Die Faktoren und Kovariaten werden aufgelistet.

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

Erstellen von Termen

Für die ausgewählten Faktoren und Kovariaten:

Interaktion Hiermit wird der Interaktionsterm mit der höchsten Ordnung von allen ausgewählten Variablen erstellt. Dies ist die Standardeinstellung.

Haupteffekte. Erstellt einen Haupteffektterm für jede ausgewählte Variable.

Alle 2-Wege. Hiermit werden alle möglichen Zweiwegeinteraktionen der ausgewählten Variablen erstellt.

Alle 3-Wege. Hiermit werden alle möglichen Dreiwegeinteraktionen der ausgewählten Variablen erstellt.

Alle 4-Wege. Hiermit werden alle möglichen Vierwegeinteraktionen der ausgewählten Variablen erstellt.

Alle 5-Wege. Hiermit werden alle möglichen Fünfwegeinteraktionen der ausgewählten Variablen erstellt.

Zusätzliche Funktionen beim Befehl PLUM

Sie können die ordinale Regression an Ihre Bedürfnisse anpassen, wenn Sie ihre Auswahl in ein Syntaxfenster einfügen und die resultierende Befehlssyntax für den Befehl `PLUM` bearbeiten. Die Befehlssyntax ermöglicht außerdem Folgendes:

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

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

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-Service-Provider verfolgt den Prozentsatz des mit Viren infizierten E-Mail-Verkehrs über die Netze im Lauf der Zeit. Ein Streudiagramm zeigt, dass eine nicht lineare 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.

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

Erläuterungen der Daten für die Kurvenanpassung

Daten. Die abhängigen und die unabhängigen Variablen müssen quantitativ sein. Wenn Sie aus dem aktiven Dataset **Zeit** als unabhängige Variable ausgewählt haben (statt eine Variable auszuwählen), generiert 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) ein Set 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 sein: 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Regression** > **Kurvenanpassung...**
2. Wählen Sie eine oder mehrere abhängige Variablen aus. Für jede abhängige Variable wird ein separates Modell erstellt.
3. Wählen Sie eine unabhängige Variable aus (wählen Sie entweder eine Variable aus dem aktiven Dataset oder wählen Sie **Zeit** aus).
4. Die folgenden Optionen sind verfügbar:

- Wählen Sie eine Variable zum Beschriften der Fälle in Streudiagrammen aus. 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. 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. 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. Ein Modell mit folgender Gleichung: Y = b0 + (b1 * t) + (b2 * t**2) + (b3 * t**3).

Potenzfunktion. 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 wie folgt lautet: Y = e**(b0 + (b1/t)) oder ln(Y) = b0 + (b1/t).

Logistisch. 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 der obere Grenzwert 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.

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

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

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 Beschriftungen in einer Tabelle im Ausgabefenster angezeigt.

Fälle vorhersagen. Wenn Sie im aktiven Dataset 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 auswählen:

- **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 zum Schätzen 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 zurzeit 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 Strukturgleichungsmodellierung darstellt und besonders nützlich ist, wenn die Prädiktorvariablen eine hohe Korrelation aufweisen oder wenn die Anzahl der Prädiktoren 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.

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 - Variable Importance in Projection) und die Schätzungen für Regressionsparameter (nach abhängiger Variablen) werden jeweils standardmäßig angegeben.

Diagramme. Die Bedeutung der Variablen in der Projektion, Faktorscores, Faktorgewichtungen für die ersten drei latenten Faktoren und die Distanz zum Modell werden jeweils über die Registerkarte Optionen erstellt.

Erläuterungen der Daten für die Regression mit partiellen kleinsten Quadraten

Messniveau. Die abhängigen und unabhängigen Variablen (Prädiktorvariablen) 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 Quellenvariablen und wählen Sie das gewünschte Messniveau im Popup-Menü aus. Kategoriale (nominale bzw. ordinale) Variablen werden von der Prozedur als äquivalent behandelt.

Codierung für kategoriale Variablen. Die Prozedur codiert vorübergehend für die Dauer ihrer Ausführung kategoriale abhängige Variablen mithilfe der "Eins-aus-*c*"-Codierung 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 Gewichtungen oder Gewichtungen 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...**

1. Wählen Sie mindestens eine abhängige Variable aus.
2. Wählen Sie mindestens eine unabhängige Variable aus.

Die folgenden Optionen sind verfügbar:

- Geben Sie eine Referenzkategorie für kategoriale (nominale bzw. ordinale) abhängige Variablen an.
- Geben Sie eine Variable an, die als eindeutige Kennung für die fallweise Ausgabe und für die gespeicherten Datasets verwendet werden soll.
- Geben Sie eine Obergrenze für die Anzahl der zu extrahierenden latenten Faktoren an.

Voraussetzungen

Die Prozedur der Regression mit partiellen kleinsten Quadraten ist ein Python-Erweiterungsbefehl und erfordert IBM SPSS Statistics - Essentials for Python, das standardmäßig mit Ihrem IBM SPSS Statistics-Produkt installiert wird. Die Regression mit partiellen kleinsten Quadraten erfordert darüber hinaus die frei erhältlichen Python-Bibliotheken NumPy und SciPy.

Anmerkung: Für Benutzer, die im verteilten Analysemodus arbeiten (erfordert IBM SPSS Statistics Server), müssen NumPy und SciPy auf dem Server installiert sein. Bitten Sie Ihren Systemadministrator um Unterstützung.

Windows- und Mac-Benutzer

Unter Windows und Mac müssen NumPy und SciPy in einer anderen Instanz von Python 2.7 installiert werden als in der, die mit IBM SPSS Statistics installiert ist. Wenn Sie nicht über eine separate Instanz von Python 2.7 verfügen, können Sie diese von http://www.python.org herunterladen. Installieren Sie anschließend NumPy und SciPy für Python Version 2.7. Die Installationsprogramme sind unter http://www.scipy.org/Download verfügbar.

Zum Aktivieren der Verwendung von NumPy und SciPy müssen Sie Ihren Speicherort für Python auf die Instanz von Python 2.7 setzen, in der Sie NumPy und SciPy installiert haben. Der Speicherort für Python wird auf der Registerkarte "Dateispeicherorte" im Dialogfeld "Optionen" (Bearbeiten> Optionen) festgelegt.

Linux-Benutzer

Es wird empfohlen, die Quelle herunterzuladen und NumPy und SciPy selbst zu erstellen. Die Quelle ist unter http://www.scipy.org/Download verfügbar. Sie können NumPy und SciPy in der Instanz von Python 2.7 installieren, die mit IBM SPSS Statistics installiert ist. Diese befindet sich im Verzeichnis `Python` an dem Speicherort, an dem IBM SPSS Statistics installiert ist.

Wenn Sie NumPy und SciPy in einer anderen Instanz von Python 2.7 installieren wollen als in der, die mit IBM SPSS Statistics installiert ist, müssen Sie Ihren Speicherort für Python so festlegen, dass er auf diese Instanz verweist. Der Speicherort für Python wird auf der Registerkarte "Dateispeicherorte" im Dialogfeld "Optionen" (Bearbeiten> Optionen) festgelegt.

Windows- und UNIX-Server

NumPy und SciPy müssen auf dem Server in einer anderen Version von Python 2.7 installiert werden als in der, die mit IBM SPSS Statistics installiert ist. Wenn auf dem Server keine separate Version von Python 2.7 vorhanden ist, kann sie von http://www.python.org heruntergeladen werden. NumPy und SciPy für Python 2.7 sind unter http://www.scipy.org/Download verfügbar. Zum Aktivieren der Verwendung von NumPy und SciPy muss der Speicherort für Python auf die Version von Python 2.7 gesetzt werden, in der NumPy und SciPy installiert sind. Der Speicherort für Python wird über IBM SPSS Statistics Administration Console gesetzt.

Modell

Modelleffekte 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 Interaktionen auswählen, die für Ihre Analyse von Interesse sind.

Erstellen von Termen

Für die ausgewählten Faktoren und Kovariaten:

Interaktion Hiermit wird der Interaktionsterm mit der höchsten Ordnung von allen ausgewählten Variablen erstellt. Dies ist die Standardeinstellung.

Haupteffekte. Erstellt einen Haupteffektterm für jede ausgewählte Variable.

Alle 2-Wege. Hiermit werden alle möglichen Zweiwegeinteraktionen der ausgewählten Variablen erzeugt.

Alle 3-Wege. Hiermit werden alle möglichen Dreiwegeinteraktionen der ausgewählten Variablen erzeugt.

Alle 4-Wege. Hiermit werden alle möglichen Vierwegeinteraktionen der ausgewählten Variablen erstellt.

Alle 5-Wege. Hiermit werden alle möglichen Fünfwegeinteraktionen der ausgewählten Variablen erzeugt.

Optionen

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

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

- **Schätzungen für einzelne Fälle speichern.** Speichert die folgenden fallweisen Modellschätzungen: vorhergesagte Werte, Residuen, Distanz zum Modell mit latenten Faktoren und Scores für latente Faktoren. Außerdem werden die Scores für latente Faktoren grafisch dargestellt.
- **Schätzungen für latente Faktoren speichern.** Speichert die Ladungen und Gewichtungen latenter Faktoren. Außerdem werden die Gewichtungen für latente Faktoren grafisch dargestellt.
- **Schätzungen für unabhängige Variablen speichern.** Speichert Schätzungen 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. Nächste-Nachbarn-Analyse

Die Nächste-Nachbarn-Analyse ist eine Methode für die Klassifikation 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 Klassifikationen der ähnlichsten Fälle – der nächstgelegenen Nachbarn – werden ermittelt und der neue Fall wird in die Kategorie eingeordnet, die die größte Anzahl nächstgelegener Nachbarn aufweist.

Sie können die Anzahl der nächstgelegenen Nachbarn angeben, die untersucht werden sollen; dieser Wert wird als *k* bezeichnet.

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

Erläuterungen der Daten für die Nächste-Nachbarn-Analyse

Ziel und Merkmale. Folgende Ziele und Merkmale sind möglich:

- *Nominal*. Eine Variable kann als nominal behandelt werden, wenn ihre Werte Kategorien darstellen, die 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.
- *Skala*. 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 Quellenvariablen und wählen Sie das gewünschte Messniveau im Popup-Menü aus.

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

Tabelle 1. Messniveausymbole

	Numerisch	**Zeichenfolge**	**Datum**	**Zeit**
Metrisch (stetig)		entfällt		
Ordinal				

Tabelle 1. Messniveausymbole (Forts.)

	Numerisch	Zeichenfolge	Datum	Zeit
Nominal				

Codierung für kategoriale Variablen. Die Prozedur codiert vorübergehend für die Dauer ihrer Ausführung kategoriale Prädiktoren und abhängige Variablen mithilfe der "Eins-aus-*c*"-Codierung 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 Merkmalbereichs. Die Gesamtanzahl der Dimensionen ist die Anzahl der an metrischen Prädiktoren plus die Anzahl der Kategorien in allen kategorialen Prädiktoren. Daher kann das Training durch dieses Codierungsschema verlangsamt werden. Wenn das Training der nächstgelegenen Nachbarn sehr langsam vorangeht, können Sie versuchen, die Anzahl der Kategorien der kategorialen Prädiktoren zu verringern, indem Sie ähnliche Kategorien zusammenfassen oder Fälle ausschließen, die extrem seltene Kategorien aufweisen, bevor Sie die Prozedur ausführen.

Jede "Eins-aus-*c*"-Codierung beruht auf den Trainingsdaten, selbst wenn eine Holdout-Stichprobe definiert wurde (siehe „Partitionen" auf Seite 90). Wenn die Holdout-Stichprobe daher Fälle mit Prädiktorkategorien 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. Jede Neuskalierung beruht auf den Trainingsdaten, selbst wenn eine Holdout-Stichprobe definiert wurde (siehe „Partitionen" auf Seite 90). 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 Kreuzvalidierungsaufteilungen. Wenn Sie Ihre Ergebnisse exakt reproduzieren wollen, müssen Sie nicht nur dieselben Einstellungen für die Prozedur, sondern auch einen Startwert für den Mersenne-Twister festlegen (siehe „Partitionen" auf Seite 90) oder Variablen für die Definition von Partitionen und Kreuzvalidierungsaufteilungen verwenden.

So definieren Sie die Analyse der nächstgelegenen Nachbarn:

Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren > **Klassifizieren** > **Nächstgelegener Nachbar...**

1. Geben Sie ein oder zwei Funktionen an, die als unabhängige Variablen oder Prädiktoren 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 Klassifikation oder Vorhersage 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)], 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 ermitteln, welche Testscores aus einem Schulbezirk – dem Fokusfall

– vergleichbar sind mit denen aus ähnlichen Schulbezirken. Er verwendet die Nächste-Nachbarn-Analyse, um die Schulbezirke zu finden, die sich hinsichtlich einer festgelegten Menge an Merkmalen am ähnlichsten sind. Anschließend vergleicht er die Testscores 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 Merkmalbereichsdiagramm, im Peerdiagramm 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 Merkmalbereichsdiagramm, im Peerdiagramm und in der Quadrantenkarte mit diesen Werten beschriftet.

Felder mit unbekanntem Messniveau

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

Daten durchsuchen. Liest die Daten im aktiven Dataset und weist allen Feldern, deren Messniveau zurzeit nicht bekannt ist, das Standardmessniveau zu. Bei großen Datasets 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 Dateneditors 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

Anzahl der nächstgelegenen Nachbarn (k). Geben Sie die Anzahl der nächstgelegenen Nachbarn an. Beachten Sie dabei, dass eine höhere Anzahl von 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 von Nachbarn in diesem Bereich ermitteln lassen. Wie die Anzahl der nächstgelegenen Nachbarn bestimmt wird, hängt davon ab, ob auf der Registerkarte "Merkmale" die Merkmalauswahl angegeben wurde.

- Wenn die Merkmalauswahl aktiviert wurde, wird für jeden Wert von *k* im angegebenen Bereich eine Merkmalauswahl durchgeführt und *k* und die zugehörige Funktionsgruppe mit der niedrigsten Fehlerrate (oder dem geringsten Quadratsummenfehler, falls das Ziel metrisch ist) werden ausgewählt.
- Wenn die Merkmalauswahl nicht aktiviert ist, wird eine *V*-fache Kreuzvalidierung angewendet, um die "beste" Anzahl Nachbarn zu ermitteln. Informationen zur Zuweisung von Aufteilungen finden Sie auf 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 Merkmale für einen Prädiktor ergibt sich aus dem Verhältnis der Fehlerrate oder dem Quadratsummenfehler des Modells, wobei der Prädiktor bis zum Quadratsummenfehler für das gesamte Modell vom Modell entfernt wird. Die normalisierte Wichtigkeit wird durch die Neugewichtung der Werte der Merkmalwichtigkeit berechnet, sodass 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

Auf der Registerkarte "Merkmale" können Sie Optionen für die Merkmalauswahl angeben, wenn auf der Registerkarte "Variablen" ein Ziel angegeben ist. Standardmäßig werden bei der Merkmalauswahl alle Merkmale berücksichtigt, Sie können optional aber auch ein Subset von Merkmalen auswählen, die in das Modell aufgenommen werden sollen.

Stoppkriterien. Bei jedem Schritt wird das Merkmal, dessen 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 von Funktionen in das Modell ein. Geben Sie eine positive Ganzzahl ein. Eine geringere Anzahl von Werten führt zu einem sparsameren Modell. Dabei läuft man allerdings Gefahr, wichtige Funktionen zu vernachlässigen. Bei einer höheren Anzahl von 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. Bei einem höheren Wert für die minimale Änderung 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 Merkmalauswahlfehler in der Ausgabe. Weitere Informationen finden Sie im Thema „Merkmalauswahl-Fehlerprotokoll” auf Seite 95.

Partitionen

Auf der Registerkarte "Partitionen" können Sie das Dataset in Trainings- und Holdout-Sets unterteilen und gegebenenfalls Kreuzvalidierungsaufteilungen Fälle zuweisen.

Training- und Holdout-Partitionen. Diese Gruppe gibt die Methode zur Partitionierung des aktiven Datasets 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 Dataset muss der Trainingsstichprobe zugewiesen werden, um ein Modell zu erhalten. Die **Holdout-Stichprobe** ist ein unabhängiges Set von Datensätzen, der zur Bewertung des end-

gültigen Modells verwendet wird; der Fehler für die Houldout-Stichprobe bietet eine "ehrliche" Schätzung der Vorhersagekraft des Modells, da die Holdout-Fälle (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 im aktiven Dataset 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.

Kreuzvalidierungsaufteilungen Um die "beste" Anzahl Nachbarn zu ermitteln, wird eine *V*-fache Kreuzvalidierung durchgeführt. Bei Merkmalauswahl ist sie aus Leistungsgründen nicht verfügbar.

Bei der Kreuzvalidierung wird die Stichprobe in mehrere Teilstichproben oder Aufteilungen gegliedert. Anschließend werden Nächste-Nachbarn-Modelle generiert; 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 nächstgelegener Nachbarn ist die Anzahl, die die wenigsten Fehler für alle Aufteilungen erzeugt.

- **Aufteilungen willkürlich Fälle zuweisen.** Geben Sie die Anzahl der Aufteilungen an, die für die Kreuzvalidierung herangezogen werden sollen. Die Prozedur weist Fälle willkürlich Aufteilungen zu und nummeriert sie von 1 bis *V*, der Anzahl der Aufteilungen.
- **Variable zum Zuweisen von Fällen verwenden.** Geben Sie eine numerische Variable an, die jeden Fall im aktiven Dataset 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, tritt ein Fehler auf.

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

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 Dateneditor 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.

- **Kreuzvalidierungsaufteilungsvariable.** Wenn Fälle auf der Registerkarte "Partitionen" Kreuzvalidierungsaufteilungen willkürlich zugewiesen werden, wird mit dieser Einstellung der Wert der Aufteilung gespeichert, der dieser Fall zugewiesen wurde.

Ausgabe

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 Auswahlfehlerprotokoll, ein Wichtigkeitsdiagramm für die Funktionen, ein Merkmalbereichsdiagramm, ein Peerdiagramm und eine Quadrantenkarte. Weitere Informationen finden Sie im Thema „Modellansicht".

Dateien

- **Modell als XML exportieren.** Anhand dieser Modelldatei können Sie die Modellinformationen zu Scoring-Zwecken 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 erstellt.

Optionen

Benutzerdefiniert 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.

Systemdefiniert fehlende Werte und fehlende Werte für metrische Variablen werden immer als ungültige Werte behandelt.

Modellansicht

Wenn Sie auf der Registerkarte "Ausgabe" die Option **Diagramme und Tabellen** wählen, erstellt die Prozedur ein Nächste-Nachbarn-Modellobjekt 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 einem bestimmten Merkmal des Modells an, wenn der Benutzer einen Teil der Hauptansicht ansteuert.

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

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

Merkmalbereich

Das Merkmalbereichsdiagramm ist ein interaktives Diagramm für den Merkmalbereich (bzw. -unterbereich, bei mehr als drei Merkmalen). Jede Achse stellt ein Merkmal im Modell dar und die Position der Punkte in der Tabelle gibt die Werte dieser Merkmale für Fälle in den Trainings- und Holdout-Partitionen an.

Schlüssel. Neben den Merkmalwerten 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 Trainingspartitionen 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 Merkmalbereich mit einer Reihe an Steuerelementen im Diagramm untersuchen.

- Sie können festlegen, welche Subsets an Funktionen im Diagramm angezeigt werden soll, und ändern, welche Funktionen in den Dimensionen dargestellt werden.
- "Fokusfälle" sind Punkte, die im Merkmalbereichsdiagramm 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 Steuertaste drücken und auf einen Punkt klicken, wird er dem Set an gewählten Punkten hinzugefügt. Verknüpfte Ansichten wie das Peerdiagramm werden automatisch mit den Fällen aktualisiert, die im Merkmalbereich ausgewählt werden.
- Sie können die Anzahl der 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 Merkmalbereich über die Schaltfläche "Zurücksetzen" wieder in seinen Originalzustand versetzen.

Hinzufügen und Entfernen von Feldern/Variablen

Sie können dem Merkmalbereich neue Felder/Variablen hinzufügen oder aktuell angezeigte Felder/Variablen entfernen.

Variablenpalette

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

1. Um den Modellviewer in den Bearbeitungsmodus zu versetzen, wählen Sie die folgenden Menübefehle aus:

   Ansicht > **Bearbeitungsmodus**
2. Klicken Sie im Bearbeitungsmodus auf einen beliebigen Fall im Merkmalbereich.
3. Zum Anzeigen der Variablenpalette wählen Sie die folgenden Menübefehle aus:

   Ansicht > **Paletten** > **Variablen**

In der Variablenpalette sind alle Variablen im Merkmalbereich aufgeführt. Das Symbol neben dem Variablennamen zeigt das Messniveau der Variablen an.

4. 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 Merkmalbereich zu "Zonen" hinzugefügt. Um die Zonen anzuzeigen, ziehen Sie eine Variable aus der Variablenpalette oder wählen **Zonen anzeigen**.

Der Merkmalbereiche 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 Variable in eine Zone zu verschieben, klicken Sie auf die Variable und ziehen Sie sie aus der Variablenpalette in die Zone. Wenn Sie **Zonen anzeigen** auswählen, 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 Grafikelement aus.

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.

Peers

Dieses Diagramm enthält die Fokusfälle und ihre *k* nächstgelegenen Nachbarn für jedes Merkmal im Ziel. Es ist verfügbar, wenn ein Fokusfall im Merkmalbereich ausgewählt ist.

Verknüpfungsverhalten. Das Peerdiagramm ist auf zwei Arten mit dem Merkmalbereich verknüpft.

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

Abstände zwischen nächstgelegenen Nachbarn

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

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 eines metrischen Merkmals auf der *x*-Achse nach Merkmalen in einzelne Felder unterteilt an. Es ist verfügbar, wenn ein Ziel vorhanden und ein Fokusfall im Merkmalbereich ausgewählt ist.

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

Merkmalauswahl-Fehlerprotokoll

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 dem Merkmal auf der *x*-Achse an (plus allen Merkmalen weiter links auf der *x*-Achse). Dieses Diagramm ist verfügbar, wenn ein Ziel und eine Merkmalauswahl aktiviert sind.

k-Auswahl-Fehlerprotokoll

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 der nächstgelegenen Nachbarn (*k*) auf der *x*-Achse an. Dieses Diagramm ist verfügbar, wenn ein Ziel und eine *k*-Auswahl aktiviert sind.

k- und Merkmalauswahl-Fehlerprotokoll

Dies sind Diagramme für die Merkmalauswahl (siehe „Merkmalauswahl-Fehlerprotokoll”), unterteilt nach *k*. Dieses Diagramm ist verfügbar, wenn ein Ziel und die *k*- und Merkmalauswahl aktiviert sind.

Klassifikationstabelle

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 zu den "Prozent insgesamt"-Werten, aber nicht zu den "Gesamtprozent korrekt"-Werten der Holdout-Stichprobe bei.

Fehlerzusammenfassung

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 Prädiktorvariablen, welche die beste Diskriminanz zwischen den Gruppen ergeben. Die Funktionen werden aus einer Stichprobe der Fälle generiert, bei denen die Gruppenzugehörigkeit bekannt ist. Diese Funktionen können dann auf neue Fälle mit Messungen für die Prädiktorvariablen, aber unbekannter Gruppenzugehörigkeit angewandt werden.

Hinweis: Die Gruppierungsvariable kann mehr als zwei Werte besitzen. Die Codes für die Gruppierungsvariable 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.

Statistik. Für jede Variable: For each 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.

Erläuterungen der Daten für die Diskriminanzanalyse

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

Annahmen. Die Fälle müssen unabhängig sein. Prädiktorvariablen 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 mehreren Gruppen) 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren > **Klassifizieren** > **Diskriminante...**

2. Wählen Sie eine Gruppierungsvariable mit ganzzahligen Werten aus und klicken Sie auf **Bereich definieren**, um die gewünschten Kategorien festzulegen.
3. Wählen Sie die unabhängigen Variablen oder Prädiktorvariablen aus. (Wenn die Gruppierungsvariable nicht ganzzahlig ist, können Sie eine Variable mit dieser Eigenschaft im Menü "Transformieren" mit dem Befehl "Automatisch umcodieren" erstellen.)
4. Wählen Sie die gewünschte Methode für die Eingabe der unabhängigen Variablen aus.
   - **Unabhängige Variablen zusammen aufnehmen.** Nimmt alle unabhängigen Variablen, welche die Toleranzkriterien erfüllen, gleichzeitig auf.
   - **Schrittweise Methode verwenden.** Verwendet eine schrittweise Analyse zur Steuerung von Variablenaufnahme und Variablenausschluss.
5. Wahlweise können Sie die Fälle auch mithilfe einer Auswahlvariablen auswählen.

Diskriminanzanalyse: Bereich definieren

Geben Sie den kleinsten (Minimum) und den größten (Maximum) Wert der Gruppierungsvariablen 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

So wählen Sie die Fälle für die Analyse aus:

1. Wählen Sie im Dialogfeld "Diskriminanzanalyse" eine Auswahlvariable aus.
2. 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 generiert. 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 Subsets von Lern- und Testfällen einteilen können, um so eine Gültigkeitsprüfung des generierten Modells durchzuführen.

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 nicht signifikanter 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 Diskriminanzscore (Klassifizierungsfunktionswert) aufweist.

- *Nicht standardisiert*. Zeigt die nicht standardisierten 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 in Pools zusammengefasste Korrelationsmatrix innerhalb der Gruppen an, die als Durchschnitt der separaten Kovarianzmatrizen für alle Gruppen vor der Berechnung der Korrelationen bestimmt wird.
- *Kovarianz innerhalb der Gruppen*. Zeigt eine Pools zusammengefasste Kovarianzmatrix innerhalb der Gruppen an, die sich von der Gesamtkovarianzmatrix 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

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-Distanz, 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 Streuung zwischen den Gruppen minimiert.
- *Mahalanobis-Distanz*. Dieses Maß gibt an, wie weit die Werte der unabhängigen Variablen eines Falls vom Mittelwert aller Fälle abweichen. Eine große Mahalanobis-Distanz 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 der Mahalanobis-Distanz 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. Verfügbare Alternativen sind **F-Wert verwenden** und **F-Wahrscheinlichkeit verwenden**. Geben Sie Werte zum Eingeben und Entfernen von Variablen ein.

- *F-Wert verwenden*. Eine Variable wird in ein Modell aufgenommen, wenn ihr F-Wert größer als der Aufnahmewert ist. Sie wird ausgeschlossen, wenn der F-Wert kleiner als der Ausschlusswert ist. 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 als der Aufnahmewert ist. Sie wird ausgeschlossen, wenn das Signifikanzniveau größer als der Ausschlusswert ist. 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. Zusammenfassung der Schritte. Hier 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.

Diskriminanzanalyse: 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", "Zusammenfassungstabelle" und "Klassifikation mit Fallauslassung".

- *Fallweise Ergebnisse*. Für jeden Fall werden Codes für die tatsächliche Gruppe, die vorhergesagte Gruppe, A-posteriori-Wahrscheinlichkeiten und Diskriminanzscores angezeigt.
- *Zusammenfassungstabelle*. Die Anzahl der Fälle, die auf Grundlage der Diskriminanzanalyse jeder der Gruppen richtig oder falsch zugeordnet werden. Zuweilen auch als Konfusionsmatrix 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 in Pools zusammengefasste 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

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 im XML-Format exportieren. Anhand dieser Modelldatei können Sie die Modellinformationen zu Scoring-Zwecken auf andere Datendateien anwenden.

Zusätzliche Funktionen beim Befehl DISCRIMINANT

Die Befehlssyntax ermöglicht außerdem 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 die 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`).

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

Kapitel 22. Faktorenanalyse

Mit der Faktorenanalyse wird versucht, die zugrunde liegenden Variablen oder **Faktoren** zu bestimmen, welche die Korrelationsmuster innerhalb eines Sets 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 Generieren 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 Faktorscores stehen drei Methoden zur Verfügung. Die Scores 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 konzeptuelle Bedeutung der Faktoren bestimmen. Zusätzlich können Sie für jeden Fall Faktorscores 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 Faktorscores vorherzusagen.

Statistik. 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 Faktorscores und Kovarianzmatrix des Faktors. Diagramme: Screeplot von Eigenwerten und Diagramm der Ladungen der ersten zwei oder drei Faktoren.

Erläuterungen der Daten für die Faktorenanalyse

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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren > **Dimensionsreduktion** > **Faktor...**

2. Wählen Sie die Variablen für die Faktorenanalyse aus.

Faktorenanalyse: Fälle auswählen

So wählen Sie die Fälle für die Analyse aus:

1. Wählen Sie eine Auswahlvariable aus.
2. 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

Statistik. Univariate Statistiken enthält den Mittelwert, die Standardabweichung und die Anzahl gültiger Fälle für jede Variable. Die **Anfangslösung** zeigt die anfänglichen Kommunalitäten, Eigenwerte und den Prozentsatz 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 Identitätsmatrix 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

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.

- *Analyse der Hauptkomponenten*. 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.
- *Methode der ungewichteten kleinsten 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 hoher Eindeutigkeit schwach und solche mit geringer Eindeutigkeit stärker gewichtet werden.
- *Maximum-Likelihood-Methode*. Eine Methode für die Faktorextraktion, die Parameterschätzungen 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 multiplen Korrelationskoeffizienten als Anfangsschätzungen der Kommunalitäten verwendet werden. Diese Faktorladungen werden benutzt, um neue Kommunalitäten zu schätzen, welche die alten Schätzungen 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 Image-Theorie 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

Methode. Hier können Sie die Methode der Faktorrotation auswählen. Die verfügbaren Methoden sind Varimax, Quartimax, Equamax, Promax oder Oblimin, direkt.

- *Varimax-Methode*. Eine orthogonale Rotationsmethode, die die Anzahl der Variablen mit hohen Ladungen für jeden Faktor minimiert. Sie vereinfacht die Interpretation der Faktoren.
- *Methode "Oblimin, direkt"*. Eine Methode für schiefe (nicht orthogonale) 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-Methode*. Eine Rotationsmethode, welche die Zahl der Faktoren minimiert, die zum Erklären aller Variablen benötigt werden. Sie vereinfacht die Interpretation der beobachteten Variablen.
- *Equamax-Methode*. 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 Datasets.

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.
- *Faktorladungsdiagramm*. 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: Faktorscores

Als Variablen speichern. Hiermit wird für jeden Faktor in der endgültigen Lösung eine neue Variable erstellt.

Methode. Alternative Methoden zur Berechnung der Faktorscores sind Regression, Bartlett und Anderson-Rubin.

- *Regressionsmethode*. Eine Methode, um Koeffizienten für Faktorscores zu schätzen. Die Faktorscores haben einen Mittelwert von 0 und eine Varianz, die der quadrierten multiplen Korrelation zwischen den geschätzten und den wahren Faktorscores entspricht. Die Scores können korreliert sein, selbst wenn die Faktoren orthogonal sind.
- *Bartlett-Scores*. Eine Methode, um Koeffizienten für Faktorscores zu schätzen. Die erzeugten Faktorscores 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 Faktorscores; 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 Faktorscores anzeigen. Hiermit werden die Koeffizienten angezeigt, mit denen die Variablen multipliziert werden, um Faktorscores zu erhalten. Hiermit werden auch die Korrelationen zwischen Faktorscores angezeigt.

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

Die Befehlssyntax ermöglicht außerdem Folgendes:

- Angeben von Konvergenzkriterien für die Iteration während der Extraktion und Rotation.
- Angeben von einzelnen rotierten Faktordiagrammen.
- Angeben der Anzahl der zu speichernden Faktorscores.

- Angeben der Diagonalwerte für die Hauptachsen-Faktorenanalyse.
- Schreiben der Korrelationsmatrizen oder der Faktorladungsmatrizen auf die Festplatte für eine spätere Analyse.
- Einlesen und Analysieren von Korrelationsmatrizen oder Faktorladungsmatrizen.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

Kapitel 23. Auswählen einer Prozedur zum Durchführen einer Clusteranalyse

Clusteranalysen können mit den Prozeduren "Two-Step-Clusteranalyse", "Hierarchische Clusteranalyse" oder "K-Means-Clusteranalyse" 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 Clustermodells helfen
- Gleichzeitiges Erstellen von Clustermodellen mit kategorialen und stetigen Variablen
- Speichern des Clustermodells in einer externen XML-Datei und anschließendem Einlesen dieser Datei und Aktualisieren des Clustermodells 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 Clusterzugehörigkeiten 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.

K-Means-Clusteranalyse. Die Prozedur "K-Means-Clusteranalyse" ist auf stetige Daten beschränkt ⌂und setzt eine Festlegung der Clusteranzahl 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 "K-Means-Clusteranalyse" 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 Datasets, die andernfalls nicht erkennbar wären. Der von der Prozedur verwendete Algorithmus verfügt über vielfältige nützliche Funktionen, durch die er sich von traditionellen Clustermethoden 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 Clusteranzahl.** 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 Clustermethoden 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.

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 "Clusterkriterium" 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. Weitere Informationen finden Sie im Thema „Two-Step-Clusteranalyse: Optionen” auf Seite 112.

Clusterkriterium. Mit dieser Auswahl legen Sie fest, wie die Anzahl der Cluster vom automatischen Clusteralgorithmus bestimmt wird. Angegeben werden kann entweder das Bayes-Informationskriterium (BIC) oder das Akaikes-Informationskriterium (AIC).

Erläuterungen der Daten für Two-Step-Clusteranalyse

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 Clusterfunktionsbaum und die endgültige Lösung gegebenenfalls 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 stattdessen 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Klassifizieren** > **Two-Step-Clusteranalyse...**
2. Wählen Sie mindestens eine kategoriale oder stetige Variable aus.

Die folgenden Optionen sind verfügbar:

- Passen Sie die Kriterien für die Erstellung der Cluster an.
- Wählen Sie Einstellungen für die Rauschverarbeitung, Speicherzuweisung, Variablenstandardisierung und Eingabe des Clustermodells aus.
- Fordern Sie die Ausgabe des Modellviewers an.
- Speichern Sie Modellergebnisse in der Arbeitsdatei oder in einer externen XML-Datei.

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 beschriftet. Der Ausreißercluster 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 Clusteralgorithmus 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 korrekte oder angegebene Anzahl der Cluster unter Umständen nicht ordnungsgemäß ermittelt werden.

Variablenstandardisierung. Mit dem Clusteralgorithmus werden standardisierte stetige 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 Clusteralgorithmus 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 von untergeordneten Knoten, über die ein Blattknoten verfügen kann.
- **Maximale Baumtiefe.** Die maximale Anzahl der Ebenen, über die ein CF-Baum verfügen kann.
- **Höchstmögliche Anzahl Knoten.** Gibt die maximale Anzahl der 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 beeinträchtigen 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. Weitere Informationen finden Sie im Thema „Two-Step-Clusteranalyse: Ausgabe” auf Seite 114.

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 im aktiven Dataset 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

Ausgabe. In diesem Gruppenfeld können Sie Optionen für die Anzeige der Ergebnisse der Clusteranalyse einstellen.

- **Pivot-Tabellen.** Die Ergebnisse werden in Pivot-Tabellen angezeigt.
- **Diagramme und Tabellen im Modellviewer.** Die Ergebnisse werden im Modellviewer angezeigt.
- **Evaluierungsfelder.** Mit dieser Option werden Clusterdaten für Variablen berechnet, die bei der Clustererstellung nicht verwendet wurden. Evaluierungsfelder können zusammen mit den Eingabemerkmalen im Modellviewer angezeigt werden, indem sie im untergeordneten Dialogfeld "Anzeigen" ausgewählt werden. Felder mit fehlenden Werten werden ignoriert.

Arbeitsdatendatei. Mit diesem Gruppenfeld können Sie Variablen im aktiven Dataset speichern.

- **Variable für Clusterzugehörigkeit erstellen.** Diese Variable enthält für jeden Fall eine Cluster-ID-Nummer. Der Name dieser Variablen lautet *tsc_n*, wobei *n* eine positive Ganzzahl ist, die auf die Ordinalzahl des aktiven Datasets 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 Scoring-Zwecken auf andere Datendateien anwenden.
- **CF-Baum exportieren.** Mit dieser Option können Sie den aktuellen Stand des Clusterbaums speichern und zu einem späteren Zeitpunkt mit neueren Daten aktualisieren.

Cluster-Viewer

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 Clusterdarstellung:

- 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 Kreditscore eine Rolle bei der Zugehörigkeit zu einem Cluster oder einem anderen?*

Wenn Sie die Hauptansicht und die zahlreichen verknüpften Ansichten im Cluster-Viewer nutzen, lassen sich diese Fragen beantworten.

Um Informationen über das Clustermodell anzuzeigen, aktivieren Sie (durch Doppelklicken) das Modellviewerobjekt im Cluster-Viewer.

Cluster-Viewer

Der Cluster-Viewer besteht aus zwei Bereichen, der Hauptansicht im linken Bereich und der verknüpften oder Hilfsansicht im rechten Bereich. Es gibt zwei Hauptansichten:

- Modellübersicht (Standard). Weitere Informationen finden Sie im Thema „Ansicht "Modellübersicht"" auf Seite 115.

- Cluster. Weitere Informationen finden Sie im Thema „Ansicht "Cluster"”.

Es gibt vier verknüpfte/Hilfsansichten:

- Prädiktoreinfluss. Weitere Informationen finden Sie im Thema „Ansicht "Prädiktoreinfluss" für Cluster” auf Seite 117.
- Clustergrößen (Standard). Weitere Informationen finden Sie im Thema „Ansicht "Clustergrößen"” auf Seite 117.
- Zellenverteilung. Weitere Informationen finden Sie im Thema „Ansicht "Zellverteilung"” auf Seite 117.
- Clustervergleich. Weitere Informationen finden Sie im Thema „Ansicht "Clustervergleich"” auf Seite 117.

Ansicht "Modellübersicht"

Die Ansicht "Modellübersicht" zeigt eine Momentaufnahme oder eine Übersicht des Clustermodells einschließlich eines schattierten Silhouettenmaßes der Clusterkohäsion und Clusterseparation, um schlechte, mittelmäßige und gute Ergebnisse anzuzeigen. Anhand dieser Momentaufnahme erkennen Sie schnell, ob die Qualität schlecht ist, sodass Sie dann gegebenenfalls zum Modellierungsknoten zurückkehren und die Clustermodelleinstellungen 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 Silhouettenmaß 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 Silhouettenkoeffizient von 1 würde bedeuten, dass alle Fälle direkt in ihren Clusterzentren liegen. Ein Wert -1 würde bedeuten, dass alle Fälle in den Clusterzentren anderer Cluster liegen. Ein Wert 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".
- **Eingabemerkmale.** Die Anzahl der Felder, auch bekannt als **Eingaben** oder **Prädiktoren**.
- **Cluster.** Die Anzahl der Cluster in der Lösung.

Ansicht "Cluster"

Die Ansicht "Cluster" 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.
- **Beschriftung.** Beschriftungen für jeden Cluster (ist standardmäßig leer). Doppelklicken Sie in die Zelle, um eine Beschriftung 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 Clusterstichprobe. Jede Größenzelle in der Tabelle zeigt einen vertikalen Balken, der den Größenprozentsatz innerhalb des Clusters, einen Größenprozentsatz in numerischem Format und die Clusterfallzahl anzeigt.

- **Strukturen.** Die einzelnen Eingaben oder Prädiktoren, standardmäßig nach Gesamtwichtigkeit sortiert. Wenn Spalten die gleiche Größe aufweisen, werden sie in aufsteigender Sortierfolge ihrer Clusternummern angezeigt.

  Die Gesamtwichtigkeit des Merkmals wird durch die 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 Beschriftung 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. Weitere Informationen finden Sie im Thema „Cluster und Merkmale transponieren".
- Merkmale sortieren. Weitere Informationen finden Sie im Thema „Merkmale sortieren".
- Cluster sortieren. Weitere Informationen finden Sie im Thema „Cluster sortieren".
- Zelleninhalte auswählen. Weitere Informationen finden Sie im Thema „Zelleninhalt".

Cluster und Merkmale transponieren: Standardmäßig werden Cluster als Spalten angezeigt und Merkmale als Zeilen. 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.

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 Dataset 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 Beschriftungen erstellt haben, stattdessen auch in alphanumerischer Beschriftungsreihenfolge.

Merkmale mit derselben Beschriftung werden nach Clustername sortiert. Wenn die Cluster nach Beschriftung sortiert sind und Sie die Beschriftung 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/Beschriftungen und das Lagemaß (zentrale Tendenz) für jede Cluster/Merkmal-Kombination an. Für stetige Felder wird der Mittelwert angezeigt und für kategorische Felder der Modus (die am häufigsten auftretende Kategorie) mit Kategorieprozentsatz.

- **Absolute Verteilungen.** Zeigt die Merkmalnamen/-beschriftungen 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 stetigen Merkmalen 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/-beschriftungen 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 "Prädiktoreinfluss" für Cluster

Die Ansicht "Prädiktoreinfluss" zeigt die relative Wichtigkeit jedes Felds bei Schätzung des Modells.

Ansicht "Clustergrößen"

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"

Die Ansicht "Zellverteilung" zeigt ein erweitertes, detaillierteres Diagramm der Datenverteilung für jede Merkmalszelle, die Sie in der Tabelle in der Clusterhauptanzeige auswählen.

Ansicht "Clustervergleich"

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 in der Clusterhauptanzeige. Wenn Sie die Steuertaste 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.
- Stetige Merkmale sind als Boxplots angezeigt, die die Gesamtmediane und die Interquartilbereiche anzeigen.

Vor diesen Hintergrundansichten sind Boxplots für ausgewählte Cluster dargestellt:

- Bei stetigen 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 im Cluster-Viewer

Der Cluster-Viewer ist eine interaktive Anzeige. Sie verfügen über folgende Möglichkeiten:

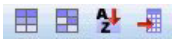
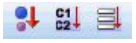
- Auswählen eines Felds oder eines Clusters für weitere Details
- Vergleichen von Clustern, um die Elemente von Interesse auszuwählen
- Verändern der Anzeige
- Transponieren von Achsen

Verwenden der Symbolleisten

Sie können die Informationen, die in den Anzeigen links und rechts erscheinen, mithilfe der Symbolleistenoptionen steuern. Mit der Symbolleistensteuerung können Sie die Ausrichtung der Anzeige ändern (oben-unten, links-rechts oder rechts-links). Außerdem können Sie den Viewer auf die Standardeinstellungen zurücksetzen und ein Dialogfeld öffnen, um den Inhalt der Ansicht "Cluster" in der Hauptanzeige zu spezifizieren.

Die Optionen **Merkmale sortieren nach**, **Cluster sortieren nach**, **Zellen** und **Anzeige** sind nur verfügbar, wenn Sie die Ansicht **Cluster** in der Hauptanzeige auswählen. Weitere Informationen finden Sie im Thema „Ansicht "Cluster"" auf Seite 115.

Tabelle 2. Symbolleistensymbole.

Symbol	Thema
	Siehe Cluster und Merkmale transponieren.
	Siehe Merkmale sortieren nach.
	Siehe Cluster sortieren nach.
	Siehe Zellen.

Anzeige "Clusteransicht steuern"

Um den Inhalt der Ansicht "Cluster" in der Hauptanzeige zu steuern, klicken Sie auf die Schaltfläche **Anzeige**. Der Anzeigedialog wird geöffnet.

Strukturen. Standardmäßig ausgewählt. Inaktivieren 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 den Modellviewer zur Evaluierung der Cluster gesendet werden); standardmäßig werden keine angezeigt. *Hinweis* Das Evaluierungsfeld muss eine Zeichenfolge mit mehreren Werten sein. Dieses Kontrollkästchen ist nicht verfügbar, wenn keine Evaluierungsfelder verfügbar sind.

Clusterbeschreibungen. Standardmäßig ausgewählt. Inaktivieren Sie das Kontrollkästchen, um alle Clusterbeschreibungszellen auszublenden.

Clustergröße. Standardmäßig ausgewählt. Inaktivieren Sie das Kontrollkästchen, um alle Clustergrößenzellen auszublenden.

Maximale Anzahl an Kategorien. Geben Sie die maximale Anzahl der Kategorien an, die in den Diagrammen der kategorischen Merkmale angezeigt werden sollen; der Standard ist 20.

Datensätze filtern

Wenn Sie weitere Informationen zu den Fällen in einem bestimmten Cluster oder einer Clustergruppe benötigen, können Sie ein Subset an Datensätzen für die weitere Analyse auf der Grundlage der ausgewählten Cluster auswählen.

1. Wählen Sie die Cluster in der Ansicht "Cluster" des Cluster-Viewers aus. Sollen mehrere Knoten ausgewählt werden, halten Sie beim Klicken die Steuertaste gedrückt.
2. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Generieren > **Datensätze filtern**
3. 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.
4. Klicken Sie auf **OK**.

Kapitel 25. Hierarchische Clusteranalyse

Mit dieser Prozedur 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" generiert. Für jeden Schritt werden Statistiken angezeigt, um Sie bei der Auswahl der besten Lösung zu unterstützen.

Beispiel. Können Gruppen von verschiedenen Fernsehshows identifiziert werden, die ein ähnliches Publikum ansprechen? Mithilfe der hierarchischen Clusteranalyse können Sie die Fernsehshows (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.

Statistik. Zuordnungsübersicht, Distanz- oder Ähnlichkeitsmatrix und Clusterzugehörigkeit für eine einzelne Lösung oder einen Bereich von Lösungen. Diagramme: Dendrogramme und Eiszapfendiagramme.

Erläuterungen der Daten für hierarchische Clusteranalyse

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 Clusterlösung(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 Clusterlösung gegebenenfalls 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Klassifizieren** > **Hierarchische Cluster...**
2. 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

Clustermethode. Verfügbar sind Linkage zwischen den Gruppen, Linkage innerhalb der Gruppen, nächstgelegener Nachbar, entferntester Nachbar, Zentroidclustering, Medianclustering 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-Quadrat-Maß und Phi-Quadrat-Maß.
- **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 generierten 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

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 Clusterschritt, in dem ein Fall (oder eine Variable) in den Cluster aufgenommen wurde.

Ähnlichkeitsmatrix. Zeigt die Distanzen oder Ähnlichkeiten zwischen den Elementen.

Clusterzugehörigkeit. 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

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

Clusterzugehörigkeit. Hiermit können Sie die Clusterzugehörigkeit 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. Die Befehlssyntax ermöglicht außerdem Folgendes:

- Verwenden mehrerer Clustermethoden in einer einzigen Analyse
- Einlesen und Analysieren einer Ähnlichkeitsmatrix

- Schreiben einer Ähnlichkeitsmatrix 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 Befehlssyntaxreferenz.

Kapitel 26. K-Means-Clusteranalyse

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 Clusterzugehörigkeit, 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 Fernsehshows identifiziert werden, die innerhalb jeder Gruppe ein ähnliches Publikum anziehen? Mit der K-Means-Clusteranalyse könnten Sie Fernsehshows (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.

Statistik. Vollständige Lösung: anfängliche Clusterzentren, ANOVA-Tabelle. Einzelne Fälle: Clusterinformationen, Distanz vom Clusterzentrum.

Erläuterungen der Daten für die K-Means-Clusteranalyse

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 potenzielle 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 der einfachen euklidischen Distanz 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 K-Means-Clusteranalyse 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 K-Means-Clusteranalyse berechnen:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Klassifizieren** > **K-Means-Clusteranalyse...**
2. Wählen Sie die Variablen für die Clusteranalyse aus.
3. 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.)
4. Wählen Sie als Methode entweder **Iterieren und klassifizieren** oder **Nur klassifizieren**.
5. Wählen Sie optional eine Identifizierungsvariable zum Beschriften der Fälle aus.

K-Means-Clusteranalyse: Effizienz

Der Befehl für die *K*-Means-Clusteranalyse 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 Dataset geschrieben und aus einer Datei oder einem Dataset ausgelesen werden. Datasets 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 Datasets müssen den Regeln zum Benennen von Variablen entsprechen.

K-Means-Clusteranalyse: Iterieren

Hinweis: Diese Optionen sind nur verfügbar, wenn Sie im Dialogfeld "K-Means-Clusteranalyse" die Methode **Iterieren und klassifizieren** auswählen.

Maximalzahl der Iterationen. Begrenzt die Anzahl der Iterationen im *K*-Means-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 vom Befehl Quick Cluster verwendeten Algorithmus vor Version 5.0 zu reproduzieren, müssen Sie **Maximalzahl der Iterationen** auf 1 setzen.

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.

K-Means-Clusteranalyse: Neue Variablen

Sie können die Informationen über die Lösung als neue Variablen speichern, um diese in nachfolgenden Analysen zu verwenden:

Clusterzugehörigkeit. Erstellt eine neue Variable, welche die endgültige Clusterzugehörigkeit 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 die euklidische Distanz zwischen jedem Fall und seinem Klassifikationszentrum anzeigt.

K-Means-Clusteranalyse: Optionen

Statistik. Sie können die folgenden Statistiken auswählen: anfängliche Clusterzentren, ANOVA-Tabelle und Clusterinformationen für jeden Fall.

- *Anfängliche Clusterzentren*. Erste Schätzung der Variablenmittelwerte 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 Clustervariable 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.
- *Clusterinformationen für jeden Fall*. Zeigt für jeden Fall die endgültige Clusterzuordnung und die euklidische Distanz zwischen dem Fall und dem Clusterzentrum an, 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 nicht fehlenden Werten berechneten Distanzen zugewiesen.

Zusätzliche Funktionen beim Befehl QUICK CLUSTER

In der Prozedur "Clusterzentrenanalyse" wird die Befehlssyntax von `QUICK CLUSTER` verwendet. Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Befehlssyntaxreferenz.

Kapitel 27. Nicht parametrische Tests

Nicht parametrische Tests gehen von Mindestannahmen über die zugrunde liegende Verteilung der Daten aus. 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 Fallset.
- Ein Test bei unabhängigen Stichproben analysiert ein Feld, das durch Kategorien eines anderen Felds gruppiert wurde.

Nicht parametrische Tests bei einer Stichprobe

Nicht parametrische Tests bei einer Stichprobe identifizieren Unterschiede in einzelnen Feldern mithilfe von einem oder mehreren nicht parametrischen Tests. Nicht parametrische Tests setzen keine Normalverteilung Ihrer Daten voraus.

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.

Berechnen nicht parametrischer Tests bei einer Stichprobe

Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren > **Nicht parametrische Tests** > **Eine Stichprobe...**

1. 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"

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.

Auswählen von Tests

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 (nicht fehlenden) 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 Flagfeldes (ein kategoriales Feld mit nur zwei Kategorien) mit der erwarteten angegebenen Binomialverteilung übereinstimmt. Sie können außerdem Konfidenzintervalle anfordern. Details zu den Testeinstellungen finden Sie in „Optionen für den Test auf Binomialverteilung".
- **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. Details zu den Testeinstellungen finden Sie in „Optionen für den Chi-Quadrat-Test" auf Seite 131.
- **Beobachtete und hypothetische Verteilung testen (Kolmogorov-Smirnov-Test).** Der Kolmogorov-Smirnov-Test wird auf stetige und ordinale Felder angewendet. Mit dieser Option wird ein Test bei einer Stichprobe erstellt, der prüft, ob die kumulative Stichprobenverteilungsfunktion für ein Feld homogen mit einer Gleich-, Normal-, Poisson- oder Exponentialverteilung ist. Details zu den Testeinstellungen finden Sie in „Optionen für den Kolmogorov-Smirnov-Test" auf Seite 131.
- **Median- und hypothetische Werte vergleichen (Wilcoxon-Test).** Der Wilcoxon-Test wird auf stetige und ordinale 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. Details zu den Testeinstellungen finden Sie in „Optionen für den Sequenztest" auf Seite 131.

Optionen für den Test auf Binomialverteilung: Der Test auf Binomialverteilung ist für Flagfelder 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 Prozeduren 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 Zeichenfolgewerten 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: 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 Zeichenfolgewerten 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: Dieses Dialogfeld gibt an, welche Verteilungen getestet werden sollten, sowie die Parameter der hypothetischen Verteilungen.

Normal. Bei Auswahl von **Stichprobendaten verwenden** werden der beobachtete Mittelwert und die Standardabweichung verwendet, mit **Benutzerdefiniert** können Sie eigene Werte bestimmen.

Gleichverteilung. Bei Auswahl von **Stichprobendaten verwenden** werden die beobachteten Mindest- und Höchstwerte verwendet, mit **Benutzerdefiniert** können Sie eigene Werte bestimmen.

Exponentialverteilung. Bei Auswahl von **Stichprobenmittelwert** wird der beobachtete Mittelwert verwendet, mit **Benutzerdefiniert** können Sie eigene Werte bestimmen.

Poisson-Verteilung. Bei Auswahl von **Stichprobenmittelwert** wird der beobachtete Mittelwert verwendet, mit **Benutzerdefiniert** können Sie eigene Werte bestimmen.

Optionen für den Sequenztest: Der Sequenztest ist für Flagfelder 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 Die folgenden Optionen sind verfügbar:

- **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 umcodieren** 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

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

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. Systemdefiniert fehlende Werte und fehlende Werte für stetige Felder werden immer als ungültige Werte behandelt.

Zusätzliche Merkmale beim Befehl NPTESTS

Die Befehlssyntax ermöglicht außerdem Folgendes:

- Angabe von Tests bei einer, bei verbundenen und bei unabhängigen Stichproben in einem einzigen Lauf der Prozedur.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

Nicht parametrische Tests bei unabhängigen Stichproben

Nicht parametrische Tests bei unabhängigen Stichproben identifizieren Unterschiede zwischen zwei oder mehr Gruppen mithilfe von einem oder mehreren nicht parametrischen Tests. Nicht parametrische Tests setzen keine Normalverteilung Ihrer Daten voraus.

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.

Berechnen nicht parametrischer Tests bei unabhängigen Stichproben

Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren > **Nicht parametrische Tests** > **Unabhängige Stichproben...**

1. 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"

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 und ordinalen 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 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 oder ordinales 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

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 nicht parametrischen Entsprechung der einfaktoriellen Varianzanalyse. Sie können optional Mehrfachvergleiche der *k*-Stichproben anfordern, entweder **alle paarweisen** Mehrfachvergleiche 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. **Klein nach groß** gibt die Alternativhypothese an, dass der Lageparameter der ersten Gruppe kleiner-gleich dem der zweiten Gruppe ist, der wiederum kleiner-gleich dem der dritten Gruppe ist usw. **Groß nach klein** gibt die Alternativhypothese an, dass der Lageparameter der ersten Gruppe größer-gleich dem der zweiten Gruppe ist, der wiederum größer-gleich dem der dritten Gruppe ist usw. Für beide Optionen setzt die Alternativhypothese auch voraus aus, dass die Lagen nicht alle gleich sind. 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 Dataset 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

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

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. Systemdefiniert fehlende Werte und fehlende Werte für stetige Felder werden immer als ungültige Werte behandelt.

Zusätzliche Merkmale beim Befehl NPTESTS

Die Befehlssyntax ermöglicht außerdem Folgendes:

- Angabe von Tests bei einer, bei verbundenen und bei unabhängigen Stichproben in einem einzigen Lauf der Prozedur.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

Nicht parametrische Tests bei verbundenen Stichproben

Identifiziert Differenzen zwischen mindestens zwei verbundenen Feldern mithilfe mindestens eines nicht parametrischen Tests. Nicht parametrische Tests setzen keine Normalverteilung Ihrer Daten voraus.

Erläuterung der Daten Jeder Datensatz entspricht einem gegebenen Befragten, für den in separaten Feldern im Dataset zwei oder mehr miteinander verbundene Messungen vorhanden sind. Beispielsweise kann eine Studie zur Wirksamkeit eines Diätplans mit nicht parametrischen 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".

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, Cochran-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 (k-Stichproben) 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.

Berechnen nicht parametrischer Tests bei verbundenen Stichproben

Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren > **Nicht parametrische Tests** > **Verbundene Stichproben...**

1. 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"

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 die Prozedur 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.

Auswählen von Tests

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, Cochran-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 zweifaktorielle ANOVA für Ränge nach Friedman 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. 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 Flagfeldern (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. Details zu den Testeinstellungen finden Sie in „McNemar-Test: Erfolg definieren” auf Seite 137. **Cochran-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* Flagfeldern (kategoriale Felder mit nur zwei Werten) gleich wahrscheinlich sind. Sie können optional Mehrfachvergleiche der *k*-Stichproben anfordern, entweder **alle paarweisen** Mehrfachvergleiche oder **schrittweise Step-down**-Vergleiche. Details zu den Testeinstellungen finden Sie in „Cochran-Q: Erfolg definieren” auf Seite 137.
- **Test auf Veränderungen in multinomialen Daten. Randhomogenitätstest (2 Stichproben)** erstellt einen Test bei verbundenen Stichproben, der prüft, ob Wertekombinationen zwischen zwei paarigen ordinalen Feldern gleich wahrscheinlich sind. Der Randhomogenitätstest 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 Feldern von 0 abweichen. Der Test wird auf stetige und ordinale Felder angewendet. Diese Tests werden nicht durchgeführt, wenn auf 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 paarigen Feldern erstellt. Der Test wird auf stetige und ordinale Felder angewendet. Der Test wird nicht durchgeführt, wenn auf der Registerkarte "Felder" mehr als zwei Felder angegeben wurden.
- **Zusammenhänge quantifizieren.** Der **Konkordanzkoeffizient 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: Der McNemar-Test ist für Flagfelder 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 Zeichenfolgewerten oder numerischen Werten an. Die Werte in der Liste müssen nicht in der Stichprobe vorhanden sein.

Cochran-Q: Erfolg definieren: Cochran-Q-Test ist für Flagfelder 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 Zeichenfolgewerten oder numerischen Werten an. Die Werte in der Liste müssen nicht in der Stichprobe vorhanden sein.

Testoptionen

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

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. Systemdefiniert fehlende Werte und fehlende Werte für stetige Felder werden immer als ungültige Werte behandelt.

Zusätzliche Merkmale beim Befehl NPTESTS

Die Befehlssyntax ermöglicht außerdem Folgendes:

- Angabe von Tests bei einer, bei verbundenen und bei unabhängigen Stichproben in einem einzigen Lauf der Prozedur.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

Modellanzeige

Modellanzeige

Die Prozedur erstellt ein Modellansichtsobjekt 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. Die ist die Standardansicht. Weitere Informationen finden Sie im Thema „Hypothesenübersicht" auf Seite 139.
- Konfidenzintervallübersicht. Weitere Informationen finden Sie im Thema „Konfidenzintervallübersicht" auf Seite 139.

Es gibt sieben verknüpfte/Hilfsansichten:

- Ansicht Test bei einer Stichprobe. Dies ist die Standardansicht, falls Tests bei einer Stichprobe angefordert wurden. Weitere Informationen finden Sie im Thema „Test bei einer Stichprobe" auf Seite 139.
- Ansicht Test bei verbundenen Stichproben. Dies ist die Standardansicht, falls keine Tests bei einer Stichprobe, sondern Tests bei mehreren verbundenen Stichproben angefordert wurden. Weitere Informationen finden Sie im Thema „Test bei verbundenen Stichproben" auf Seite 140.
- Ansicht Test bei unabhängigen Stichproben. Dies ist die Standardansicht, falls keine Tests bei mehreren verbundenen Stichproben oder Tests bei einer Stichprobe angefordert wurden. Weitere Informationen finden Sie im Thema „Test bei unabhängigen Stichproben" auf Seite 141.
- Informationen über kategoriales Feld. Weitere Informationen finden Sie im Thema „Informationen über kategoriales Feld" auf Seite 142.
- Informationen über stetiges Feld. Weitere Informationen finden Sie im Thema „Informationen über stetiges Feld," auf Seite 142.
- Paarweise Vergleiche. Weitere Informationen finden Sie im Thema „Paarweise Vergleiche" auf Seite 142.
- Homogene Subsets. Weitere Informationen finden Sie im Thema „Homogene Subsets" auf Seite 142.

Hypothesenübersicht

Mit der Ansicht "Modellzusammenfassung" erhalten Sie eine momentane, übersichtliche Zusammenfassung der nicht parametrischen 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 den Modellviewer über die Schaltfläche **Zurücksetzen** wieder in ihren Originalzustand versetzen.
- Die Dropdown-Liste **Feldfilter** ermöglicht es, nur diejenigen Tests anzuzeigen, die das ausgewählte Feld betreffen.

Konfidenzintervallübersicht

Die Konfidenzintervallübersicht zeigt alle Konfidenzintervalle an, die von den nicht parametrischen 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 nicht parametrischen 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

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 einer QuickInfo die Prozentsätze 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

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

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

Der Sequenztest zeigt ein Diagramm und eine Testtabelle an.

- Das Diagramm zeigt eine Normalverteilung an, in der die beobachtete Anzahl von 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

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 nicht parametrischen 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

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 2x2-Tabelle an.
- Die Tabelle zeigt Details zum Test an.

Vorzeichentest

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

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.

Randhomogenitätstest

Der Randhomogenitätstest 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.

Cochran-Q-Test

Cochran-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 einer QuickInfo die Prozentsätze der Kategorien angezeigt.
- Die Tabelle zeigt Details zum Test an.

Zweifaktorielle Varianzanalyse für Ränge nach Friedman

Die zweifaktorielle Varianzanalyse für Ränge nach Friedman 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.

Konkordanzkoeffizient nach Kendall

Die Ansicht Konkordanzkoeffizient 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 nicht parametrischen 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

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

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

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

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

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

Der Test auf Extremreaktionen nach Moses zeigt Boxplots und eine Testtabelle an.

- Für jede Kategorie des Gruppierungsfelds werden separate Boxplots angezeigt. Die Punktbeschriftungen können angezeigt oder ausgeblendet werden, indem Sie auf die Schaltfläche **Datensatz-ID** klicken.
- Die Tabelle zeigt Details zum Test an.

Mediantest

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

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 einer QuickInfo die Prozentsätze der Kategorien angezeigt.

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

Die Ansicht Paarweise Vergleiche zeigt ein Abstandsnetzdiagramm und eine Vergleichstabelle an, die von nicht parametrischen Tests bei *k* Stichproben erstellt werden, wenn paarweise Mehrfachvergleiche angefordert werden.

- Das Abstandsnetzdiagramm ist eine grafische Darstellung der Vergleichstabelle, in der die Abstände zwischen Knoten im Netz den Unterschieden zwischen Stichproben entsprechen. Gelbe Linien entsprechen statistisch signifikanten Unterschieden, schwarze Linien nicht signifikanten Unterschieden. Wenn Sie die Maus über eine Linie im Netz 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 Subsets

Die Ansicht "Homogene Subsets" zeigt eine Vergleichstabelle an, die von nicht parametrischen 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 unterschiedlich sind, werden in gleichfarbigen Subsets gruppiert. Für jede identifizierte Untergruppe ist eine separate Spalte vorhanden. Wenn alle Stichproben statistisch signifikant unterschiedlich sind, ist für jede Stichprobe ein separates Subset vorhanden. Wenn keine der Stichproben statistisch signifikant unterschiedlich ist, ist nur ein Subset vorhanden.
- Für jedes Subset mit mehreren Stichproben werden eine Teststatistik, ein Signifikanzwert und ein angepasster Signifikanzwert berechnet.

Zusätzliche Merkmale beim Befehl NPTESTS

Die Befehlssyntax ermöglicht außerdem Folgendes:

- Angabe von Tests bei einer, bei verbundenen und bei unabhängigen Stichproben in einem einzigen Lauf der Prozedur.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

Veraltete Dialogfelder

Es gibt einige "veraltete" Dialogfelder, die ebenfalls nicht parametrische 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, eine Exponentialverteilung oder 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 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 nicht fehlender 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 von 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.

Statistik. Mittelwert, Standardabweichung, Minimum, Maximum und Quartile. Die Anzahl und der Prozentsatz nicht fehlender und fehlender Fälle, die Anzahl der für jede Kategorie beobachteten und erwarteten Fälle, Residuen und die Chi-Quadrat-Statistik.

Erläuterungen der Daten für den Chi-Quadrat-Test

Daten. Verwenden Sie geordnete oder nicht geordnete numerische kategoriale Variablen (nominales oder ordinales Niveau der Messwerte). Verwenden Sie zum Umwandeln von Zeichenfolgevariablen in numerische Variablen den Befehl "Automatisch umcodieren" im Menü "Transformieren".

Annahmen. Nicht parametrische 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Nicht parametrische Tests** > **Veraltete Dialogfelder** > **Chi-Quadrat...**
2. Wählen Sie mindestens eine Testvariable aus. Mit jeder Variablen wird ein separater Test erzeugt.
3. 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 **Angegebenen Bereich 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

Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistiken.** Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nicht fehlenden 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)

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Befehlssyntaxreferenz.

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.

Statistik. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nicht fehlenden Fälle und Quartile.

Erläuterungen der Daten für den Test auf Binomialverteilung

Daten. Die getesteten Variablen müssen numerisch und dichotom sein. Verwenden Sie zum Umwandeln von Zeichenfolgevariablen in numerische Variablen den Befehl "Automatisch umcodieren" im Menü "Transformieren". Eine **dichotome Variable** ist eine Variable, die nur zwei mögliche Werte annehmen kann: *ja* oder *nein*, *wahr* oder *falsch*, 0 oder 1 usw. Der erste in dem Dataset 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. Nicht parametrische 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren > **Nicht parametrische Tests** > **Veraltete Dialogfelder** > **Binomial...**

2. Wählen Sie mindestens eine numerische Testvariable.
3. 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

Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistiken.** Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nicht fehlenden 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)

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Befehlssyntaxreferenz.

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.

Statistik. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nicht fehlenden Fälle und Quartile.

Erläuterungen der Daten für Sequenzentest

Daten. Die Variablen müssen numerisch sein. Verwenden Sie zum Umwandeln von Zeichenfolgevariablen in numerische Variablen den Befehl "Automatisch umcodieren" im Menü "Transformieren".

Annahmen. Nicht parametrische Tests erfordern keine Annahmen über die Form der zugrunde liegenden Verteilung. Verwenden Sie Stichproben aus stetigen Wahrscheinlichkeitsverteilungen.

So lassen Sie einen Sequenzentest berechnen:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Nicht parametrische Tests** > **Veraltete Dialogfelder** > **Sequenzen...**
2. Wählen Sie mindestens eine numerische Testvariable.

3. 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

Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistiken.** Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nicht fehlenden 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)

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Befehlssyntaxreferenz.

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.

Statistik. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nicht fehlenden Fälle und Quartile.

Erläuterungen der Daten für den Kolmogorov-Smirnov-Anpassungstest

Daten. Die Variablen müssen auf Intervall- oder Verhältnismessniveau quantitativ sein.

Annahmen. Für den Kolmogorov-Smirnov-Test wird angenommen, dass die Parameter der Testverteilung 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Nicht parametrische Tests** > **Veraltete Dialogfelder** > **K-S bei einer Stichprobe...**
2. Wählen Sie mindestens eine numerische Testvariable. Mit jeder Variablen wird ein separater Test erzeugt.
3. 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

Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistiken.** Bei dieser Option werden Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nicht fehlenden 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 Befehlssyntaxsprache können Sie auch die Parameter der Testverteilung angeben (mit dem Unterbefehl `K-S`).

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

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.

Statistik. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nicht fehlenden Fälle und Quartile. Tests: Mann-Whitney-*U*-Test, Extremreaktionen nach Moses, Kolmogorov-Smirnov-*Z*-Test, Sequenztest nach Wald-Wolfowitz.

Erläuterungen der Daten für Tests bei zwei unabhängigen Stichproben

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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Nicht parametrische Tests** > **Veraltete Dialogfelder** > **2 unabhängige Stichproben...**
2. Wählen Sie mindestens eine numerische Variable aus.
3. Wählen Sie eine Gruppierungsvariable 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 Score aus Gruppe 2 und wie oft ein Wert aus Gruppe 2 einem Score 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 getrimmt.

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

Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistiken.** Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nicht fehlenden 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 Befehlssyntaxsprache kann die Anzahl der Fälle angegeben werden, die für den Moses-Test getrimmt werden sollen.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

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.

Statistik. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nicht fehlenden Fälle und Quartile. Tests: Wilcoxon-Test, Vorzeichentest, McNemar. Wenn die Option "Exakte Tests" installiert ist (nur unter Windows-Betriebssystemen verfügbar) steht außerdem der Randhomogenitätstest zur Verfügung.

Erläuterungen der Daten für Tests bei zwei verbundenen Stichproben

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 paarigen Differenzen als symmetrisch angenommen.

So lassen Sie Tests bei zwei verbundenen Stichproben berechnen:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Nicht parametrische Tests** > **Veraltete Dialogfelder** > **2 verbundene Stichproben...**
2. 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 Rücklaufquote am Anfang (vor dem Ereignis) gleich der Rücklaufquote 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 **Randhomogenitätstest**. 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 Randhomogenitätstest ist nur verfügbar, wenn Sie die Option Exact Tests installiert haben.

Optionen für Tests bei zwei verbundenen Stichproben

Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistiken.** Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nicht fehlenden 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 Befehlssyntaxsprache können Sie außerdem eine Variable mit jeder Variable auf einer Liste überprüfen.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

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.

Statistik. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nicht fehlenden Fälle und Quartile. Tests: Kruskal-Wallis-*H*, Median.

Erläuterungen der Daten für Tests bei mehreren unabhängigen Stichproben

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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Nicht parametrische Tests** > **Veraltete Dialogfelder** > **K unabhängige Stichproben...**
2. Wählen Sie mindestens eine numerische Variable aus.

3. Wählen Sie eine Gruppierungsvariable aus und klicken Sie auf **Bereich definieren**, um die ganzzahligen Minimal- und Maximalwerte der Gruppierungsvariablen 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 nicht parametrische 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 Zusatzmodul Exact Tests installiert haben.

Tests bei mehreren unabhängigen Stichproben: Bereich definieren

Um den Bereich zu definieren, geben Sie für **Minimum** und **Maximum** ganzzahlige Werte ein, die der niedrigsten und höchsten Kategorie der Gruppierungsvariablen 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

Statistik. Sie können eine oder beide Auswertungsstatistiken wählen.

- **Deskriptive Statistiken.** Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nicht fehlenden 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 Befehlssyntaxsprache 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 Befehlssyntaxreferenz.

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.

Statistik. Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl der nicht fehlenden Fälle und Quartile. Tests: Friedman, Kendall-*W* und Cochran-*Q*.

Erläuterungen der Daten für Tests bei mehreren verbundenen Stichproben

Daten. Verwenden Sie numerische Variablen, die geordnet werden können.

Annahmen. Nicht parametrische Tests erfordern keine Annahmen über die Form der zugrunde liegenden Verteilung. Verwenden Sie abhängige Zufallsstichproben.

So lassen Sie Tests bei mehreren verbundenen Stichproben berechnen:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Nicht parametrische Tests** > **Veraltete Dialogfelder** > **K verbundene Stichproben...**
2. 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 **Friedman-Test** stellt das nicht parametrische 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 die *k* Variablen in eine Rangordnung von 1 bis *k* gebracht. Die Teststatistik beruht auf dieser Rangordnung.

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

Sie können Statistiken auswählen.

- **Deskriptive Statistiken.** Zeigt Mittelwert, Standardabweichung, Minimum, Maximum und Anzahl der nicht fehlenden 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 Befehlssyntaxreferenz.

Kapitel 28. Analyse von Mehrfachantworten

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 Mehrfachantwortsets 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 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 mehrere Antworten ankreuzen kann. Diese Frage kann allerdings nicht direkt codiert werden, weil eine Variable nur einen Wert je Fall aufweisen 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 mithilfe eines Codes angegeben. Beim Durchsehen einer Stichprobe von Fragebogen 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 codiert 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 Mehrfachantwortsets definieren. Jedes Set muss über einen eigenen eindeutigen Namen verfügen. Sie können ein Set entfernen, indem Sie es in der Liste der Mehrfachantwortsets 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 von dichoto-

men 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 von 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. Mit der Prozedur werden alle unterschiedlichen ganzzahligen Werte in dem einschließenden Bereich aller Komponentenvariablen addiert. Leere Kategorien werden nicht in Tabellen übernommen.

Sie müssen jedem Mehrfachantwortset einen eindeutigen Namen zuweisen, der aus bis zu sieben Zeichen bestehen darf. Die Prozedur stellt dem Namen, den Sie zuweisen, ein Dollarzeichen ($) voran. Folgende reservierte Namen können Sie nicht verwenden: *casenum*, *sysmis*, *jdate*, *date*, *time*, *length* und *width*. Der Name des Mehrfachantwortsets ist nur zur Verwendung in Mehrfachantworten-Prozeduren vorgesehen. In anderen Prozeduren können Sie sich nicht auf Namen von Mehrfachantwortsets beziehen. Wahlweise können Sie für das Mehrfachantwortset eine aussagekräftige Variablenbeschriftung eingeben. Die Beschriftung kann bis zu 40 Zeichen lang sein.

So definieren Sie Mehrfachantwortsets

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Mehrfachantwort** > **Variablensets definieren...**
2. Wählen Sie mindestens zwei Variablen aus.
3. Wenn Ihre Variablen als Dichotomien codiert sind, geben Sie an, welcher Wert gezählt werden soll. Wenn Ihre Variablen als Kategorien codiert sind, legen Sie den Bereich für die Kategorien fest.
4. Geben Sie einen eindeutigen Namen für jedes Mehrfachantwortset ein.
5. Klicken Sie auf **Hinzufügen**, um das Mehrfachantwortset zur Liste der definierten Sets hinzuzufügen.

Mehrfachantworten: Häufigkeiten

Mit der Prozedur "Mehrfachantworten: Häufigkeiten" erstellen Sie Häufigkeitstabellen für Mehrfachantwortsets. Zuvor müssen Sie mindestens ein Mehrfachantwortset definieren (siehe "Mehrfachantworten: Sets definieren").

Bei Sets aus dichotomen Variablen entsprechen die in der Ausgabe gezeigten Kategorienamen den Variablenbeschriftungen, die für die elementaren Variablen in der Gruppe festgelegt wurden. Wenn keine Variablenbeschriftungen festgelegt wurden, werden die Variablennamen als Beschriftungen verwendet. Bei Sets aus kategorialen Variablen entsprechen die Kategoriebeschriftungen den Wertbeschriftungen der ersten Variable in der Gruppe. Wenn Kategorien, die bei der ersten Variable fehlen, bei anderen Variablen in der Gruppe vorhanden sind, müssen Sie eine Wertbeschriftung 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 Mehrfachantwortsets, die als Sets aus dichotomen Variablen definiert wurden. In der Standardeinstellung gilt ein Fall in einem Set von 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 Mehrfachantwortsets, die als Sets aus kategorialen Variablen definiert wurden. In der Standardeinstellung gilt ein Fall in einem Set von kategorialen Variablen nur als fehlend, wenn keine der Komponenten des Falls gültige Werte innerhalb des definierten Bereichs enthält.

Beispiel. Jede Variable, die sich aus einer Umfrage ergibt, ist eine elementare Variable. Zum Analysieren der Mehrfachantworten müssen Sie die Variablen in einem der beiden möglichen Typen von Mehrfachantwortsets zusammenfassen: in einem Set von dichotomen Variablen oder in einem Set von 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 von dichotomen Variablen** definiert, dann würde jede der drei Variablen im Set zu einer Kategorie der Gruppenvariablen werden. Die Angaben zu Anzahl und Prozentsatz 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 von 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 Prozentsatz für jede Fluggesellschaft werden zusammen in einer Häufigkeitstabelle angezeigt.

Statistik. Häufigkeitstabellen mit den Häufigkeiten, Prozentsätzen der Antworten, Prozentsätzen der Fälle, der Anzahl gültiger Fälle und der Anzahl fehlender Fälle.

Erläuterungen der Daten für Mehrfachantworten - Häufigkeiten

Daten. Verwenden Sie Mehrfachantwortsets.

Annahmen. Die Häufigkeiten und Prozentsätze geben nützliche Beschreibungen für Daten mit beliebigen Verteilungen.

Verwandte Prozeduren. Mit der Prozedur "Mehrfachantworten: Sets definieren" können Sie Mehrfachantwortsets definieren.

So berechnen Sie Häufigkeiten mit Mehrfachantworten:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Mehrfachantwort** > **Häufigkeiten...**
2. Wählen Sie mindestens ein Mehrfachantwortset aus.

Mehrfachantworten: Kreuztabellen

Mit der Prozedur "Mehrfachantworten: Kreuztabellen" können Kreuztabellen für definierte Mehrfachantwortsets, elementare Variablen oder eine Kombination dieser Elemente berechnet werden. Sie können außerdem Prozentsätze für Zellen basierend auf Fällen oder Antworten berechnen lassen, die Verarbeitung von fehlenden Werten ändern oder paarige Kreuztabellen erstellen lassen. Zuvor müssen Sie mindestens ein Mehrfachantwortset definieren (siehe "So definieren Sie Mehrfachantwortsets").

Bei Sets aus dichotomen Variablen entsprechen die in der Ausgabe gezeigten Kategorienamen den Variablenbeschriftungen, die für die elementaren Variablen in der Gruppe festgelegt wurden. Wenn keine Variablenbeschriftungen festgelegt wurden, werden die Variablennamen als Beschriftungen verwendet. Bei Sets aus kategorialen Variablen entsprechen die Kategoriebeschriftungen den Wertbeschriftungen der ersten Variable in der Gruppe. Wenn Kategorien, die bei der ersten Variable fehlen, bei anderen Variablen in der Gruppe vorhanden sind, müssen Sie eine Wertbeschriftung für die fehlenden Kategorien festlegen. Die Prozedur zeigt Kategoriebeschriftungen für Spalten in drei Zeilen mit bis zu acht Zeichen je Zeile an. Wenn Sie vermeiden möchten, dass Wörter getrennt werden, können Sie die Anordnung von Zeilen und Spalten umdrehen oder die Beschriftungen 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 Reisende um folgende Informationen gebeten: "Kreuzen Sie bitte alle 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.

Statistik. Kreuztabellen mit Häufigkeiten pro Zelle, Zeile, Spalte und Gesamt sowie Prozentsätzen für Zellen, Zeilen, Spalten und Gesamt. Die Prozentsätze für die Zellen können auf Fällen oder auf Antworten basieren.

Erläuterungen der Daten für Mehrfachantworten - Kreuztabellen

Daten. Verwenden Sie Mehrfachantwortsets 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 Mehrfachantwortsets definieren.

So berechnen Sie Kreuztabellen mit Mehrfachantworten:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Mehrfachantwort** > **Kreuztabellen...**
2. Wählen Sie mindestens eine numerische Variable oder mindestens ein Mehrfachantwortset für jede Dimension der Kreuztabelle aus.
3. Definieren Sie den Bereich jeder elementaren Variablen.

Außerdem können Sie eine Zweiwege-Kreuztabelle für jede Kategorie einer Kontrollvariablen oder eines Mehrfachantwortsets berechnen lassen. Wählen Sie mindestens einen Eintrag für die Liste "Schicht(en)" aus.

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

Prozentsätze für Zellen. Die Zellenhäufigkeiten werden immer angezeigt. Sie können aber auch Spalten- und Zeilenprozentsätze sowie Prozentsätze für Zweiwegetabellen (Gesamtwerte) anzeigen lassen.

Prozentsätze bezogen auf. Sie können festlegen, dass die Prozentsätze für die Zellen auf Fällen (oder Befragten) basieren. Diese Option ist nicht verfügbar, wenn Sie Variablen aus verschiedenen Sets von kategorialen Variablen abgleichen. 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 Mehrfachantwortsets, die als Sets aus dichotomen Variablen definiert wurden. In der Standardein-

stellung gilt ein Fall in einem Set von 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 Mehrfachantwortsets, die als Sets aus kategorialen Variablen definiert wurden. In der Standardeinstellung gilt ein Fall in einem Set von kategorialen Variablen nur als fehlend, wenn keine der Komponenten des Falls gültige Werte innerhalb des definierten Bereichs enthält.

Wenn zwei Sets von kategorialen Variablen in eine Kreuztabelle aufgenommen werden, tabuliert die Prozedur standardmäßig jede Variable in der ersten Gruppe mit jeder Variablen in der zweiten Gruppe und addiert die Anzahlen für die einzelnen Zellen; daher können manche Antworten mehrmals in einer Tabelle vorkommen. Sie können die folgende Option auswählen:

Variablen aus den Antwortsets abgleichen. Hiermit wird die erste Variable aus der ersten Gruppe mit der ersten Variable aus der zweiten Gruppe abgeglichen 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 die Paarbildung (Abgleich) nicht zur Verfügung.

Zusätzliche Funktionen beim Befehl MULT RESPONSE

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Wertbeschriftungen unterdrückt werden.

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

Kapitel 29. Ergebnisberichte

Ergebnisberichte

Auflistungen von Fällen und deskriptive Statistiken sind wichtige Hilfsmittel zur Untersuchung und Darstellung von Daten. Mit dem Dateneditor oder der Prozedur "Berichte" können Sie Fälle auflisten, mit den Prozeduren "Häufigkeiten" Häufigkeitszähler und deskriptive Statistiken erstellen und mit der Prozedur "Mittelwert" Statistiken für Teilgesamtheiten 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 (Breakvariablen) 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.

Breakspalten. Hier werden optionale Breakvariablen aufgelistet, die den Bericht in Gruppen aufteilen, und Einstellungen für die Auswertungsstatistik sowie Anzeigeformate für Breakspalten festgelegt. Bei mehreren Breakvariablen wird für jede Kategorie einer Breakvariablen eine getrennte Gruppe innerhalb der Kategorien der vorhergehenden Breakvariablen in der Liste erzeugt. Die Breakvariablen müssen diskrete kategoriale Variablen sein, welche die Fälle in eine begrenzte Anzahl von sinnvollen Kategorien aufteilen. Die Einzelwerte jeder Breakvariablen 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, Seitennummerierung und Titel.

Fälle anzeigen. Hiermit werden für jeden Fall die aktuellen Werte (oder Wertbeschriftungen) 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 Breakvariablen muss die Datendatei vor dem Erstellen des Berichts nach den Werten der Breakvariablen sortiert werden. Wenn Ihre Datendatei bereits nach den Werten der Breakvariablen sortiert ist, können Sie durch Auswählen dieser Option Verarbeitungszeit einsparen. Diese Option ist besonders hilfreich, wenn Sie bereits einen Bericht für die Vorschau erstellt haben.

Erstellen eines Zusammenfassungsberichts: Bericht in Zeilen

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Berichte** > **Bericht in Zeilen...**
2. Wählen Sie mindestens eine Variable für die Datenspalten aus. Für jede ausgewählte Variable wird eine Spalte im Bericht generiert.
3. Wählen Sie bei sortierten und nach Untergruppen angezeigten Berichten mindestens eine Variable für die Breakspalten aus.
4. Bei Berichten mit Auswertungsstatistiken für Untergruppen, die durch Breakvariablen definiert wurden, wählen Sie in der Liste "Breakspaltenvariablen" die Breakvariablen aus und klicken Sie im Gruppenfeld "Breakspalten" auf **Auswertung**, um das (die) Auswertungsmaß(e) festzulegen.
5. Bei Berichten mit zusammenfassenden Auswertungsstatistiken klicken Sie auf **Auswertung**, um das (die) Auswertungsmaß(e) festzulegen.

Datenspaltenformat/Breakformat in Berichten

In den Formatdialogfeldern werden Spaltentitel, Spaltenbreite, Textausrichtung sowie Anzeige der Datenwerte oder Wertbeschriftungen festgelegt. Mit "Datenspaltenformat" wird das Format der Datenspalten auf der rechten Seite des Berichtes festgelegt. Das Format der Breakspalten auf der linken Seite wird mit "Breakformat" festgelegt.

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 Wertbeschriftungen in der Spalte festgelegt. Die Ausrichtung der Werte oder Beschriftungen 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 Wertbeschriftungen der ausgewählten Variablen. Für Werte ohne definierte Wertbeschriftungen 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 Breakgruppen und für den gesamten Bericht fest. Mit "Auswertung" können Sie Einstellungen bezüglich der Untergruppenstatistik für jede durch die Breakvariablen definierte Kategorie vornehmen. Mit "Endgültige Auswertungszeilen" können Sie Einstellungen für die am Ende des Berichts angezeigte Gesamtstatistik vornehmen.

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: Breakoptionen

Mit "Breakoptionen" werden Abstand und Seitenaufteilung der Informationen in den Breakkategorien festgelegt.

Seiteneinstellung. Hiermit werden Abstand und Seitenaufteilung für Kategorien der ausgewählten Breakvariablen festgelegt. Sie können eine Anzahl von Leerzeilen zwischen den Breakkategorien festlegen oder eine Breakkategorie an einen neuen Seitenanfang legen.

Leerzeilen vor Zusammenfassung. Hiermit legen Sie die Anzahl der Leerzeilen zwischen Beschriftungen oder Daten von Breakkategorien und Auswertungsstatistiken fest. Dies bietet sich besonders für kombinierte Berichte mit Listen von einzelnen Fällen und Auswertungsstatistiken für Breakkategorien 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.

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.

Seitennummerierung 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.

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.

Breakspalten. Hiermit wird die Anzeige der Breakspalten festgelegt. Wenn mehrere Breakvariablen festgelegt wurden, können sie sich in getrennten Spalten oder in der ersten Spalte befinden. Das Anordnen aller Breakvariablen 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 Breakbeschriftungen zu Beginn jeder Breakkategorie. Die erste Informationszeile in der Datenspalte kann entweder in der gleichen Zeile wie die Beschriftung der Breakkategorie oder nach einer festgelegten Anzahl von Zeilen nach der Beschriftung der Breakkategorie 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.

Wenn Sie in Titeln oder Fußzeilen Variablen eingeben, wird die aktuelle Wertbeschriftung oder der Wert der Variablen im Titel oder in der Fußzeile angezeigt. In Titeln wird die Wertbeschriftung angezeigt, die dem Wert der Variablen am Beginn der Seite entspricht. In den Fußzeilen wird die Wertbeschriftung angezeigt, die dem Wert der Variablen am Ende der Seite entspricht. Ist keine Wertbeschriftung 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.

Breakspalten. Hiermit werden optionale Breakvariablen, die den Bericht in Gruppen aufteilen, aufgelistet und das Anzeigeformat der Breakspalten festgelegt. Bei mehreren Breakvariablen wird für jede Kategorie einer Breakvariablen eine getrennte Gruppe innerhalb der Kategorien der vorhergehenden Breakvariablen in der Liste erzeugt. Die Breakvariablen 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, Seitennummerierung 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 Breakvariablen muss die Datendatei vor dem Erstellen des Berichts nach den Werten der Breakvariablen sortiert werden. Wenn Ihre Datendatei bereits nach den Werten der Breakvariablen sortiert ist, können Sie durch Auswählen dieser Option Verarbeitungszeit einsparen. Diese Option ist besonders hilfreich, wenn Sie bereits einen Bericht für die Vorschau erstellt haben.

Erstellen eines Zusammenfassungsberichts: Bericht in Spalten

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Berichte** > **Bericht in Spalten...**
2. Wählen Sie mindestens eine Variable für die Datenspalten aus. Für jede ausgewählte Variable wird eine Spalte im Bericht generiert.
3. Um das Auswertungsmaß für eine Variable zu ändern, wählen Sie die Variable in der Liste "Datenspaltenvariablen" aus und klicken Sie auf **Auswertung**.
4. Um mehr als ein Auswertungsmaß für eine Variable berechnen zu lassen, wählen Sie die Variable in der Quellenliste aus und übernehmen diese für jedes gewünschte Auswertungsmaß in die Liste "Datenspaltenvariablen".
5. 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.
6. Wählen Sie bei sortierten und nach Untergruppen angezeigten Berichten mindestens eine Variable für die Breakspalten aus.

Datenspalten: Auswertungsfunktion

Im Dialogfeld "Auswertung" wird die angezeigte Auswertungsstatistik der ausgewählten Datenspaltenvariablen festgelegt.

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 Gesamtauswertungsstatistik festgelegt, die zwei oder mehr Datenspalten zusammenfasst.

Die folgenden Gesamtauswertungsstatistiken 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.

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 Breakspalten für "Bericht in Spalten" entsprechen den Optionen für "Bericht in Zeilen".

Bericht: Breakoptionen für Bericht in Spalten

Mit "Breakoptionen" werden Anzeige der Zwischenergebnisse, Abstand und Seitenaufteilung für Breakkategorien festgelegt.

Zwischenergebnis. Hiermit wird die Anzeige der Zwischenergebnisse für Breakkategorien festgelegt.

Seiteneinstellung. Hiermit werden Abstand und Seitenaufteilung für Kategorien der ausgewählten Breakvariablen festgelegt. Sie können eine Anzahl von Leerzeilen zwischen den Breakkategorien festlegen oder eine Breakkategorie an einen neuen Seitenanfang legen.

Leerzeilen vor Zwischenergebnis. Hiermit legen Sie die Anzahl leerer Zeilen zwischen den Daten der Breakkategorien und den Zwischenergebnissen fest.

Bericht: Optionen für Bericht in Spalten

Mit "Optionen" werden Anzeige der Gesamtergebnisse, Anzeige der fehlenden Werte und Seitenaufteilung in Auswertungsberichten in Spalten festgelegt.

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 Layoutoptionen für "Bericht in Spalten" entsprechen den Optionen für "Bericht in Zeilen".

Zusätzliche Funktionen beim Befehl REPORT

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Befehlssyntaxreferenz.

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. Intraklassen-Korrelationskoeffizienten können verwendet werden, um Reliabilitätsschätzungen 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.

Statistik. Deskriptive Statistiken für jede Variable und für die Skala, Auswertungsstatistik für mehrere Items, Inter-Item-Korrelationen und Inter-Item-Kovarianzen, Reliabilitätsschätzungen, ANOVA-Tabelle, Intraklassen-Korrelationskoeffizienten, 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.

Erläuterungen der Daten für die Reliabilitätsanalyse

Daten. Die Daten können dichotom, ordinal- oder intervallskaliert sein. Sie müssen jedoch numerisch codiert 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 Gesamtscore verhält.

Verwandte Prozeduren. Wenn Sie die Dimensionalität der Skalen-Items untersuchen möchten (um herauszufinden, ob mehr als ein Konstrukt nötig ist, um das Muster der Item-Scores 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Skala** > **Reliabilitätsanalyse...**
2. Wählen Sie mindestens zwei Variablen als potenzielle Komponenten einer additiven Skala aus.
3. Wählen Sie aus dem Dropdown-Listenfeld "Modell" ein Modell aus.

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 Reliabilitätsschätzungen:

- **Alpha-Modelle.** Bei dichotomen Daten entspricht dies dem Kuder-Richardson-20-Koeffizienten (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, Schätzungen der Fehlervarianz, der Gesamtvarianz und der wahren Varianz, geschätzte gemeinsame Inter-Item-Korrelation, geschätzte Reliabilität und unverzerrte Schätzung 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 Cronbach-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. Angezeigt werden der kleinste, der größte und der durchschnittliche Item-Mittelwert, der Bereich und die Varianz der Item-Mittelwerte sowie das Verhältnis zwischen dem größten und dem kleinsten Item-Mittelwert.
- *Varianzen*. 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*. Auswertungsstatistik 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*. Auswertungsstatistik 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 Konkordanzkoeffizienten 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 den Cochran-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 Interaktion besteht.

Intraklassen-Korrelationskoeffizient. Erzeugt ein Maß der Konsistenz oder Werteübereinstimmung innerhalb von Fällen.

- **Modell.** Wählen Sie das Modell für die Berechnung des Intraklassen-Korrelationskoeffizienten aus. Verfügbar sind die Modelle "Zweifach, gemischt", "Zweifach, zufällig" und "Einfach, zufällig". Wählen Sie **Zweifach, gemischt** aus, wenn die Personeneffekte zufällig und die Item-Effekte fest sind. Wählen Sie **Zweifach, zufällig** aus, wenn die Personeneffekte und die Item-Effekte zufällig sind. Wählen Sie **Einfach, zufällig** aus, wenn die Personeneffekte zufällig sind.
- **Typ.** Wählen Sie den Indextyp aus. "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

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Befehlssyntaxreferenz.

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 Bewertungen 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 Bewertungen 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.

Statistik. Für jedes Modell: Datenmatrix, optimal skalierte Datenmatrix, S-Stress (Young), Stress (Kruskal), RSQ, Stimuluskoordinaten, durchschnittlicher Stress und RSQ für jeden Stimulus (RMDS-Modelle). Für Modelle der individuellen Differenzen (INDSCAL): Subjektgewichtungen und Seltsamkeitsindex ("weirdness index") für jedes Subjekt. Für jede Matrix in replizierten Modellen für die multidimensionale Skalierung: Stress und RSQ für jeden Stimulus. Diagramme: Stimuluskoordinaten (zwei- oder dreidimensional), Streudiagramm der Unähnlichkeiten über Distanzen.

Erläuterungen der Daten für die multidimensionale Skalierung

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 *K*-Means-Clusteranalyse ergänzen.

So berechnen Sie eine multidimensionale Skalierung:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren > **Skala** > **Multidimensionale Skalierung...**

2. Wählen Sie für die Analyse mindestens vier numerische Variablen aus.
3. Wählen Sie in der Gruppe "Distanzen" entweder **Daten sind Distanzen** oder **Distanzen aus Daten erstellen** aus.
4. Wenn Sie **Distanzen aus Daten erstellen** auswählen, können Sie für einzelne Matrizen auch eine Gruppierungsvariable auswählen. Die Gruppierungsvariable kann eine numerische Variable oder eine Zeichenfolgevariable sein.

Außerdem sind die folgenden Optionen verfügbar:

- Geben Sie die Form der Distanzmatrix 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

Wenn das aktive Dataset Distanzen innerhalb eines Sets von Objekten oder zwischen zwei Sets 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

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**. Euklidische Distanz, quadrierte euklidische Distanz, Tschebyscheff, Block, Minkowski oder ein benutzerdefiniertes Maß.
- **Häufigkeiten**. Chi-Quadrat-Maß oder Phi-Quadrat-Maß.
- **Binär**. Euklidische Distanz, quadrierte euklidische Distanz, 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

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 Dimensionalität für die Skalierungslösung(en) festlegen. Für jede Zahl im Bereich wird eine Lösung berechnet. Legen Sie ganze Zahlen im Bereich von 1 bis 6 fest. Ein Minimum von 1 ist nur möglich, wenn Sie als Skalierungsmodell **Euklidische Distanz** 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 "Euklidische Distanz" oder "Euklidische Distanz mit individuell gewichteten Differenzen" (auch als INDSCAL bekannt) verfügbar. Beim Modell "Euklidische Distanz mit individuell gewichteten Differenzen" können Sie **Negative Subjektgewichtungen zulassen** auswählen, wenn dies für Ihre Daten geeignet ist.

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 Subjektdiagramme", "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

Die Befehlssyntax ermöglicht außerdem Folgendes:

- 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 Befehlssyntaxreferenz.

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 Gruppierungsvariablen 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.

Statistik. 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.

Erläuterungen der Daten für die Verhältnisstatistik

Daten. Verwenden Sie zum Codieren von Gruppierungsvariablen (nominales oder ordinales Messniveau) numerische Codes oder Zeichenfolgen

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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Deskriptive Statistiken** > **Verhältnis...**
2. Wählen Sie eine Zählervariable.
3. Wählen Sie eine Nennervariable.

Die folgenden Optionen sind verfügbar:

- Wählen Sie eine Gruppierungsvariable 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

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 entspricht dem Prozentsatz von Verhältnissen, 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 Klassifikationsschemas 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.

Statistik. Fläche unter der ROC-Kurve mit Konfidenzintervall und Koordinatenpunkten 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.

Erläuterungen der Daten für ROC-Kurven

Daten. Die Testvariablen sind quantitativ. Die Testvariablen setzen sich oft aus Wahrscheinlichkeiten aus der Diskriminanzanalyse bzw. logistischen Regression zusammen oder sie werden aus Scores auf einer willkürlichen Skala zusammengesetzt, die anzeigen, wie sehr ein Bewerter 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 Bewerters 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:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **ROC-Kurve...**
2. Wählen Sie mindestens eine Wahrscheinlichkeitsvariable für den Test aus.
3. Wählen Sie eine Zustandsvariable aus.
4. Legen Sie den *positiven* Wert für die Zustandsvariable fest.

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 zurzeit keine Auswirkungen auf die Ausgabe.

Testrichtung. 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 nicht parametrische 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.

Kapitel 34. Simulation

Bei Vorhersagemodellen, wie beispielsweise der linearen Regression, ist eine Menge bekannter Eingaben erforderlich, um ein Ergebnis bzw. einen Zielwert vorherzusagen. In vielen Anwendungen in der Praxis sind die Werte der Eingaben jedoch unsicher. Durch die Simulation können Sie die Unsicherheit in den Eingaben für Vorhersagemodelle berücksichtigen und die Wahrscheinlichkeit verschiedener Ausgaben des Modells bei Vorhandensein dieser Unsicherheit einschätzen. Nehmen wir beispielsweise an, Sie verwenden ein Profitmodell, bei dem die Materialkosten als Eingaben verwendet werden, aufgrund von Marktschwankungen besteht jedoch eine gewisse Unsicherheit in Bezug auf diese Kosten. Mithilfe der Simulation können Sie diese Unsicherheit modellieren und ihre Auswirkung auf den Profit bestimmen.

Bei der Simulation in IBM SPSS Statistics wird die Monte-Carlo-Methode verwendet. Unsichere Eingaben werden mit Wahrscheinlichkeitsverteilungen (z. B. Dreiecksverteilung) modelliert und simulierte Werte für diese Eingaben werden durch Ziehen aus diesen Verteilungen generiert. Bei Eingaben mit bekannten Werten werden stets die bekannten Werte verwendet (feste Eingaben). Das Vorhersagemodell wird jeweils mit einem simulierten Wert für jede unsichere Eingabe und mit festen Werten für die bekannten Eingaben ausgewertet, um das Ziel (bzw. die Ziele) des Modells zu berechnen. Dieser Prozess wird viele Male wiederholt (üblicherweise mehrere Zehntausend oder Hunderttausend Mal), was zu einer Verteilung der Zielwerte führt, die zur Beantwortung probabilistischer Fragen verwendet werden kann. Im Rahmen von IBM SPSS Statistics generiert jede Wiederholung des Prozesses einen separaten Fall (Datensatz) von Daten, der aus dem Set der simulierten Werte für die unsicheren Eingaben, den Werten für die festen Eingaben und dem vorhergesagten Ziel (bzw. den vorhergesagten Zielen) des Modells besteht.

Sie können Daten auch ohne Vorhersagemodell simulieren, indem Sie für zu simulierende Variablen Wahrscheinlichkeitsverteilungen angeben. Jeder generierte Fall von Daten besteht aus dem Set simulierter Werte für die angegebenen Variablen.

Zur Ausführung einer Simulation müssen Sie Details angeben, wie beispielsweise das Vorhersagemodell, die Wahrscheinlichkeitsverteilungen für die unsicheren Eingaben, Korrelationen zwischen diesen Eingaben sowie Werte für etwaige feste Eingaben. Nachdem Sie alle Details für eine Simulation angegeben haben, können Sie sie ausführen und die Spezifikationen bei Bedarf in einer **Simulationsplan**-Datei speichern. Sie können den Simulationsplan für andere Benutzer freigeben, die dadurch die Simulation ausführen können, ohne im Detail wissen zu müssen, wie sie erstellt wurde.

Für die Arbeit mit Simulationen stehen zwei Schnittstellen zur Verfügung. Der Simulation Builder ist eine erweiterte Schnittstelle für Benutzer, die Simulationen entwerfen und ausführen. Er stellt alle Funktionen bereit, die zum Entwerfen einer Simulation, zum Speichern der Spezifikationen in einer Simulationsplandatei, zur Angabe der Ausgaben sowie für die Ausführung der Simulation erforderlich sind. Sie können eine Simulation auf der Grundlage einer IBM SPSS-Modelldatei oder einer Menge benutzerdefinierter Gleichungen erstellen, die Sie im Simulation Builder festlegen. Sie können auch einen bestehenden Simulationsplan in den Simulation Builder laden, beliebige Einstellungen ändern und die Simulation ausführen und dabei bei Bedarf den aktualisierten Plan speichern. Bei Benutzern, die einen Simulationsplan besitzen und in erster Linie die Simulation ausführen möchten, steht eine einfachere Schnittstelle zur Verfügung. Mit dieser Schnittstelle können Sie Einstellungen bearbeiten, mit denen Sie die Simulation unter anderen Bedingungen ausführen können, sie bietet jedoch nicht den vollen Funktionsumfang des Simulation Builder für den Entwurf von Simulationen.

Entwerfen einer Simulation auf der Grundlage einer Modelldatei

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Simulation...**
2. Klicken Sie auf **SPSS-Modelldatei auswählen** und klicken Sie dann auf **Weiter**.
3. Öffnen Sie die Modelldatei.

   Die Modelldatei ist eine XML-Datei, die Modell-PMML enthält, die aus IBM SPSS Statistics oder IBM SPSS Modeler erstellt wurde. Weitere Informationen finden Sie im Thema „Registerkarte "Modell"" auf Seite 182.
4. Geben Sie auf der Registerkarte "Simulation" (im Simulation Builder) Wahrscheinlichkeitsverteilungen für simulierte Eingaben und Werte für feste Eingaben an. Wenn das aktive Dataset historische Daten für simulierte Eingaben enthält, klicken Sie auf **Alle anpassen**, um automatisch die am besten angepasste Verteilung für jede dieser Eingaben sowie Korrelationen zwischen diesen Eingaben zu bestimmen. Für jede simulierte Eingabe, die nicht an historische Daten angepasst wird, müssen Sie explizit eine Verteilung angeben, indem Sie einen Verteilungstyp auswählen und die erforderlichen Parameter eingeben.
5. Klicken Sie auf **Ausführen**, um die Simulation auszuführen. Der Simulationsplan, der die Details der Simulation angibt, wird standardmäßig an der in den Speichereinstellungen angegebenen Position gespeichert.

Die folgenden Optionen sind verfügbar:

- Ändern Sie den Speicherort für den Simulationsplan.
- Geben Sie bekannte Korrelationen zwischen simulierten Eingaben an.
- Berechnen Sie eine Kontingenztabelle mit Zuordnungen zwischen kategorialen Eingaben und Verwendung dieser Zuordnungen automatisch, wenn Daten für diese Eingaben generiert werden.
- Geben Sie eine Sensitivitätsanalyse zur Untersuchung des Effekts an, der durch Variieren des Werts einer festen Eingabe oder durch Variieren eines Verteilungsparameters für eine simulierte Eingabe erzeugt wird.
- Geben Sie erweiterte Optionen an, wie die Festlegung der maximalen Anzahl der zu generierenden Fälle oder die Anforderung einer Stichprobenziehung aus der Flanke.
- Passen Sie die Ausgabe an.
- Speichern Sie die simulierten Daten in einer Datendatei.

Entwerfen einer Simulation auf der Grundlage benutzerdefinierter Gleichungen

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Simulation...**
2. Klicken Sie auf **Gleichungen eintippen** und klicken Sie dann auf **Weiter**.
3. Klicken Sie auf der Registerkarte "Modell" (im Simulation Builder) auf **Neue Gleichung**, um die einzelnen Gleichungen in Ihrem Vorhersagemodell zu definieren.
4. Klicken Sie auf die Registerkarte "Simulation" und geben Sie Wahrscheinlichkeitsverteilungen für simulierte Eingaben und Werte für feste Eingaben an. Wenn das aktive Dataset historische Daten für simulierte Eingaben enthält, klicken Sie auf **Alle anpassen**, um automatisch die am besten angepasste Verteilung für jede dieser Eingaben sowie Korrelationen zwischen diesen Eingaben zu bestimmen. Für jede simulierte Eingabe, die nicht an historische Daten angepasst wird, müssen Sie explizit eine Verteilung angeben, indem Sie einen Verteilungstyp auswählen und die erforderlichen Parameter eingeben.
5. Klicken Sie auf **Ausführen**, um die Simulation auszuführen. Der Simulationsplan, der die Details der Simulation angibt, wird standardmäßig an der in den Speichereinstellungen angegebenen Position gespeichert.

Die folgenden Optionen sind verfügbar:

- Ändern Sie den Speicherort für den Simulationsplan.
- Geben Sie bekannte Korrelationen zwischen simulierten Eingaben an.
- Berechnen Sie eine Kontingenztabelle mit Zuordnungen zwischen kategorialen Eingaben und Verwendung dieser Zuordnungen automatisch, wenn Daten für diese Eingaben generiert werden.
- Geben Sie eine Sensitivitätsanalyse zur Untersuchung des Effekts an, der durch Variieren des Werts einer festen Eingabe oder durch Variieren eines Verteilungsparameters für eine simulierte Eingabe erzeugt wird.
- Geben Sie erweiterte Optionen an, wie die Festlegung der maximalen Anzahl der zu generierenden Fälle oder die Anforderung einer Stichprobenziehung aus der Flanke.
- Passen Sie die Ausgabe an.
- Speichern Sie die simulierten Daten in einer Datendatei.

Entwerfen einer Simulation ohne Vorhersagemodell

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Simulation...**
2. Klicken Sie auf **Simulierte Daten erstellen** und klicken Sie dann auf **Weiter**.
3. Wählen sie Sie auf der Registerkarte "Modell" (im Simulation Builder) die Felder aus, die Sie simulieren wollen. Sie können Felder aus dem aktiven Dataset auswählen oder Sie können neue Felder definieren, indem Sie auf **Neu** klicken.
4. Klicken Sie auf die Registerkarte "Simulation" und geben Sie Wahrscheinlichkeitsverteilungen für die zu simulierenden Felder an. Wenn das aktive Dataset historische Daten für eines der Felder enthält, klicken Sie auf **Alle anpassen**, um automatisch die am besten angepasste Verteilung sowie Korrelationen zwischen den Feldern zu bestimmen. Für Felder, die nicht an historische Daten angepasst werden, müssen Sie explizit eine Verteilung angeben, indem Sie einen Verteilungstyp auswählen und die erforderlichen Parameter eingeben.
5. Klicken Sie auf **Ausführen**, um die Simulation auszuführen. Die simulierten Daten werden standardmäßig in dem neuen in den Speichereinstellungen angegebenen Dataset gespeichert. Darüber hinaus wird der Simulationsplan, der die Details der Simulation angibt, standardmäßig an der in den Speichereinstellungen angegebenen Position gespeichert.

Die folgenden Optionen sind verfügbar:

- Ändern Sie den Speicherort für die simulierten Daten oder den gespeicherten Simulationsplan.
- Geben Sie bekannte Korrelationen zwischen simulierten Feldern an.
- Berechnen Sie eine Kontingenztabelle mit Zuordnungen zwischen kategorialen Feldern und Verwendung dieser Zuordnungen automatisch, wenn Daten für diese Felder generiert werden.
- Geben Sie eine Sensitivitätsanalyse zur Untersuchung des Effekts an, der durch Variieren eines Verteilungsparameters für ein simuliertes Feld Eingabe erzeugt wird.
- Geben Sie erweiterte Optionen wie die Festlegung der Anzahl der zu generierenden Fälle an.

Ausführen einer Simulation über einen Simulationsplan

Für die Ausführung einer Simulation über einen Simulationsplan stehen zwei Optionen zur Verfügung. Sie können das Dialogfeld "Simulation ausführen" verwenden, das hauptsächlich für die Ausführung über einen Simulationsplan gedacht ist, oder Sie können den Simulation Builder verwenden.

So verwenden Sie das Dialogfeld "Simulation ausführen":

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Analysieren > **Simulation...**
2. Klicken Sie auf **Bestehenden Simulationsplan öffnen**.

3. Stellen Sie sicher, dass das Kontrollkästchen **In Simulation Builder öffnen** nicht aktiviert ist, und klicken Sie auf **Weiter**.
4. Öffnen Sie den Simulationsplan.
5. Klicken Sie im Dialogfeld "Simulation ausführen" auf **Ausführen**.

Zum Ausführen der Simulation über den Simulation Builder gehen Sie wie folgt vor:

1. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Analysieren > **Simulation...**
2. Klicken Sie auf **Bestehenden Simulationsplan öffnen**.
3. Aktivieren Sie das Kontrollkästchen **In Simulation Builder öffnen** und klicken Sie auf **Weiter**.
4. Öffnen Sie den Simulationsplan.
5. Nehmen Sie alle gewünschten Änderungen an den Einstellungen auf der Registerkarte "Simulation" vor.
6. Klicken Sie auf **Ausführen**, um die Simulation auszuführen.

Optional können Sie folgende Aktionen ausführen:

- Einrichten oder Ändern einer Sensitivitätsanalyse zur Untersuchung des Effekts, der durch Variieren des Werts einer festen Eingabe bzw. durch Variieren eines Verteilungsparameters für eine simulierte Eingabe erzeugt wird
- Erneutes Anpassen von Verteilungen und Korrelationen für simulierte Eingaben an neue Daten
- Ändern der Verteilung für eine simulierte Eingabe
- Passen Sie die Ausgabe an.
- Speichern Sie die simulierten Daten in einer Datendatei.

Simulation Builder

Der Simulation Builder stellt alle Funktionen zum Entwerfen und Ausführen von Simulationen bereit. Er ermöglicht Ihnen die Ausführung folgender allgemeiner Aufgaben:

- Entwerfen und Ausführen einer Simulation für ein in einer PMML-Modelldatei definiertes IBM SPSS-Modell
- Entwerfen und Ausführen einer Simulation für ein Vorhersagemodell, das durch von Ihnen angegebene benutzerdefinierte Gleichungen definiert wurde
- Entwerfen und Ausführen einer Simulation, die ohne Vorhersagemodell Daten generiert.
- Ausführen einer Simulation auf der Grundlage eines bestehenden Simulationsplans, optional mit Änderungen an den Planeinstellungen.

Registerkarte "Modell"

Für auf Vorhersagemodellen basierende Simulationen gibt die Registerkarte "Modell" die Quelle des Modells an. Für Simulationen, die kein Vorhersagemodell umfassen, gibt die Registerkarte "Modell" die Felder an, die simuliert werden sollen.

SPSS-Modelldatei auswählen. Diese Option gibt an, dass das Vorhersagemodell in einer IBM SPSS-Modelldatei definiert ist. Eine IBM SPSS-Modelldatei ist eine XML-Datei, die Modell-PMML enthält, die aus IBM SPSS Statistics oder IBM SPSS Modeler erstellt wurde. Vorhersagemodelle werden durch Prozeduren, wie beispielsweise lineare Regression und Entscheidungsbäume in IBM SPSS Statistics, erstellt und können in eine Modelldatei exportiert werden. Sie können eine andere Modelldatei verwenden, indem Sie auf **Durchsuchen** klicken und zu der gewünschten Datei navigieren.

Von Simulation unterstützte PMML-Modelle

- Lineare Regression

- Verallgemeinertes lineares Modell
- Allgemeines lineares Modell
- Binäre logistische Regression
- Multinomiale logistische Regression
- Ordinal-multinomiale Regression
- Cox-Regression
- Baum
- Verstärkter Baum (C5.0)
- Diskriminanz
- Two-Step-Clusteranalyse
- K-Means-Clusteranalyse
- Neuronales Netz
- Regelset (Entscheidungsliste)

Anmerkung:
- PMML-Modelle mit mehreren Zielfeldern (Variablen) bzw. Aufteilungen werden für die Verwendung bei einer Simulation nicht unterstützt.
- Werte von Zeichenfolgeeingaben für binäre logistische Regressionsmodelle sind im Modell auf 8 Byte begrenzt. Wenn Sie diese Eingabezeichenfolgen an das aktive Dataset anpassen, müssen Sie sicherstellen, dass die Werte in den Daten maximal 8 Byte lang sind. Datenwerte, die länger als 8 Byte sind, werden aus der zugehörigen kategorialen Verteilung für die Eingabe ausgeschlossen und werden in der Ausgabetabelle "Kategorien ohne Entsprechung" als ohne Entsprechung angezeigt.

Gleichungen für das Modell eintippen. Diese Option gibt an, dass das Vorhersagemodell aus einer oder mehreren benutzerdefinierten Gleichungen besteht, die von Ihnen erstellt werden müssen. Erstellen Sie Gleichungen, indem Sie auf **Neue Gleichung** klicken. Dadurch wird der Gleichungseditor geöffnet. Sie können bestehende Gleichungen bearbeiten, sie zur Verwendung als Vorlagen für neue Gleichungen kopieren, sie neu anordnen oder löschen.
- Der Simulation Builder unterstützt keine Systeme mit simultanen Gleichungen oder Gleichungen, die in der Zielvariablen nicht linear sind.
- Benutzerdefinierte Gleichungen werden in der Reihenfolge ihrer Angabe ausgewertet. Wenn die Gleichung für ein bestimmtes Ziel von einem anderen Ziel abhängt, muss das andere Ziel durch eine vorangehende Gleichung definiert sein.

  Beispielsweise hängt bei den unten stehenden drei Gleichungen die Gleichung für *Gewinn* von den Werten für *Einnahmen* und *Ausgaben* ab, sodass die Gleichungen für *Einnahmen* und *Ausgaben* der Gleichung für *Gewinn* vorangehen müssen.

```
Einnahmen = Preis * Volumen
Ausgaben = feste + Volumen * (Stückkosten_Material + Stückkosten_Arbeit)
Gewinn = Einnahmen - Ausgaben
```

Simulierte Daten ohne Modell erstellen. Wählen Sie diese Option aus, um Daten ohne Vorhersagemodell zu simulieren. Geben Sie die zu simulierenden Felder an, indem Sie Felder aus dem aktiven Dataset auswählen oder auf **Neu** klicken, um neue Felder zu definieren.

Gleichungseditor

Mit dem Gleichungseditor können Sie eine benutzerdefinierte Gleichung für Ihr Vorhersagemodell erstellen oder bearbeiten.
- Der Ausdruck für die Gleichung kann Felder aus dem aktiven Dataset oder neue Eingabefelder enthalten, die Sie im Gleichungseditor definieren.
- Sie können Eigenschaften des Ziels angeben, beispielsweise das Messniveau, die Wertbeschriftungen und ob Ausgaben für das Ziel generiert werden.

- Sie können Ziele aus zuvor definierten Gleichungen als Eingaben für die aktuelle Gleichung verwenden und so gekoppelte Gleichungen erstellen.
- Sie können einen beschreibenden Kommentar zu der Gleichung hinzufügen. Die Kommentare werden zusammen mit der Gleichung auf der Registerkarte "Modell" angezeigt.

1. Geben Sie den Namen des Ziels ein. Klicken Sie optional unter dem Textfeld "Ziel" auf **Bearbeiten**, um das Dialogfeld "Definierte Eingaben" zu öffnen, in dem Sie die Standardeigenschaften des Ziels ändern können.
2. Um einen Ausdruck zu erstellen, fügen Sie Komponenten in das Feld "Numerischer Ausdruck" ein oder geben den Ausdruck direkt in dieses Feld ein.

- Sie können Ihren Ausdruck mithilfe von Feldern aus dem aktiven Dataset erstellen oder Sie können neue Eingaben definieren, indem Sie auf **Neu** klicken. Dadurch wird das Dialogfeld "Eingaben definieren" geöffnet.
- Sie können Funktionen einfügen, indem Sie eine Gruppe aus der Liste "Funktionsgruppe" auswählen und in der Liste "Funktionen" auf die Funktion doppelklicken (oder die Funktion auswählen und auf den Pfeil neben der Liste "Funktionsgruppe" klicken). Geben Sie alle durch Fragezeichen gekennzeichneten Parameter ein. Die Funktionsgruppe mit der Beschriftung **Alle** bietet eine Auflistung aller verfügbaren Funktionen. Eine kurze Beschreibung der aktuell ausgewählten Funktion wird in einem speziellen Bereich des Dialogfelds angezeigt.
- Zeichenfolgekonstanten müssen in Anführungszeichen eingeschlossen werden.
- Wenn die Werte Dezimalstellen enthalten, muss ein Punkt (.) als Dezimaltrennzeichen verwendet werden.

Hinweis: Simulation unterstützt keine benutzerdefinierten Gleichungen mit Zeichenfolgezielen.

Definierte Eingaben: Im Dialogfeld "Definierte Eingaben" können Sie neue Eingaben definieren und Eigenschaften für Ziele festlegen.

- Wenn eine Eingabe, die in einer Gleichung verwendet werden soll, im aktiven Dataset nicht vorhanden ist, müssen Sie sie definieren, um sie in der Gleichung verwenden zu können.
- Wenn Sie Daten ohne Vorhersagemodell simulieren, müssen Sie alle simulierten Eingaben definieren, die im aktiven Dataset nicht vorhanden sind.

Name. Geben Sie den Namen für ein Ziel oder eine Eingabe an.

Ziel. Sie können das Messniveau eines Ziels angeben. Die Standardeinstellung für das Messniveau ist "stetig". Außerdem können Sie angeben, ob Ausgaben für dieses Ziel erstellt werden sollen. Bei einem Set gekoppelter Gleichungen sind Sie beispielsweise vielleicht nur an Ausgaben aus dem Ziel der letzten Gleichung interessiert und unterdrücken daher die Ausgaben aus den anderen Zielen.

Eingabe wird simuliert. Hiermit wird angegeben, dass die Werte der Eingabe gemäß einer angegebenen Wahrscheinlichkeitsverteilung simuliert werden (die Wahrscheinlichkeitsverteilung wird auf der Registerkarte "Simulation" angegeben). Das Messniveau legt fest, welche Verteilungen standardmäßig berücksichtigt werden, wenn nach der am besten angepassten Verteilung für die Eingabe gesucht wird (durch Klicken auf **Anpassung** bzw. **Alle anpassen** auf der Registerkarte "Simulation"). Beispielsweise wird bei einem stetigen Messniveau die Normalverteilung (geeignet für stetige Daten) berücksichtigt, nicht jedoch die Binomialverteilung.

Anmerkung: Wählen Sie ein Messniveau des Typs "Zeichenfolge" für Zeichenfolgeeingaben aus. Zu simulierende Zeichenfolgeeingaben sind auf die kategoriale Verteilung beschränkt.

Fester Wert als Eingabe. Dadurch wird angegeben, dass der Wert der Eingabe bekannt ist und stets dieser Wert verwendet wird. Feste Eingaben können vom Typ her numerisch oder Zeichenfolgen sein. Geben Sie einen Wert für die feste Eingabe an. Zeichenfolgewerte sollten nicht in Anführungszeichen eingeschlossen werden.

Wertbeschriftungen. Sie können Wertbeschriftungen für Ziele, simulierte Eingaben und feste Eingaben angeben. Wertbeschriftungen werden in Ausgabediagrammen und -tabellen verwendet.

Registerkarte "Simulation"

Auf der Registerkarte "Simulation" werden, abgesehen vom Vorhersagemodell, alle Eigenschaften der Simulation angegeben. Auf der Registerkarte "Simulation" können Sie folgende allgemeine Aufgaben ausführen:

- Angabe von Wahrscheinlichkeitsverteilungen für simulierte Eingaben und von Werten für feste Eingaben.
- Angabe von Korrelationen zwischen simulierten Eingaben. Bei kategorialen Eingaben können Sie angeben, dass zwischen diesen Eingaben im aktiven Dataset bestehende Zuordnungen verwendet werden sollen, wenn Daten für die Eingaben generiert werden.
- Angabe erweiterter Optionen, wie beispielsweise Stichprobenziehung aus der Flanke und Kriterien zur Anpassung von Verteilungen an historische Daten.
- Passen Sie die Ausgabe an.
- Angabe des Speicherorts für den Simulationsplan und optionale Speicherung der simulierten Daten.

Simulierte Felder

Um eine Simulation ausführen zu können, müssen die einzelnen Eingabefelder als fest oder simuliert angegeben werden. Simulierte Eingaben sind Eingaben, deren Werte unsicher sind und die durch Stichprobenziehung aus einer angegebenen Wahrscheinlichkeitsverteilung generiert werden. Wenn historische Daten für die Eingaben, die Sie simulieren möchten, verfügbar sind, können Sie automatisch die am besten angepassten Verteilungen ermitteln sowie Korrelationen zwischen diesen Eingaben bestimmen. Sie können die Verteilungen bzw. Korrelationen auch manuell angeben, wenn keine historischen Daten verfügbar sind oder Sie bestimmte Verteilungen oder Korrelationen benötigen.

Feste Eingaben sind Eingaben, deren Werte bekannt sind und die für jeden in der Simulation generierten Fall konstant bleiben. Nehmen wir beispielsweise an, Sie verfügen über ein lineares Regressionsmodell für die Umsätze als Funktion einer Reihe von Eingaben, wie dem Preis, und möchten den Preis beim aktuellen Marktpreis konstant halten. In diesem Fall geben Sie den Preis als feste Eingabe ein.

Für Simulationen, die auf Vorhersagemodellen basieren, ist jeder Prädiktor im Modell ein Eingabefeld für die Simulation. Für Simulationen, die kein Vorhersagemodell umfassen, stellen die auf der Registerkarte "Modell" angegebenen Felder die Eingaben für die Simulation dar.

Automatische Anpassung von Verteilungen und Berechnung von Korrelationen für simulierte Eingaben. Wenn das aktive Dataset historische Daten für die Eingaben enthält, die Sie simulieren möchten, können Sie automatisch die am besten angepassten Verteilungen für diese Eingaben ermitteln sowie Korrelationen zwischen diesen Eingaben bestimmen. Gehen Sie dazu wie folgt vor:

1. Prüfen Sie, ob alle zu simulierenden Eingaben jeweils dem richtigen Feld im aktiven Dataset zugeordnet sind. Die Eingaben sind in der Spalte "Eingabe" aufgeführt und in der Spalte "Anpassen an" wird das zugeordnete Feld im aktiven Dataset angezeigt. Sie können eine Eingabe einem anderen Feld im aktiven Dataset zuordnen, indem Sie in der Dropdown-Liste "Anpassen an" einen anderen Eintrag auswählen.

   Der Wert *-Ohne-* in der Spalte "Anpassen an" gibt an, dass die Eingabe nicht automatisch einem Feld im aktiven Dataset zugeordnet werden konnte. Standardmäßig werden Eingaben je nach Name, Messniveau und Typ (numerisch oder Zeichenfolge) Datasetfeldern zugeordnet. Wenn das aktive Dataset keine historischen Daten für die Eingabe enthält, geben Sie die Verteilung für die Eingabe manuell an oder geben Sie die Eingabe als feste Eingabe an, wie unten beschrieben.
2. Klicken Sie auf **Alle anpassen**.

Die am besten angepasste Verteilung und die damit verknüpften Parameter werden zusammen mit einer grafischen Darstellung der Verteilung, die über ein Histogramm (oder Balkendiagramm) der historischen Daten gelegt ist, angezeigt. Korrelationen zwischen simulierten Eingaben werden in den Korrelationsein-

stellungen angezeigt. Sie können die Anpassungsergebnisse untersuchen und die automatische Verteilungsanpassung für eine bestimmte Eingabe individuell gestalten, indem Sie die Zeile für die Eingabe auswählen und auf **Anpassungsdetails** klicken. Weitere Informationen finden Sie im Thema „Anpassungsdetails" auf Seite 188.

Sie können eine automatische Verteilungsanpassung für eine bestimmte Eingabe ausführen, indem Sie die Zeile für die Eingabe auswählen und auf **Anpassung** klicken. Es werden auch automatisch Korrelationen für alle simulierten Eingaben, die mit Feldern im aktiven Dataset übereinstimmen, berechnet.

Anmerkung:

- Fälle mit fehlenden Werten für simulierte Eingaben werden von der Verteilungsanpassung, der Berechnung von Korrelationen und der Berechnung der optionalen Kontingenztabelle ausgeschlossen (bei Eingaben mit kategorialer Verteilung). Sie können optional angeben, ob benutzerdefiniert fehlende Werte für Eingaben mit einer kategorialen Verteilung als gültig behandelt werden. Standardmäßig werden sie als fehlend behandelt. Weitere Informationen finden Sie im Thema „Erweiterte Optionen" auf Seite 190.
- Wenn bei stetigen und ordinalen Eingaben für keine der getesteten Verteilungen eine akzeptable Anpassung gefunden wird, wird die empirische Verteilung als bestmögliche Anpassung vorgeschlagen. Bei stetigen Eingaben ist die empirische Verteilung die kumulative Verteilungsfunktion der historischen Daten. Bei ordinalen Eingaben ist die empirische Verteilung die kategoriale Verteilung der historischen Daten.

Manuelle Angabe von Verteilungen. Sie können die Wahrscheinlichkeitsverteilung für jede beliebige simulierte Eingabe manuell angeben, indem Sie die Verteilung aus der Dropdown-Liste **Typ** auswählen und die Verteilungsparameter in das Parameterraster eingeben. Nachdem Sie die Parameter für eine Verteilung eingegeben haben, wird ein Beispieldiagramm der Verteilung auf der Grundlage der angegebenen Parameter neben dem Parameterraster angezeigt. Hier einige Anmerkungen zu bestimmten Verteilungen:

- **Kategorial.** Die kategoriale Verteilung beschreibt ein Eingabefeld, das eine feste Anzahl von Werten, sogenannten Kategorien, aufweist. Jeder Kategorie ist eine Wahrscheinlichkeit zugeordnet, dergestalt, dass die Summe der Wahrscheinlichkeiten über alle Kategorien gleich 1 ist. Klicken Sie zur Eingabe einer Kategorie auf die linke Spalte im Parameterraster und geben Sie den Kategoriewert an. Geben Sie in der rechten Spalte die Wahrscheinlichkeit ein, die der Kategorie zugeordnet ist.

  Anmerkung: Kategoriale Eingaben aus einem PMML-Modell weisen Kategorien auf, die durch das Modell festgelegt sind und nicht geändert werden können.
- **Negativ binomial – Fehler.** Beschreibt die Verteilung der Anzahl der Fehlversuche in einer Abfolge von Versuchen, bevor eine angegebene Anzahl von Erfolgen beobachtet wird. Der Parameter *thresh* ist die angegebene Anzahl von Erfolgen und der Parameter *prob* ist die Erfolgswahrscheinlichkeit für den jeweiligen Versuch.
- **Negativ binomial – Versuche.** Beschreibt die Verteilung der Anzahl von Versuchen, die erforderlich ist, bevor eine angegebene Anzahl von Erfolgen beobachtet wird. Der Parameter *thresh* ist die angegebene Anzahl von Erfolgen und der Parameter *prob* ist die Erfolgswahrscheinlichkeit für den jeweiligen Versuch.
- **Bereich.** Diese Verteilung besteht aus einem Set von Intervallen, denen jeweils eine Wahrscheinlichkeit zugewiesen ist, sodass die Summe der Wahrscheinlichkeiten über alle Intervalle hinweg gleich 1 ist. Die Werte innerhalb eines Intervalls werden jeweils aus einer für dieses Intervall definierten Gleichverteilung gezogen. Die Intervalle werden durch Eingabe eines Mindest- und Höchstwerts und einer zugeordneten Wahrscheinlichkeit angegeben.

  Nehmen wir beispielsweise an, Sie glauben, dass die Kosten für einen Rohstoff mit einer Wahrscheinlichkeit von 40 % im Bereich von 10 bis 15 Euro pro Einheit liegen und mit einer Wahrscheinlichkeit von 60 % im Bereich von 15 bis 20 Euro pro Einheit. Die Kosten könnten mit einer Bereichsverteilung modelliert werden, die aus den beiden Intervallen [10–15] und [15–20] besteht, wobei die dem ersten Intervall zugeordnete Wahrscheinlichkeit auf 0,4 und die Wahrscheinlichkeit für das zweite Intervall

auf 0,6 gesetzt wird. Die Intervalle müssen nicht aneinander angrenzen und sie können sich sogar überschneiden. Sie könnten also beispielsweise auch die Intervalle 10 bis 15 und 20 bis 25 Euro oder 10 bis 15 und 13 bis 16 Euro angeben.

- **Weibull.** Der Parameter *c* ist ein optionaler Lageparameter, der angibt, wo sich der Ursprung der Verteilung befindet.

Die Parameter für die folgenden Verteilungen haben dieselbe Bedeutung wie in den zugehörigen Funktionen für Zufallsvariablen, die im Dialogfeld "Variable berechnen" verfügbar sind: Bernoulli, Beta, Binomial, Exponentiell, Gamma, Lognormal, Negativ Binomial (Versuche und Fehler), Normal, Poisson und Gleichverteilung.

Angabe fester Eingaben. Sie können eine feste Eingabe angeben, indem Sie den Wert "Fest" aus der Dropdown-Liste **Typ** in der Spalte "Verteilung" auswählen und den festen Wert eingeben. Es kann sich um einen numerischen Wert oder einen Zeichenfolgewert handeln, je nachdem, ob die Eingabe numerisch oder eine Zeichenfolge ist. Zeichenfolgewerte sollten nicht in Anführungszeichen eingeschlossen werden.

Angabe von Grenzen für simulierte Werte. Die meisten Verteilungen unterstützen die Angabe von Ober- und Untergrenzen für die simulierten Werte. Zur Angabe einer Untergrenze geben Sie einen Wert in das Textfeld **Min** ein und zur Angabe einer Obergrenze geben Sie einen Wert in das Textfeld **Max** ein.

Eingaben sperren. Durch Sperren einer Eingabe (durch Aktivieren des Kontrollkästchens in der Spalte mit dem Schlosssymbol) wird die Eingabe von der automatischen Verteilungsanpassung ausgeschlossen. Dies ist besonders dann nützlich, wenn Sie eine Verteilung oder einen festen Wert manuell angeben und sicherstellen wollen, dass diese nicht durch die automatische Verteilungsanpassung beeinträchtigt werden. Sperren ist auch sinnvoll, wenn Sie vorhaben, Ihren Simulationsplan für andere Benutzer freizugeben, die ihn im Dialogfeld "Simulation ausführen" verwenden, und etwaige Änderungen an bestimmten Eingaben verhindern wollen. Spezifikationen für gesperrte Eingaben können im Dialogfeld "Simulation ausführen" nicht geändert werden.

Sensitivitätsanalyse. Mit der Sensitivitätsanalyse können Sie den Effekt systematischer Änderungen in einer festen Eingabe oder in einem Verteilungsparameter für eine simulierte Eingabe untersuchen, indem Sie ein unabhängiges Set simulierter Fälle – also im Grunde eine separate Simulation – für jeden angegebenen Wert generieren. Zur Angabe der Sensitivitätsanalyse wählen Sie eine feste oder simulierte Eingabe aus und klicken Sie auf **Sensitivitätsanalyse**. Die Sensitivitätsanalyse ist auf eine einzelne feste Eingabe oder einen einzelnen Verteilungsparameter für eine simulierte Eingabe beschränkt. Weitere Informationen finden Sie im Thema „Sensitivitätsanalyse” auf Seite 189.

Symbole für den Anpassungsstatus

Symbole in der Spalte "Anpassen an" geben den Anpassungsstatus für die einzelnen Eingabefelder an.

Tabelle 3. Statussymbole.

Symbol	Beschreibung
	Für die Eingabe wurde keine Verteilung angegeben und die Eingabe wurde auch nicht als feste Eingabe angegeben. Um die Simulation ausführen zu können, müssen Sie entweder eine Verteilung für diese Eingabe angeben oder sie als feste Eingabe definieren und den festen Wert angeben.
	Die Eingabe wurde zuvor an ein Feld angepasst, das im aktiven Dataset (aktives Dataset) nicht vorhanden ist. Es sind keine Maßnahmen erforderlich, es sei denn, Sie möchten eine Neuanpassung der Verteilung für die Eingabe an das aktive Dataset durchführen.
	Die am besten angepasste Verteilung wurde durch eine alternative Verteilung aus dem Dialogfeld "Anpassungsdetails" ersetzt.

Tabelle 3. Statussymbole (Forts.).

Symbol	Beschreibung
	Die Eingabe ist auf die am besten angepasste Verteilung gesetzt.
	Die Verteilung wurde manuell angegeben oder es wurden Iterationen der Sensitivitätsanalyse für diese Eingabe angegeben.

Anpassungsdetails: Im Dialogfeld "Anpassungsdetails" werden die Ergebnisse der automatischen Verteilungsanpassung für eine bestimmte Eingabe angezeigt. Die Verteilungen sind nach Anpassungsgüte sortiert, beginnend mit der am besten angepassten Verteilung. Sie können die am besten angepasste Verteilung überschreiben, indem Sie das Optionsfeld für die gewünschte Verteilung in der Spalte "Verwenden" auswählen. Durch die Auswahl eines Optionsfelds in der Spalte "Verwenden" wird außerdem eine grafische Darstellung der Verteilung, die über ein Histogramm (oder Balkendiagramm) der historischen Daten für die betreffende Eingabe gelegt ist, angezeigt.

Anpassungsstatistik. Standardmäßig und für stetige Felder wird der Anderson-Darling-Test zur Ermittlung der Anpassungsgüte verwendet. Alternativ können Sie (nur für stetige Felder) den Kolmogorow-Smirnow-Test für die Anpassungsgüte verwenden, indem Sie diese Option in den Einstellungen unter "Erweiterte Optionen" angeben. Für stetige Eingaben werden die Ergebnisse beider Tests in der Spalte "Anpassungsstatistik" ("A" für "Anderson-Darling" und "K" für "Kolmogorow-Smirnow") angezeigt, wobei der ausgewählte Test zur Sortierung der Verteilungen dient. Für ordinale und nominale Eingaben wird der Chi-Quadrat-Test verwendet. Die dem Test zugeordneten p-Werte werden ebenfalls angezeigt.

Parameter. Die den einzelnen angepassten Verteilungen zugeordneten Verteilungsparameter werden in der Spalte "Parameter" angezeigt. Die Parameter für die folgenden Verteilungen haben dieselbe Bedeutung wie in den zugehörigen Funktionen für Zufallsvariablen, die im Dialogfeld "Variable berechnen" verfügbar sind: Bernoulli, Beta, Binomial, Exponentiell, Gamma, Lognormal, Negativ Binomial (Versuche und Fehler), Normal, Poisson und Gleichverteilung.

Weitere Informationen finden Sie im Thema . Bei der kategorialen Beschreibung sind die Parameternamen die Kategorien und die Parameterwerte sind die zugeordneten Wahrscheinlichkeiten.

Erneute Anpassung mit einem benutzerdefinierten Verteilungsset. Standardmäßig wird das Messniveau der Eingabe verwendet, um zu bestimmen, welche Verteilungen bei der automatischen Verteilungsanpassung berücksichtigt werden. Stetige Verteilungen, wie "lognormal" und "gamma", werden beispielsweise bei der Anpassung einer stetigen Eingabe berücksichtigt, nicht jedoch diskrete Verteilungen, wie "Poisson" und "binomial". Sie können ein Subset der Standardverteilungen auswählen, indem Sie die Verteilungen in der Spalte "Neu anpassen" auswählen. Sie können auch das als Standard vorgegebene Verteilungsset außer Kraft setzen, indem Sie in der Dropdown-Liste **Behandeln als (Messniveau)** ein anderes Messniveau auswählen und die Verteilungen in der Spalte "Neu anpassen" auswählen. Klicken Sie auf **Neuanpassung ausführen**, um eine Neuanpassung mit dem benutzerdefinierten Verteilungsset durchzuführen.

Anmerkung:

- Fälle mit fehlenden Werten für simulierte Eingaben werden von der Verteilungsanpassung, der Berechnung von Korrelationen und der Berechnung der optionalen Kontingenztabelle ausgeschlossen (bei Eingaben mit kategorialer Verteilung). Sie können optional angeben, ob benutzerdefiniert fehlende Werte für Eingaben mit einer kategorialen Verteilung als gültig behandelt werden. Standardmäßig werden sie als fehlend behandelt. Weitere Informationen finden Sie im Thema „Erweiterte Optionen” auf Seite 190.
- Wenn bei stetigen und ordinalen Eingaben für keine der getesteten Verteilungen eine akzeptable Anpassung gefunden wird, wird die empirische Verteilung als bestmögliche Anpassung vorgeschlagen.

Bei stetigen Eingaben ist die empirische Verteilung die kumulative Verteilungsfunktion der historischen Daten. Bei ordinalen Eingaben ist die empirische Verteilung die kategoriale Verteilung der historischen Daten.

Sensitivitätsanalyse: Mit der Sensitivitätsanalyse können Sie den Effekt untersuchen, der durch Variieren einer festen Eingabe oder durch Variieren eines Verteilungsparameters für eine simulierte Eingabe über einer angegebenen Menge an Werten erzeugt wird. Für jeden angegebenen Wert wird ein unabhängiges Set simulierter Fälle (also im Grunde eine separate Simulation) erzeugt, wodurch Sie den Effekt der Eingabevariation untersuchen können. Die einzelnen Sets an simulierten Fällen werden als **Iteration** bezeichnet.

Iterieren. Mit dieser Option können Sie das Werteset angeben, über das die Eingabe variiert werden soll.

- Wenn Sie den Wert eines Verteilungsparameters variieren, wählen Sie den Parameter aus der Dropdown-Liste aus. Geben Sie das Werteset in das Raster "Parameterwert in Abhängigkeit von der Iteration" ein. Durch Klicken auf **Weiter** werden die angegebenen Werte zum Parameterraster für die zugeordnete Eingabe hinzugefügt, mit einem Index, der die Iterationsnummer des Werts angibt.
- Für die Verteilungstypen "Kategorial" und "Bereich" können die Wahrscheinlichkeitswerte der Kategorien bzw. Intervalle variiert werden, nicht jedoch die Werte der Kategorien und die Endpunkte der Intervalle. Wählen Sie eine Kategorie oder ein Intervall aus der Dropdown-Liste aus und geben Sie das Set der Wahrscheinlichkeitswerte im Raster "Parameterwert in Abhängigkeit von der Iteration" ein. Die Wahrscheinlichkeitswerte für die anderen Kategorien bzw. Intervalle werden automatisch angepasst.

Keine Iterationen. Verwenden Sie diese Option, um die Iterationen für eine Eingabe abzubrechen. Durch Klicken auf **Weiter** werden die Iterationen entfernt.

Korrelationen

Zwischen zu simulierenden Eingabefeldern liegen bekanntlich häufig Korrelationen vor, beispielsweise zwischen Größe und Gewicht. Korrelationen zwischen zu simulierenden Eingaben müssen berücksichtigt werden, um sicherzustellen, dass diese Korrelationen in den simulierten Werten beibehalten werden.

Korrelationen bei der Anpassung neu berechnen. Diese Option gibt an, dass Korrelationen zwischen simulierten Eingaben automatisch berechnet werden, wenn Verteilungen über die Aktionen **Alle anpassen** bzw. **Anpassung** in den Einstellungen für simulierte Felder an das aktive Dataset angepasst werden.

Korrelationen bei der Anpassung nicht neu berechnen. Wählen Sie diese Option, wenn Sie Korrelationen manuell angeben und verhindern möchten, dass sie bei der automatischen Anpassung von Verteilungen an das aktive Dataset überschrieben werden. Die im Korrelationsraster eingegebenen Werte müssen zwischen -1 und 1 liegen. Der Wert 0 gibt an, dass keine Korrelation zwischen dem zugehörigen Eingabenpaar besteht.

Zurücksetzen. Dadurch werden alle Korrelationen auf 0 zurückgesetzt.

Angepasste Mehrwegekontingenztabellen für Eingaben mit einer kategorialen Verteilung verwenden. Für Eingaben mit einer kategorialen Verteilung können Sie aus dem aktiven Dataset automatisch eine Mehrwegekontingenztabelle berechnen, die die Zuordnungen zwischen diesen Eingaben beschreibt. Die Kontingenztabelle wird dann verwendet, wenn Daten für diese Eingaben generiert werden. Wenn Sie den Simulationsplan speichern wollen, wird die Kontingenztabelle in der Plandatei gespeichert und beim Ausführen des Plans verwendet.

- **Kontingenztabelle aus dem aktiven Dataset berechnen.** Wenn Sie mit einem vorhandenen Simulationsplan arbeiten, der eine Kontingenztabelle enthält, können Sie die Kontingenztabelle aus dem aktiven Dataset neu berechnen. Durch diese Aktion wird die Kontingenztabelle aus der geladenen Plandatei überschrieben.

- **Kontingenztabelle aus geladenem Simulationsplan verwenden.** Wenn Sie einen Simulationsplan laden, der eine Kontingenztabelle enthält, wird die Tabelle aus dem Plan verwendet. Sie können die Kontingenztabelle aus dem aktiven Dataset neu berechnen, indem Sie **Kontingenztabelle aus dem aktiven Dataset berechnen** auswählen.

Erweiterte Optionen

Maximale Anzahl an Fällen. Dadurch wird die maximal zu generierende Anzahl von Fällen mit simulierten Daten (sowie die zugehörigen Zielwerte) angegeben. Wenn Sensitivitätsanalyse angegeben wurde, ist dies die maximale Anzahl von Fällen in jeder Iteration.

Ziel für Stoppkriterien Wenn Ihr Vorhersagemodell mehrere Ziele enthält, können Sie das Ziel auswählen, auf das Stoppkriterien angewendet werden sollen.

Stoppkriterien. Hier können Kriterien für das Stoppen der Simulation angegeben werden, und zwar möglicherweise, bevor die maximale Anzahl zulässiger Fälle generiert wurde.

- **Bis Erreichen des Höchstwerts fortfahren.** Hiermit wird angegeben, dass so lange simulierte Fälle generiert werden, bis die maximale Anzahl von Fällen erreicht ist.
- **Stoppen, wenn Stichprobenziehung aus Flanken abgeschlossen.** Verwenden sie diese Option, wenn Sie sicherstellen möchten, dass aus einer der Flanken eine angemessenen Stichprobe gezogen wurde. Es werden so lange simulierte Fälle generiert, bis die angegebene Stichprobenziehung aus der Flanke abgeschlossen ist oder die maximale Anzahl von Fällen erreicht wurde. Wenn Ihr Vorhersagemodell mehrere Ziele enthält, wählen Sie in der Dropdown-Liste **Ziel für Stoppkriterien** das Ziel aus, auf das diese Kriterien angewendet werden sollen.

  Typ. Sie können die Grenze des Flankenbereichs definieren, indem Sie einen Wert für das Ziel angeben, wie beispielsweise 10.000.000, oder ein Perzentil, wie beispielsweise das 99. Perzentil. Wenn Sie in der Dropdown-Liste **Typ** die Option "Wert" auswählen, müssen Sie anschließend den Wert der Grenze in das Textfeld "Wert" eingeben und mithilfe der Dropdown-Liste **Seite** angeben, ob es sich um den linken oder den rechten Flankenbereich handelt. Wenn Sie in der Dropdown-Liste **Typ** die Option "Perzentil" auswählen, müssen Sie anschließend einen Wert in das Textfeld "Perzentil" eingeben.

  Häufigkeit. Geben sie an, wie viele Werte des Ziels im Flankenbereich liegen müssen, um sicherzustellen, dass eine angemessene Stichprobe aus der Flanke gezogen wurde. Bei Erreichen dieses Werts wird die Fallerzeugung gestoppt.
- **Stoppen, wenn das Konfidenzintervall des Mittelwerts innerhalb des angegebenen Schwellenwerts liegt.** Verwenden Sie diese Option, wenn Sie sicherstellen möchten, dass der Mittelwert des Ziels mit einer bestimmten Genauigkeit bekannt ist. Es werden so lange simulierte Fälle generiert, bis der angegebene Genauigkeitsgrad oder die maximale Anzahl von Fällen erreicht wurde. Zur Verwendung dieser Option geben Sie ein Konfidenzniveau und einen Schwellenwert an. Es werden so lange simulierte Fälle generiert, bis das dem angegebenen Niveau zugeordnete Konfidenzintervall innerhalb des Schwellenwerts liegt. Beispielsweise können Sie mit dieser Option angeben, dass so lange Fälle generiert werden, bis das Konfidenzintervall des Mittelwerts bei einem Konfidenzniveau von 95 % im Bereich von 5 % um den Mittelwert liegt. Wenn Ihr Vorhersagemodell mehrere Ziele enthält, wählen Sie in der Dropdown-Liste **Ziel für Stoppkriterien** das Ziel aus, auf das diese Kriterien angewendet werden sollen.

  Schwellenwerttyp. Sie können den Schwellenwert als numerischen Wert oder als Prozentsatz des Mittelwerts angeben. Wenn Sie in der Dropdown-Liste **Schwellenwerttyp** die Option "Wert" auswählen, müssen Sie den Schwellenwert anschließend in das Textfeld "Schwellenwert als Wert" eingeben. Wenn Sie in der Dropdown-Liste **Schwellenwerttyp** die Option "Prozent" auswählen, müssen Sie anschließend einen Wert in das Textfeld "Schwellenwert als Prozent" eingeben.

Anzahl der Fälle in Stichprobe. Hier können Sie angeben, wie viele Fälle für die automatische Anpassung von Verteilungen für simulierte Eingaben an das aktive Dataset verwendet werden sollen. Wenn Ihr Dataset sehr groß ist, kann es sinnvoll sein, die Anzahl der Fälle, die für die Verteilungsanpassung verwendet werden, zu begrenzen. Bei Auswahl von **Auf N Fälle begrenzen** werden die ersten N Fälle verwendet.

Anpassungsgütekriterien (stetig). Bei stetigen Eingaben können Sie den Anderson-Darling-Test oder den Kolmogorow-Smirnow-Test für die Anpassungsgüte verwenden, um bei der Anpassung der Verteilungen für simulierte Eingaben an das aktive Dataset eine Rangfolge der Verteilungen zu erstellen. Der Anderson-Darling-Test wird standardmäßig ausgewählt und wird insbesondere dann empfohlen, wenn Sie die bestmögliche Anpassung in den Flankenbereichen sicherstellen möchten.

Empirische Verteilung. Bei stetigen Eingaben ist die empirische Verteilung die kumulative Verteilungsfunktion der historischen Daten. Sie können angeben, wie viele Klassen für die Berechnung der empirischen Verteilung für stetige Eingaben verwendet werden sollen. Die Standardeinstellung ist 100 und der Höchstwert ist 1000.

Ergebnisse reproduzieren. Durch Einstellen eines Startwerts für Zufallszahlen kann die Simulation reproduziert werden. Geben Sie eine ganze Zahl ein oder klicken Sie auf **Generieren**. Dadurch wird eine pseudozufällige Ganzzahl zwischen 1 und 2147483647 (einschließlich) erstellt. Der Standardwert ist 629111597.

Benutzerdefiniert fehlende Werte für Eingaben mit einer kategorialen Verteilung. Diese Steuerelemente geben an, ob benutzerdefiniert fehlende Werte für Eingaben mit einer kategorialen Verteilung als gültig behandelt werden. Systemdefiniert fehlende Werte und benutzerdefiniert fehlende Werte für alle anderen Typen von Eingaben werden immer als ungültige Werte behandelt. Alle Eingaben müssen gültige Werte für einen Fall aufweisen, um in die Verteilungsanpassung, die Berechnung von Korrelationen und die Berechnung der optionalen Kontingenztabelle aufgenommen zu werden.

Dichtefunktionen

Mit diesen Einstellungen können Sie Ausgaben für Wahrscheinlichkeitsdichtefunktionen und kumulative Verteilungsfunktionen für stetige Ziele benutzerdefiniert gestalten, ebenso wie Balkendiagramme vorhergesagter Werte für kategoriale Ziele.

Wahrscheinlichkeitsdichtefunktion (PDF) Die Wahrscheinlichkeitsdichtefunktion zeigt die Verteilung der Zielwerte an. Bei stetigen Zielen können Sie damit die Wahrscheinlichkeit bestimmen, mit der das Ziel in einem bestimmten Bereich liegt. Bei kategorialen Zielen (Zielen mit nominalem oder ordinalem Messniveau) wird ein Balkendiagramm generiert, in dem der Prozentsatz der Fälle angezeigt wird, die jeweils auf die einzelnen Kategorien des Ziels entfallen. Zusätzliche Optionen für kategoriale Ziele von PMML-Modellen stehen mit der weiter unten beschriebenen Einstellung "Zu berichtende Kategoriewerte" zur Verfügung.

Bei Two-Step-Clustermodellen und Clusterzentrenmodellen wird ein Balkendiagramm für die Clusterzugehörigkeit erstellt.

Kumulative Verteilungsfunktion (CDF). Die kumulative Verteilungsfunktion zeigt die Wahrscheinlichkeit an, mit der der Wert des Ziels kleiner oder gleich einem angegebenen Wert ist. Diese Option ist nur für stetige Ziele verfügbar.

Schiebereglerpositionen. Sie können die ursprünglichen Speicherorte der verschiebbaren Bezugslinien in PDF- und CDF-Diagrammen angeben. Für die untere und die obere Linie angegebene Werte beziehen sich auf Speicherorte an der horizontalen Achse, nicht auf Perzentile. Sie können die untere Linie entfernen, indem Sie **-Unendlich** auswählen, oder Sie können die obere Linie entfernen, indem Sie **Unendlich** auswählen. Standardmäßig befinden sich die Linien am 5. und am 95. Perzentil. Wenn mehrere Verteilungsfunktionen in demselben Diagramm angezeigt werden (da mehrere Ziele oder Ergebnisse aus Iterationen der Sensitivitätsanalyse vorliegen), bezieht sich dies standardmäßig auf die Verteilung für die erste Iteration oder das erste Ziel.

Bezugslinien (stetig). Sie können verschiedene vertikale Bezugslinien anfordern, die zu Wahrscheinlichkeitsdichtefunktionen und kumulativen Verteilungsfunktionen für stetige Ziele hinzugefügt werden können.

- **Sigmas.** Sie können Bezugslinien bei plus und minus einer angegebenen Anzahl von Standardabweichungen vom Mittelwert eines Ziels hinzufügen.

- **Perzentile.** Sie können Bezugslinien bei einem oder zwei Perzentilwerten der Verteilung eines Ziels hinzufügen, indem Sie Werte in die Textfelder "Unten" und "Oben" eingeben. Der Wert "95" im Textfeld "Oben" steht beispielsweise für das 95. Perzentil, also den Wert, unter dem 95 % der Beobachtungen liegen). Der Wert "5" im Textfeld "Unten" steht für das 5. Perzentil, also den Wert, unter dem 5 % der Beobachtungen liegen).
- **Benutzerdefinierte Bezugslinien.** Sie können Bezugslinien an bestimmten Werten des Ziels hinzufügen.

Anmerkung: Wenn mehrere Verteilungsfunktionen in demselben Diagramm angezeigt werden (da mehrere Ziele oder Ergebnisse aus Iterationen der Sensitivitätsanalyse vorliegen), gelten die Bezugslinien nur für die Verteilung für die erste Iteration oder das erste Ziel. Über das Dialogfeld "Diagrammoptionen", auf das Sie über das PDF- oder CDF-Diagramm zugreifen, können Sie Bezugslinien zu den anderen Verteilungen hinzufügen.

Ergebnisse aus separaten stetigen Zielen überlagern. Wenn mehrere stetige Ziele vorliegen, wird hiermit angegeben, ob Verteilungsfunktionen für alle diese Ziele in demselben Diagramm angezeigt werden, mit einem Diagramm für Wahrscheinlichkeitsdichtefunktionen und einem weiteren für kumulative Verteilungsfunktionen. Wenn diese Option nicht aktiviert ist, werden die Ergebnisse für die einzelnen Ziele jeweils in einem gesonderten Diagramm angezeigt.

Zu berichtende Kategoriewerte. Bei PMML-Modellen mit kategorialen Zielen ist das Ergebnis des Modells ein Set von vorhergesagten Wahrscheinlichkeiten (eine für jede Kategorie) dafür, dass der Zielwert auf die einzelnen Kategorien entfällt. Die Kategorie mit der höchsten Wahrscheinlichkeit wird als vorhergesagte Kategorie und beim Generieren des Balkendiagramms verwendet, das für die oben angegebene Einstellung **Wahrscheinlichkeitsdichtefunktion** beschrieben ist. Durch Auswahl von **Vorhergesagte Kategorie** wird das Balkendiagramm generiert. Durch Auswahl von **Geschätzte Wahrscheinlichkeiten** werden Histogramme der Verteilung der vorhergesagten Wahrscheinlichkeiten für die einzelnen Kategorien des Ziels generiert.

Gruppierung für Sensitivitätsanalyse. Simulationen, die eine Sensitivitätsanalyse beinhalten, generieren ein unabhängiges Set vorhergesagter Zielwerte für jede von der Analyse definierte Iteration (eine Iteration für jeden variierten Eingabewert). Wenn Iterationen vorliegen, wird das Balkendiagramm der vorhergesagten Kategorie für ein kategoriales Ziel als gruppiertes Balkendiagramm angezeigt, das die Ergebnisse für alle Iterationen beinhaltet. Sie können auswählen, dass die Kategorien gruppiert werden sollen, oder Sie können die Iterationen gruppieren.

Ausgabe

Tornado-Diagramme. Tornado-Diagramme sind Balkendiagramme, die anhand verschiedener Metriken Beziehungen zwischen Zielen und simulierten Eingaben anzeigen.

- **Korrelation zwischen Ziel und Eingabe.** Mit dieser Option wird ein Tornado-Diagramm der Korrelationskoeffizienten zwischen einem Ziel und seinen einzelnen simulierten Eingaben erstellt. Dieser Typ von Tornado-Diagramm unterstützt Ziele mit einem nominalen oder ordinalen Messniveau oder simulierten Eingaben für eine kategoriale Verteilung nicht.
- **Beitrag zu Varianz.** Mit dieser Option wird ein Tornado-Diagramm erstellt, das den Beitrag angibt, den ein Ziel ausgehend von jeder seiner simulierten Eingaben zur Varianz leistet, sodass Sie einschätzen können, in welchem Grad die einzelnen Eingaben zur Gesamtunsicherheit des Ziels beitragen. Dieser Typ von Tornado-Diagramm unterstützt keine Ziele mit ordinalem oder nominalem Messniveau und auch keine simulierten Eingaben mit einer der folgenden Verteilungen: kategoriale Verteilung, Bernoulli-Verteilung, binomiale Verteilung, Poisson-Verteilung oder negativ binomiale Verteilung.
- **Änderungssensitivität des Ziels.** Mit dieser Option wird ein Tornado-Diagramm erstellt, das den Effekt auf das Ziel anzeigt, der entsteht, wenn die einzelnen simulierten Eingaben um plus oder minus einer angegebenen Anzahl von Standardabweichungen der mit der Eingabe verknüpften Verteilung moduliert werden. Dieser Typ von Tornado-Diagramm unterstützt keine Ziele mit ordinalem oder no-

minalem Messniveau und auch keine simulierten Eingaben mit einer der folgenden Verteilungen: kategoriale Verteilung, Bernoulli-Verteilung, binomiale Verteilung, Poisson-Verteilung oder negativ binomiale Verteilung.

Boxplots der Zielverteilungen. Boxplots sind für stetige Ziele verfügbar. Wählen Sie die Option **Ergebnisse aus separaten stetigen Zielen überlagern**, wenn Ihr Vorhersagemodell mehrere stetige Ziele aufweist und Sie die Boxplots für alle Ziele in demselben Diagramm anzeigen möchten.

Streudiagramme der Ziele in Abhängigkeit von den Eingaben. Streudiagramme der Ziele in Abhängigkeit von den simulierten Eingaben stehen sowohl für stetige als auch für kategoriale Ziele zur Verfügung und beinhalten auch Streudiagramme des Ziels sowohl mit stetigen als auch mit kategorialen Eingaben. Streudiagramme, die ein kategoriales Ziel oder eine kategoriale Eingabe enthalten, werden als Heat-Map angezeigt.

Tabelle der Perzentilwerte erstellen. Bei stetigen Zielen können Sie eine Tabelle mit angegebenen Perzentilen der Zielverteilungen abrufen. Quartile (das 25., 50. und 75. Perzentil) unterteilen die Beobachtungen in vier gleich große Gruppen. Falls Sie eine gleiche Anzahl von Gruppen wünschen, die von vier abweicht, klicken Sie auf **Intervalle** und geben Sie die Anzahl an. Wählen Sie die Option **Benutzerdefinierte Perzentile** zur Angabe einzelner Perzentile, beispielsweise des 99. Perzentils.

Deskriptive Statistiken der Zielverteilungen. Mit dieser Option werden Tabellen mit deskriptiven Statistiken für stetige und kategoriale Ziele sowie für stetige Eingaben erstellt. Bei stetigen Zielen beinhaltet die Tabelle Mittelwert, Standardabweichung, Median, Minimum und Maximum, Konfidenzintervall des Mittelwerts auf dem angegebenen Niveau sowie das 5. und das 95. Perzentil der Zielverteilung. Bei kategorialen Zielen beinhaltet die Tabelle die Prozentsätze der Fälle, die auf die einzelnen Kategorien des Ziels entfallen. Bei kategorialen Zielen der PMML-Modelle beinhaltet die Tabelle außerdem jeweils die mittlere Wahrscheinlichkeit der einzelnen Kategorien des Ziels. Bei stetigen Eingaben beinhaltet die Tabelle Mittelwert, Standardabweichung, Minimum und Maximum.

Korrelationen und Kontingenztabellen für Eingaben. Diese Option zeigt eine Tabelle mit Korrelationskoeffizienten zwischen simulierten Eingaben an. Wenn Eingaben mit kategorialen Verteilungen aus einer Kontingenztabelle erstellt werden, wird auch die Kontingenztabelle der Daten angezeigt, die für diese Eingaben generiert werden.

In die Ausgabe aufzunehmende simulierte Eingaben. In der Standardeinstellung sind alle simulierten Eingaben in der Ausgabe enthalten. Sie können ausgewählte simulierte Eingaben aus der Ausgabe ausschließen. Dadurch werden sie aus Tornado-Diagrammen, Streudiagrammen und Tabellenausgaben ausgeschlossen.

Grenzwertbereiche für stetige Ziele. Sie können den Bereich gültiger Werte für mindestens ein stetiges Ziel angeben. Werte außerhalb des angegebenen Bereichs werden von allen Ausgaben und Analysen ausgeschlossen, die den Zielen zugeordnet sind. Wenn Sie eine Untergrenze setzen wollen, wählen Sie in der Spalte "Grenzwert" die Option **Unterer** aus und geben Sie einen Wert in der Spalte "Minimum" ein. Wenn Sie eine Obergrenze setzen wollen, wählen Sie in der Spalte "Grenzwert" die Option **Oberer** aus und geben Sie einen Wert in der Spalte "Maximum" ein. Um sowohl eine Untergrenze als auch eine Obergrenze zu setzen, wählen Sie in der Spalte "Grenzwert" die Option **Beides** aus und geben Sie in den Spalten "Minimum" und "Maximum" Werte ein.

Anzeigeformate. Sie können festlegen, welches Format bei der Anzeige der Werte für Ziele und Eingaben (sowohl feste Eingaben als auch simulierte Eingaben) verwendet werden soll.

Speichern

Plan für diese Simulation speichern. Sie können die aktuellen Spezifikationen für Ihre Simulation in einer Simulationsplandatei speichern. Simulationsplandateien tragen die Erweiterung *.splan*. Sie können den Plan im Simulation Builder erneut öffnen, bei Bedarf Änderungen vornehmen und die Simulation ausführen. Sie können den Simulationsplan für andere Benutzer freigeben, die ihn dann im Dialogfeld "Simulati-

on ausführen" ausführen können. Simulationspläne enthalten alle Spezifikationen, ausgenommen folgende: Einstellungen für Dichtefunktionen; Ausgabeeinstellungen für Diagramme und Tabellen; Einstellungen für erweiterte Optionen für Anpassung, empirische Verteilung und Startwert für Zufallszahlen.

Speichern der simulierten Daten als neue Datendatei. Sie können simulierte Eingaben, feste Eingaben und vorhergesagte Zielwerte in einer SPSS Statistics-Datendatei, in einem neuen Dataset in der aktuellen Sitzung oder in einer Excel-Datei speichern. Jeder Fall (Zeile) der Datendatei besteht aus den vorhergesagten Werten der Ziele sowie den simulierten Eingaben und festen Eingaben, die die Zielwerte generieren. Wenn die Sensitivitätsanalyse angegeben ist, erzeugt jede Iteration ein zusammenhängendes Set von Fällen, die mit der Iterationsnummer beschriftet sind.

Dialogfeld "Simulation ausführen"

Das Dialogfeld "Simulation ausführen" ist für Benutzer konzipiert, die einen Simulationsplan besitzen und in erster Linie die Simulation ausführen möchten. Es enthält auch die erforderlichen Funktionen für die Ausführung der Simulation unter anderen Bedingungen. Er ermöglicht Ihnen die Ausführung folgender allgemeiner Aufgaben:

- Einrichten oder Ändern einer Sensitivitätsanalyse zur Untersuchung des Effekts, der durch Variieren des Werts einer festen Eingabe bzw. durch Variieren eines Verteilungsparameters für eine simulierte Eingabe erzeugt wird
- Erneutes Anpassen von Wahrscheinlichkeitsverteilungen für unsichere Eingaben (und Korrelationen zwischen diesen Eingaben) an neue Daten
- Ändern der Verteilung für eine simulierte Eingabe
- Passen Sie die Ausgabe an.
- Ausführen der Simulation

Registerkarte "Simulation"

Auf der Registerkarte "Simulation" können Sie die Sensitivitätsanalyse angeben, die Wahrscheinlichkeitsverteilungen für simulierte Eingaben und Korrelationen zwischen simulierten Eingaben an neue Daten anpassen und die einer simulierten Eingabe zugeordnete Wahrscheinlichkeitsverteilung ändern.

Das Raster "Simulierte Eingaben" enthält einen Eintrag für jedes Eingabefeld, das im Simulationsplan definiert ist. In jedem Eintrag werden der Name der Eingabe und der der Eingabe zugeordnete Wahrscheinlichkeitsverteilungstyp angezeigt, ebenso wie ein Beispieldiagramm der zugehörigen Verteilungskurve. Die einzelnen Eingaben weisen jeweils auch ein zugehöriges Statussymbol (einen farbigen Kreis mit Häkchen) auf, das nützlich ist, wenn Sie Verteilungen an neue Daten anpassen. Außerdem können die Eingaben ein Schlosssymbol enthalten, das angibt, dass die Eingabe gesperrt ist und nicht im Dialogfeld "Simulation ausführen" geändert oder an neue Daten angepasst werden kann. Zur Bearbeitung gesperrter Eingaben muss der Simulationsplan im Simulation Builder geöffnet werden.

Die einzelnen Eingaben sind entweder simuliert oder fest. Simulierte Eingaben sind Eingaben, deren Werte unsicher sind und die durch Stichprobenziehung aus einer angegebenen Wahrscheinlichkeitsverteilung generiert werden. Feste Eingaben sind Eingaben, deren Werte bekannt sind und die für jeden in der Simulation generierten Fall konstant bleiben. Um mit einer bestimmten Eingabe zu arbeiten, wählen Sie den Eintrag für die Eingabe im Raster "Simulierte Eingaben" aus.

Angabe einer Sensitivitätsanalyse

Mit der Sensitivitätsanalyse können Sie den Effekt systematischer Änderungen in einer festen Eingabe oder in einem Verteilungsparameter für eine simulierte Eingabe untersuchen, indem Sie ein unabhängiges Set simulierter Fälle – also im Grunde eine separate Simulation – für jeden angegebenen Wert generieren. Zur Angabe der Sensitivitätsanalyse wählen Sie eine feste oder simulierte Eingabe aus und klicken Sie auf

Sensitivitätsanalyse. Die Sensitivitätsanalyse ist auf eine einzelne feste Eingabe oder einen einzelnen Verteilungsparameter für eine simulierte Eingabe beschränkt. Weitere Informationen finden Sie im Thema „Sensitivitätsanalyse” auf Seite 189.

Erneutes Anpassen von Verteilungen an neue Daten

So passen Sie Wahrscheinlichkeitsverteilungen für simulierte Eingaben (und Korrelationen zwischen simulierten Eingaben) automatisch an Daten im aktiven Dataset an:

1. Vergewissern Sie sich, dass alle Modelleingaben jeweils dem richtigen Feld im aktiven Dataset zugeordnet sind. Die einzelnen simulierten Eingaben werden an das Feld im aktiven Dataset angepasst, das in der mit der betreffenden Eingabe verknüpften Dropdown-Liste **Feld** angegeben wurde. Sie können problemlos nicht zugeordnete Eingaben ermitteln, indem Sie nach Eingaben suchen, deren Statussymbol ein Häkchen mit einem Fragezeichen aufweist, wie unten gezeigt.

2. Ändern Sie alle erforderlichen Feldzuordnungen, indem Sie die Option **An Feld im Dataset anpassen** aktivieren und das Feld aus der Liste auswählen.
3. Klicken Sie auf **Alle anpassen**.

Für jede Eingabe, für die die Anpassung durchgeführt wurde, wird jeweils die am besten an die Daten angepasste Verteilung angezeigt, ebenso wie eine grafische Darstellung der Verteilung, die über ein Histogramm (oder Balkendiagramm) der historischen Daten gelegt ist. Wenn keine akzeptable Anpassung gefunden wird, wird die empirische Verteilung verwendet. Bei Eingaben, die an die empirische Verteilung angepasst sind, wird nur ein Histogramm der historischen Daten angezeigt, da die empirische Verteilung letztlich durch dieses Histogramm dargestellt wird.

Hinweis: Eine vollständige Liste der Statussymbole finden Sie im Thema „Simulierte Felder” auf Seite 185.

Ändern der Wahrscheinlichkeitsverteilungen

Sie können die Wahrscheinlichkeitsverteilung für eine simulierte Eingabe bearbeiten und optional eine simuliert Eingabe in eine feste Eingabe ändern oder umgekehrt.

1. Wählen Sie die Eingabe aus und klicken Sie auf **Verteilung manuell festlegen**.
2. Wählen Sie den Verteilungstyp aus und geben Sie die Verteilungsparameter an. Um eine simulierte Eingabe in eine feste Eingabe zu ändern, wählen Sie in der Dropdown-Liste **Typ** die Option "Fest" aus.

Nachdem Sie die Parameter für eine Verteilung eingegeben haben, wird das Beispieldiagramm der Verteilung (im Eintrag für die Eingabe angezeigt) entsprechend Ihren Änderungen aktualisiert. Weitere Informationen zur manuellen Angabe von Wahrscheinlichkeitsverteilungen finden Sie im Thema „Simulierte Felder” auf Seite 185.

Benutzerdefiniert fehlende Werte von kategorialen Eingabe wenn passend einschließen. Hiermit wird angegeben, ob benutzerdefiniert fehlende Werte für Eingaben mit einer kategorialen Verteilung als gültig behandelt werden, wenn Sie eine Neuanpassung an Daten im aktiven Dataset vornehmen. Systemdefiniert fehlende Werte und benutzerdefiniert fehlende Werte für alle anderen Typen von Eingaben werden immer als ungültige Werte behandelt. Alle Eingaben müssen gültige Werte für einen Fall aufweisen, um in die Verteilungsanpassung und die Berechnung von Korrelationen aufgenommen zu werden.

Registerkarte "Ausgabe"

Auf der Registerkarte "Ausgabe" können Sie die von der Simulation generierte Ausgabe anpassen.

Dichtefunktionen. Dichtefunktionen sind die wichtigste Methode zur Untersuchung des Ergebnissets aus Ihrer Simulation.

- **Wahrscheinlichkeitsdichtefunktion.** In der Wahrscheinlichkeitsdichtefunktion wird die Verteilung der Zielwerte angezeigt, sodass Sie ermitteln können, mit welcher Wahrscheinlichkeit das Ziel innerhalb eines bestimmten Bereichs liegt. Bei Zielen mit einem festen Ergebnisset, z. B. "schlechter Service", "mittelmäßiger Service", "guter Service" und "hervorragender Service", wird ein Balkendiagramm generiert, das jeweils anzeigt, welcher Prozentsatz an Fällen auf die einzelnen Kategorien des Ziels entfällt.
- **Kumulative Verteilungsfunktion.** Die kumulative Verteilungsfunktion zeigt die Wahrscheinlichkeit an, mit der der Wert des Ziels kleiner oder gleich einem angegebenen Wert ist.

Tornado-Diagramme. Tornado-Diagramme sind Balkendiagramme, die anhand verschiedener Metriken Beziehungen zwischen Zielen und simulierten Eingaben anzeigen.

- **Korrelation zwischen Ziel und Eingabe.** Mit dieser Option wird ein Tornado-Diagramm der Korrelationskoeffizienten zwischen einem Ziel und seinen einzelnen simulierten Eingaben erstellt.
- **Beitrag zu Varianz.** Mit dieser Option wird ein Tornado-Diagramm erstellt, das den Beitrag angibt, den ein Ziel ausgehend von jeder seiner simulierten Eingaben zur Varianz leistet, sodass Sie einschätzen können, in welchem Grad die einzelnen Eingaben zur Gesamtunsicherheit des Ziels beitragen.
- **Änderungssensitivität des Ziels.** Mit dieser Option wird ein Tornado-Diagramm erstellt, das den Effekt auf das Ziel anzeigt, der entsteht, wenn die einzelnen simulierten Eingaben um plus oder minus einer Standardabweichung der mit der Eingabe verknüpften Verteilung moduliert werden.

Streudiagramme der Ziele in Abhängigkeit von den Eingaben. Mit dieser Option werden Streudiagramme der Ziele in Abhängigkeit von simulierten Eingaben generiert.

Boxplots der Zielverteilungen. Mit dieser Option werden Boxplots der Zielverteilungen generiert.

Quartiltabelle. Mit dieser Option wird eine Tabelle mit den Quartilen der Zielverteilungen generiert. Bei den Quartilen einer Verteilung handelt es sich um das 25., 50. und 75. Perzentil der Verteilung. Die Beobachtungen werden somit in vier gleich große Gruppen unterteilt.

Korrelationen und Kontingenztabellen für Eingaben. Diese Option zeigt eine Tabelle mit Korrelationskoeffizienten zwischen simulierten Eingaben an. Eine Kontingenztabelle mit Zuordnungen zwischen Eingaben mit einer kategorialen Verteilung wird angezeigt, wenn der Simulationsplan angibt, dass kategoriale Daten aus einer Kontingenztabelle generiert werden.

Ergebnisse aus separaten Zielen überlagern. Wenn das Vorhersagemodell, das Sie simulieren, mehrere Ziele enthält, können Sie angeben, ob die Ergebnisse von separaten Zielen in demselben Diagramm angezeigt werden sollen. Diese Einstellung gilt für Diagramme von Wahrscheinlichkeitsdichtefunktionen, kumulativen Verteilungsfunktionen und Boxplots. Bei Auswahl dieser Option werden beispielsweise die Wahrscheinlichkeitsdichtefunktionen für alle Ziele in demselben Diagramm angezeigt.

Plan für diese Simulation speichern. Sie können alle Änderungen an Ihrer Simulation in einer Simulationsplandatei speichern. Simulationsplandateien tragen die Erweiterung *.splan*. Sie können den Plan im Dialogfeld "Simulation ausführen" oder im Simulation Builder erneut öffnen. Simulationspläne beinhalten alle Spezifikationen mit Ausnahme der Ausgabeeinstellungen.

Speichern der simulierten Daten als neue Datendatei. Sie können simulierte Eingaben, feste Eingaben und vorhergesagte Zielwerte in einer SPSS Statistics-Datendatei, in einem neuen Dataset in der aktuellen Sitzung oder in einer Excel-Datei speichern. Jeder Fall (Zeile) der Datendatei besteht aus den vorhergesagten Werten der Ziele sowie den simulierten Eingaben und festen Eingaben, die die Zielwerte generieren. Wenn die Sensitivitätsanalyse angegeben ist, erzeugt jede Iteration ein zusammenhängendes Set von Fällen, die mit der Iterationsnummer beschriftet sind.

Wenn Sie die Ausgabe in größerem Umfang benutzerdefiniert anpassen möchten als hier möglich, sollten Sie die Simulation über den Simulation Builder ausführen. Weitere Informationen finden Sie im Thema „Ausführen einer Simulation über einen Simulationsplan" auf Seite 181.

Arbeiten mit Diagrammausgaben aus der Simulation

Einige der aus einer Simulation generierten Diagramme weisen interaktive Funktionen auf, mit denen Sie die Anzeige anpassen können. Interaktive Funktionen stehen durch Aktivieren (Doppelklick) des Diagrammobjekts im Ausgabeviewer zur Verfügung. Alle Simulationsdiagramme sind Grafiktafelvisualisierungen.

Diagramme der Wahrscheinlichkeitsdichtefunktionen für stetige Ziele. Dieses Diagramm enthält zwei verschiebbare vertikale Bezugslinien, die das Diagramm in separate Bereiche unterteilen. In der Tabelle unter dem Diagramm wird die Wahrscheinlichkeit angezeigt, mit der sich das Ziel in den einzelnen Bereichen befindet. Wenn mehrere Dichtefunktionen im selben Diagramm angezeigt werden, enthält die Tabelle eine gesonderte Zeile für die den einzelnen Dichtefunktionen zugeordneten Wahrscheinlichkeiten. Für jede Bezugslinie gibt es einen Schieberegler (umgedrehtes Dreieck), mit dem Sie die Linie problemlos verschieben können. Durch Klicken auf **Diagrammoptionen** unten im Diagramm steht eine Reihe weiterer Funktionen zur Verfügung. Insbesondere können Sie explizit die Positionen der Schieberegler festlegen, feste Bezugslinien hinzufügen und die Diagrammansicht von einer stetigen Kurve in ein Histogramm ändern oder umgekehrt. Weitere Informationen finden Sie im Thema „Diagrammoptionen".

Diagramme der kumulativen Verteilungsfunktionen für stetige Ziele. Dieses Diagramm enthält dieselben beiden verschiebbaren vertikalen Bezugslinien und zugehörigen Tabellen, die oben für das Diagramm der Wahrscheinlichkeitsdichtefunktionen beschrieben wurden. Es bietet ebenfalls Zugriff auf das Dialogfeld "Diagrammoptionen", in dem Sie explizit die Position der Schieberegler festlegen, feste Bezugslinien hinzufügen und angeben können, ob die kumulative Verteilungsfunktion als steigende Funktion (Standardeinstellung) oder fallende Funktion dargestellt werden soll. Weitere Informationen finden Sie im Thema „Diagrammoptionen".

Balkendiagramme für kategoriale Ziele mit Sensitivitätsanalyseiterationen. Bei kategorialen Zielen mit Sensitivitätsanalyseiterationen werden die Ergebnisse für die vorhergesagte Zielkategorie als gruppiertes Balkendiagramm angezeigt, das die Ergebnisse für sämtliche Iterationen enthält. Das Diagramm beinhaltet eine Dropdown-Liste, mit der Sie die Gruppierung nach Kategorie oder nach Iteration durchführen können. Bei Two-Step-Clustermodellen und Clusterzentrenmodellen können Sie die Gruppierung nach Clusternummer oder Iteration durchführen.

Boxplots für mehrere Ziele mit Sensitivitätsanalyseiterationen. Bei Vorhersagemodellen mit mehreren stetigen Zielen und Sensitivitätsanalyseiterationen führt die Auswahl, dass die Boxplots für alle Ziele im selben Diagramm angezeigt werden sollen, zur Erstellung eines gruppierten Boxplots. Das Diagramm beinhaltet eine Dropdown-Liste, mit der Sie die Gruppierung nach Ziel oder nach Iteration durchführen können.

Diagrammoptionen

Im Dialogfeld "Diagrammoptionen" können Sie die Anzeige der aktivierten Diagramme von Wahrscheinlichkeitsdichtefunktionen und kumulativen Verteilungsfunktionen, die aus einer Simulation erstellt wurden, benutzerdefiniert anpassen.

Ansicht. Die Dropdown-Liste **Ansicht** gilt nur für das Diagramm der Wahrscheinlichkeitsdichtefunktion. Sie können damit die Diagrammansicht zwischen einer stetigen Kurve und einem Histogramm umschalten. Diese Funktion steht nicht zur Verfügung, wenn mehrere Dichtefunktionen in demselben Diagramm angezeigt werden. In diesem Fall können die Dichtefunktionen nur als stetige Kurven angezeigt werden.

Reihenfolge. Die Dropdown-Liste **Reihenfolge** gilt nur für das Diagramm der kumulativen Verteilungsfunktion. Sie gibt an, ob die kumulative Verteilungsfunktion als steigende Funktion (Standardeinstellung)

oder fallende Funktion angezeigt wird. Bei der Anzeige als fallende Funktion gibt der Wert der Funktion an einem bestimmten Punkt auf der horizontalen Achse die Wahrscheinlichkeit an, mit der das Ziel rechts von diesem Punkt liegt.

Schiebereglerpositionen. Sie können die Positionen der verschiebbaren Bezugslinien explizit festlegen, indem Sie Werte in die Textfelder "Oberer Bereich" und "Unterer Bereich" eingeben. Sie können die linke Linie entfernen, indem Sie **Minus unendlich** auswählen und somit die Position auf minus unendlich setzen. Die rechte Linie kann durch Auswahl von **Unendlich**, wodurch die Position auf unendlich gesetzt wird, entfernt werden.

Bezugslinien. Sie können verschiedene vertikale Bezugslinien zu Wahrscheinlichkeitsdichtefunktionen und kumulativen Verteilungsfunktionen hinzufügen. Wenn mehrere Funktionen in demselben Diagramm angezeigt werden, da mehrere Ziele oder Ergebnisse aus Iterationen der Sensitivitätsanalyse vorliegen, können Sie angeben, auf welche Funktionen die Linien jeweils angewendet werden.

- **Sigmas.** Sie können Bezugslinien bei plus und minus einer angegebenen Anzahl von Standardabweichungen vom Mittelwert eines Ziels hinzufügen.
- **Perzentile.** Sie können Bezugslinien bei einem oder zwei Perzentilwerten der Verteilung eines Ziels hinzufügen, indem Sie Werte in die Textfelder "Unten" und "Oben" eingeben. Der Wert "95" im Textfeld "Oben" steht beispielsweise für das 95. Perzentil, also den Wert, unter dem 95 % der Beobachtungen liegen). Der Wert "5" im Textfeld "Unten" steht für das 5. Perzentil, also den Wert, unter dem 5 % der Beobachtungen liegen).
- **Benutzerdefinierte Positionen.** Sie können Bezugslinien an bestimmten Werten auf der horizontalen Achse hinzufügen.

Bezugslinienbeschriftungen. Diese Option steuert, ob die ausgewählten Bezugslinien beschriftet werden.

Bezugslinien werden entfernt, indem die zugehörige Auswahl im Dialogfeld "Diagrammoptionen" gelöscht und auf **Weiter** geklickt wird.

Kapitel 35. Georäumliche Modellierung

Georäumliche Modellierungsverfahren ermitteln Muster in Daten, die eine räumliche Komponente (Kartenkomponente) enthalten. Der Geomodellierungsassistent stellt Methoden für die Analyse von Geodaten mit und ohne Zeitkomponente bereit.

Assoziationen auf der Basis von Ereignis- und Geodaten suchen (Geoassoziationsregeln)
Mithilfe von Geoassoziationsregeln können Sie Muster in Daten auf der Basis von sowohl räumlichen als auch nicht räumlichen Eigenschaften suchen. Sie können beispielsweise Muster in kriminologischen Daten nach Position und demografischen Attributen ermitteln. Anhand dieser Muster können Sie Regeln erstellen, die vorhersagen, wo bestimmte Arten von Verbrechen wahrscheinlich auftreten.

Vorhersagen mithilfe von Zeitreihen und Geodaten erstellen (räumlich-temporale Vorhersage).
Die räumlich-temporale Vorhersage verwendet Daten, die Positionsdaten, Eingabefelder für die Vorhersage (Prädiktoren), mindestens ein Zeitfeld und ein Zielfeld enthalten. Jede Position enthält mehrere Zeilen in den Daten, die die Werte jedes Prädiktors und das Ziel in den einzelnen Zeitintervallen darstellen.

Verwenden des Geomodellierungsassistenten

1. Wählen Sie in den Menüs Folgendes aus:
   Analysieren > **Räumliche und temporale Modellierung** > **Räumliche Modellierung**
2. Führen Sie die Schritte im Assistenten aus.

Auswählen von Karten

Die georäumliche Modellierung kann eine oder mehrere Kartendatenquellen verwenden. Kartendatenquellen enthalten Informationen, die geografische Bereiche oder andere geografische Objekte wie Straßen oder Flüsse definieren. Viele Kartenquellen enthalten darüber hinaus auch demografische oder andere beschreibende Daten und Ereignisdaten wie Kriminalberichte oder Arbeitslosenquoten. Hier können Sie eine zuvor definierte Kartenspezifikationsdatei verwenden oder Kartenspezifikationen definieren und diese Spezifikationen für die spätere Verwendung speichern.

Kartenspezifikation laden
Lädt eine zuvor definierte Kartenspezifikationsdatei (.mplan). Von Ihnen hier definierte Kartendatenquellen können in einer Kartenspezifikationsdatei gespeichert werden. Wenn Sie bei der räumlich-temporalen Vorhersage eine Kartenspezifikationsdatei auswählen, die mehrere Karten angibt, werden Sie aufgefordert, eine Karte aus der Datei auszuwählen.

Kartendatei hinzufügen
Fügen Sie eine ESRI-Shapefile (.shp) oder ein ZIP-Archiv hinzu, das eine ESRI-Shapefile enthält.

- An dem Speicherort der .shp-Datei muss sich eine entsprechende .dbf-Datei befinden, die denselben Stammnamen wie die .shp-Datei aufweist.
- Wenn die Datei ein ZIP-Archiv ist, müssen die .shp-Datei und die .dbf-Datei denselben Stammnamen wie das ZIP-Archiv aufweisen.
- Wenn es keine entsprechende Projektionsdatei (.prj) gibt, werden Sie zur Auswahl eines Projektionssystems aufgefordert.

Beziehung
Diese Spalte definiert für Geoassoziationsregeln, in welchem Bezug Ereignisse zu den Merkmalen in der Karte stehen. Diese Einstellung ist nicht für räumlich-temporale Vorhersagen verfügbar.

Nach oben, Nach unten
Die Reihenfolge der Schichten der Kartenelemente wird von der Reihenfolge ihrer Anzeige in der Liste bestimmt. Die erste Karte in der Liste ist die Basisschicht.

Auswählen einer Karte

Wenn Sie bei der räumlich-temporalen Vorhersage eine Kartenspezifikationsdatei auswählen, die mehrere Karten angibt, werden Sie aufgefordert, eine Karte aus der Datei auszuwählen. Die räumlich-temporale Vorhersage unterstützt nicht mehrere Karten.

Georäumliche Beziehung

Das Dialogfeld **Georäumliche Beziehung** definiert für Geoassoziationsregeln, in welchem Bezug Ereignisse zu den Merkmalen in der Karte stehen.

- Diese Einstellung gilt nur für Geoassoziationsregeln.
- Diese Einstellung wirkt sich nur auf Datenquellen aus, die Karten zugeordnet sind, die im Schritt zur Auswahl von Datenquellen als Kontextdaten angegeben wurden.

Beziehung

Nahe Das Ereignis tritt nahe einem angegebenen Punkt oder Bereich in der Karte auf.

In Das Ereignis tritt in einem angegebenen Bereich in der Karte auf.

Enthält
Der Ereignisbereich enthält ein Kartenkontextobjekt.

Überschneidung
Positionen, an denen sich Linien oder Bereiche der verschiedenen Karten überschneiden.

Kreuz Bei mehreren Karten Positionen, an denen sich Linien (für Straßen, Flüsse oder Schienen) verschiedener Karten kreuzen.

Nördlich von, Südlich von, Östlich von, Westlich von
Das Ereignis tritt in einem Bereich nördlich, südlich, östlich oder westlich eines angegebenen Punkts in der Karte auf.

Festlegen des Koordinatensystems

Wenn keine Projektionsdatei (.prj) für die Karte vorhanden ist oder Sie zwei Felder einer Datenquelle als Koordinatensatz definieren, müssen Sie das Koordinatensystem festlegen.

Standardkoordinatensystem (Längengrad und Breitengrad)
Das Koordinatensystem besteht aus Längen- und Breitengrad.

Einfaches kartesisches Koordinatensystem (X und Y)
Das Koordinatensystem besteht aus einfachen X- und Y-Koordinaten.

WKID verwenden
WKID-Wert für gemeinsame Projektionen.

Koordinatensystemnamen verwenden
Das Koordinatensystem basiert auf der angegebenen Projektion. Der Name steht in runden Klammern.

Festlegen der Projektion

Wenn das Projektionssystem nicht durch die mit der Karte bereitgestellten Informationen festgelegt werden kann, müssen Sie es angeben. Die häufigste Ursache hierfür ist, dass der Karte keine Projektionsdatei (.prj) zugeordnet wurde oder eine Projektionsdatei nicht verwendet werden kann.

- **Eine Stadt, eine Region oder ein Land (Mercator)**
- **Ein großes Land, mehrere Länder oder Kontinente (Winkel Tripel)**

- **Ein Bereich nahe des Äquators (Mercator)**
- **Ein Bereich in Polnähe (stereografisch)**

Die Mercator-Projektion ist eine einheitliche Projektion, die in vielen Karten verwendet wird. Bei dieser Projektion wird der Globus als Zylinder behandelt, der auf einer ebenen Fläche ausgerollt wird. Die Mercator-Projektion verzerrt die Größe und Form großer Objekte. Diese Verzerrung nimmt vom Äquator zu den Polen immer weiter zu. Bei der Winkel Tripel-Projektion und der stereografischen Projektion wird durch Anpassungen berücksichtigt, dass eine Karte einen Teil einer dreidimensionalen Kugel darstellt, die zweidimensional angezeigt wird.

Projektions- und Koordinatensystem

Wenn Sie mehrere Karten auswählen, die unterschiedliche Projektions- und Koordinatensysteme haben, müssen Sie die Karte mit dem Projektionssystem auswählen, das Sie verwenden wollen. Das Projektionssystem wird für alle Karten verwendet, wenn sie in der Ausgabe kombiniert werden.

Datenquellen

Eine Datenquelle kann eine dBase-Datei sein, die mit der Shapefile, einer IBM SPSS Statistics-Datendatei oder einem geöffneten Dataset in der aktuellen Sitzung bereitgestellt wird.

Kontextdaten. Kontextdaten geben Merkmale in der Karte an. Kontextdaten können auch Felder enthalten, die als Eingaben für das Modell verwendet werden können. Wenn Sie eine dBase-Kontextdatei (.dbf) verwenden wollen, die einer Kartenshapedatei (.shp) zugeordnet ist, muss sich die dBase-Kontextdatei an derselben Position befinden wie die Shapefile und sie muss denselben Stammnamen aufweisen. Wenn die Shapefile beispielsweise `geodata.shp` heißt, muss die dBase-Datei den Namen `geodata.dbf` aufweisen.

Ereignisdaten. Ereignisdaten enthalten Informationen zu auftretenden Ereignissen, beispielsweise zu Verbrechen oder Unfällen. Diese Option ist nur für Geoassoziationsregeln verfügbar.

Punktdichte. Zeitintervall und Koordinatendaten für Kerndichteschätzungen. Diese Option ist nur für räumlich-temporale Vorhersagen verfügbar.

Hinzufügen. Öffnet ein Dialogfeld, in dem Datenquellen hinzugefügt werden können. Eine Datenquelle kann eine dBase-Datei sein, die mit der Shapefile, einer IBM SPSS Statistics-Datendatei oder einem geöffneten Dataset in der aktuellen Sitzung bereitgestellt wird.

Zuordnen. Öffnet ein Dialogfeld, in dem die Kennungen (Koordinaten oder Schlüssel) angegeben werden können, die zum Zuordnen von Daten zu Karten verwendet werden. Jede Datenquelle muss mindestens eine Kennung enthalten, die die Daten der Karte zuordnet. dBase-Dateien, die zusammen mit einer Shapefile bereitgestellt werden, enthalten normalerweise ein Feld, das automatisch als Standardkennung verwendet wird. Bei anderen Datenquellen müssen Sie die Felder angeben, die als Kennungen verwendet werden.

Schlüssel validieren. Öffnet ein Dialogfeld, in dem der Schlüsselabgleich zwischen der Karte und der Datenquelle validiert werden kann.

Geoassoziationsregeln

- Mindestens eine Datenquelle muss eine Ereignisdatenquelle sein.
- Alle Ereignisdatenquellen müssen dasselbe Format der Kartenassoziationskennungen verwenden: Koordinaten oder Schlüsselwerte.
- Wenn die Ereignisdatenquellen den Karten mit Schlüsselwerten zugeordnet sind, müssen alle Ereignisquellen denselben Kartenmerkmaltyp (beispielsweise Polygone, Punkte, Linien) verwenden.

Räumlich-temporale Vorhersage

- Es muss eine Kontextdatenquelle vorhanden sein.
- Wenn es nur eine Datenquelle gibt (eine Datendatei ohne zugeordnete Karte), muss sie Koordinatenwerte enthalten.
- Wenn Sie über zwei Datenquellen verfügen, muss eine Datenquelle Kontextdaten enthalten und die andere Datenquelle muss Punktdichtedaten enthalten.
- Sie können können maximal zwei Datenquellen einschließen.

Hinzufügen einer Datenquelle

Eine Datenquelle kann eine dBase-Datei sein, die mit der Shapefile und der Kontextdatei, einer IBM SPSS Statistics-Datendatei oder einem geöffnetem Dataset in der aktuellen Sitzung bereitgestellt wird.

Sie können eine Datenquelle mehrmals hinzufügen, wenn Sie jeweils verschiedene räumliche Assoziationen verwenden wollen.

Daten- und Kartenassoziation

Jede Datenquelle muss mindestens eine Kennung enthalten, die die Daten der Karte zuordnet.

Koordinaten
: Die Datenquelle enthält Felder, die kartesische Koordinaten darstellen. Wählen Sie die Felder aus, die die X- und Y-Koordinaten darstellen. Für Geoassoziationsregeln kann optional auch eine Z-Koordinate verfügbar sein.

Schlüsselwerte
: Die Schlüsselwerte in den Feldern der Datenquelle entsprechen den ausgewählten Kartenschlüsseln. Eine Karte der Regionen kann beispielsweise eine Namenskennung (Kartenschlüssel) umfassen, die jede Region kennzeichnet. Diese Kennung entspricht einem Feld in den Daten, das ebenfalls die Namen der Regionen (Datenschlüssel) enthält. Die Felder werden in der Reihenfolge mit den Kartenschlüsseln abgeglichen, in der sie in den beiden Listen angezeigt werden.

Schlüssel validieren

Das Dialogfeld **Schlüssel validieren** bietet eine Zusammenfassung des Datensatzabgleichs zwischen der Karte und der Datenquelle auf der Basis der ausgewählten Kennungsschlüssel. Wenn es Datenschlüsselwerte ohne Entsprechung gibt, können Sie sie manuell mit Kartenschlüsselwerten abgleichen.

Geoassoziationsregeln

Bei Geoassoziationsregeln bleiben nach der Definition von Karten und Datenquellen folgende Schritte im Assistenten:

- Wenn es mehrere Ereignisdatenquellen gibt, definieren Sie die Zusammenführung von Ereignisdatenquellen.
- Wählen Sie Felder aus, die als Bedingungen und Vorhersagen in der Analyse verwendet werden sollen.

Außerdem sind die folgenden Optionen verfügbar:

- Wählen Sie andere Ausgabeoptionen aus.
- Speichern Sie eine Scoring-Modelldatei.
- Erstellen Sie neue Felder für vorhergesagte Werte und Regeln in den Datenquellen, die im Modell verwendet werden.
- Passen Sie die Einstellungen für die Erstellung von Assoziationsregeln an.
- Passen Sie die Klassierungs- und Aggregationseinstellungen an.

Definition von Ereignisdatenfeldern

Bei Geoassoziationsregeln werden die Ereignisdatenquellen zusammengeführt, wenn mehrere Ereignisdatenquellen vorhanden sind.

- Standardmäßig werden nur Felder einbezogen, die allen Ereignisdatenquellen gemeinsam sind.
- Sie können eine Liste aller gemeinsamen Felder, der Felder für eine bestimmte Datenquelle oder der Felder von allen Datenquellen anzeigen und die Felder auswählen, die Sie einbeziehen möchten.
- Bei gemeinsamen Feldern müssen **Typ** und **Messung** für alle Datenquellen identisch sein. Bei Konflikten können Sie den Typ und das Messniveau angeben, der bzw. das für jedes gemeinsame Feld verwendet werden soll.

Auswählen von Feldern

Die Liste verfügbarer Felder umfasst Felder von den Ereignisdatenquellen und Felder von den Kontextdatenquellen.

- Sie können die Liste der angezeigten Felder steuern, indem Sie eine Datenquelle in der Liste **Datenquellen** auswählen.
- Sie müssen mindestens zwei Felder auswählen. Mindestens ein Feld muss eine Bedingung und mindestens ein Feld muss eine Vorhersage sein. Sie können diese Anforderung auf mehrere Arten erfüllen, beispielsweise können Sie für die Liste **Beides (Bedingung und Vorhersage)** zwei Felder auswählen.
- Assoziationsregeln sagen Werte der Vorhersagefelder basierend auf den Werten der Bedingungsfelder voraus. In der Regel "If x=1 and y=2, then z=3" sind die Werte von x und y beispielsweise Bedingungen und der Wert von z ist die Vorhersage.

Ausgabe

Regeltabellen
: Jede Regeltabelle zeigt die besten Regeln und Werte für Konfidenz, Regelunterstützung, Lift, Bedingungsunterstützung und Bereitstellbarkeit an. Jede Tabelle ist nach Werten des ausgewählten Kriteriums sortiert. Sie können alle Regeln anzeigen oder nur eine **Anzahl** der besten Regeln auf der Basis des ausgewählten Kriteriums.

Sortierbare Wortcloud
: Eine Liste der besten Regeln auf der Basis der Werte des ausgewählten Kriteriums. Die Größe des Texts gibt die relative Wichtigkeit der Regel an. Das interaktive Ausgabeobjekt enthält die besten Regeln für Konfidenz, Regelunterstützung, Lift, Bedingungsunterstützung und Bereitstellbarkeit. Das ausgewählte Kriterium legt fest, welche Regelliste standardmäßig angezeigt wird. Sie können in der Ausgabe interaktiv ein anderes Kriterium auswählen. **Max. anzuzeigende Regeln** legt die Anzahl der in der Ausgabe angezeigten Regeln fest.

Karte
: Interaktives Balkendiagramm und Karte der besten Regeln auf der Basis des ausgewählten Kriteriums. Jedes interaktive Ausgabeobjekt enthält die besten Regeln für Konfidenz, Regelunterstützung, Lift, Bedingungsunterstützung und Bereitstellbarkeit. Das ausgewählte Kriterium legt fest, welche Regelliste standardmäßig angezeigt wird. Sie können in der Ausgabe interaktiv ein anderes Kriterium auswählen. **Max. anzuzeigende Regeln** legt die Anzahl der in der Ausgabe angezeigten Regeln fest.

Modellinformationstabellen

Feldtransformationen.
: Beschreibt die Transformationen, die auf in der Analyse verwendete Felder angewendet werden.

Datensatzzusammenfassung.
: Anzahl und Prozentsatz der einbezogenen und ausgeschlossenen Datensätze.

Regelstatistik.
Auswertungsstatistik für Bedingungsunterstützung, Konfidenz, Regelunterstützung, Lift und Bereitstellbarkeit. Die Statistik umfasst folgende Informationen: Mittelwert, Minimum, Maximum und Standardabweichung.

Häufigste Elemente.
Elemente, die am häufigsten vorkommen. Ein Element ist in einer Bedingung oder Vorhersage in einer Regel eingeschlossen. Beispiel: Alter < 18 oder Geschlecht=weiblich.

Häufigste Felder.
Felder die in den Regeln am häufigsten vorkommen.

Ausgeschlossene Eingaben.
Felder, die von der Analyse ausgeschlossen sind, sowie der Grund für den Ausschluss der einzelnen Felder.

Kriterien für Regeltabellen, Wordcloud und Karten

Konfidenz.
Der Prozentsatz der richtigen Regelvorhersagen.

Regelunterstützung.
Der Prozentsatz der Fälle, für die die Regel wahr ist. Wenn die Regel beispielsweise "If x=1 and y=2, then z=3" lautet, ist die Regelunterstützung der tatsächliche Prozentsatz der Fälle, in denen x=1, y=2 und z=3 zutrifft.

Lift. Der Lift gibt an, wie sehr eine Regel die Vorhersage im Unterschied zur zufälligen Auswahl verbessert. Er ist das Verhältnis richtiger Vorhersagen zum Gesamtvorkommen des vorhergesagten Werts. Der Wert muss größer als 1 sein. Wenn der vorhergesagte Wert beispielsweise 20 % der Zeit ist und die Konfidenz in der Vorhersage 80 % ist, ist der Liftwert 4.

Bedingungsunterstützung.
Der Prozentsatz der Fälle, für die die Regelbedingung vorhanden ist. Wenn die Regel beispielsweise "If x=1 and y=2,then z=3" lautet, ist die Bedingungsunterstützung der Anteil der Fälle in den Daten, für die x=1 und y=2 zutrifft.

Bereitstellbarkeit.
Der Prozentsatz der falschen Vorhersagen, wenn die Bedingungen wahr sind. Die Bereitstellbarkeit ist gleich (1-Konfidenz) multipliziert mit der Bedingungsunterstützung oder Bedingungsunterstützung minus Regelunterstützung.

Speichern

Karten- und Kontextdaten als Kartenspezifikation speichern
Speichern Sie die Kartenspezifikation in einer externen Datei (.mplan). Sie können diese Kartenspezifikationsdatei für spätere Analysen in den Assistenten laden. Sie können die Kartenspezifikationsdatei auch mit dem Befehl `SPATIAL ASSOCIATION RULES` verwenden.

Karte und Kontextdatendateien in die Spezifikation kopieren
Daten aus Kartenshapedateien, externen Datendateien und in der Kartenspezifikation verwendeten Datasets werden in der Kartenspezifikationsdatei gespeichert.

Scoring
Speichert die besten Regelwerte, Konfidenzwerte für die Regeln und Werte für die numerische ID für die Regeln als neue Felder in der angegebenen Datenquelle.

Zu scorende Datenquelle
Die Datenquelle(n), in der bzw. in denen die neuen Felder erstellt werden. Wenn die Datenquelle nicht in der aktuellen Sitzung geöffnet ist, wird sie in der aktuellen Sitzung geöffnet. Sie müssen die geänderte Datei explizit speichern, um die neuen Felder zu speichern.

Zielwerte

Erstellen Sie neue Felder für die ausgewählten Zielfelder (Vorhersagefelder).

- Für jedes Zielfeld werden zwei neue Felder erstellt: für den vorhergesagten Wert und den Konfidenzwert.
- Für stetige (metrische) Zielfelder ist der vorhergesagte Wert eine Zeichenfolge, die einen Wertebereich beschreibt. Ein Wert im Format "[Wert1, Wert2]" bedeutet "größer als Wert1 und kleiner-gleich Wert2".

Anzahl der besten Regeln

Erstellen Sie neue Felder für die angegebene Anzahl der besten Regeln. Für jede Regel werden drei neue Felder erstellt: für den Regelwert, den Konfidenzwert und den Wert für die numerische ID für die Regel.

Namenspräfix

Für die neuen Feldnamen zu verwendendes Präfix.

Regelerstellung

Regelerstellungsparameter legen die Kriterien für die erstellten Assoziationsregeln fest.

Elemente pro Regel

Anzahl der Feldwerte, die in Regelbedingungen und Vorhersagen einbezogen werden können. Die Gesamtzahl der Elemente darf 10 nicht überschreiten. Die Regel "If x=1 and y=2, then z=3" enthält beispielsweise zwei Bedingungselemente und ein Vorhersageelement.

Max. Vorhersagen.

Maximale Anzahl der Feldwerte, die in den Vorhersagen für eine Regel vorkommen können.

Max. Bedingungen.

Maximale Anzahl der Feldwerte, die in den Bedingungen für eine Regel vorkommen können.

Paar ausschließen

Schließt die angegebenen Paare von Feldern aus, sodass sie nicht dieselbe Regel einbezogen werden.

Regelkriterien

Konfidenz.

Die Konfidenz, die eine Regel mindestens aufweisen muss, damit sie in die Ausgabe einbezogen wird. Konfidenz ist der Prozentsatz richtiger Vorhersagen.

Regelunterstützung.

Die Regelunterstützung, die eine Regel mindest aufweisen muss, damit sie in die Ausgabe einbezogen wird. Der Wert stellt den Prozentsatz der Fälle dar, für die die Regel in den beobachteten Daten wahr ist. Wenn die Regel beispielsweise "If x=1 and y=2, then z=3" lautet, ist die Regelunterstützung der tatsächliche Prozentsatz der Fälle, in denen x=1, y=2 und z=3 zutrifft.

Bedingungsunterstützung.

Die Bedingungsunterstützung, die eine Regel mindest aufweisen muss, damit sie in die Ausgabe einbezogen wird. Der Wert stellt den Prozentsatz der Fälle dar, für die die Bedingung vorhanden ist. Wenn die Regel beispielsweise "If x=1 and y=2,then z=3" lautet, ist die Bedingungsunterstützung der Prozentsatz der Fälle in den Daten, für die x=1 und y=2 zutrifft.

Lift. Der Lift, den eine Regel mindestens aufweisen muss, damit sie in die Ausgabe einbezogen wird. Der Lift gibt an, wie sehr eine Regel die Vorhersage im Unterschied zur zufälligen Auswahl verbessert. Er ist das Verhältnis richtiger Vorhersagen zum Gesamtvorkom-

men des vorhergesagten Werts. Wenn der vorhergesagte Wert beispielsweise 20 % der Zeit ist und die Konfidenz in der Vorhersage 80 % ist, ist der Liftwert 4.

Als identisch behandeln
Gibt Feldpaare an, die als identisches Feld behandelt werden sollen.

Klassierung und Aggregation

- Die Aggregation ist erforderlich, wenn es mehr Datensätze in den Daten gibt als Merkmale in der Karte. Sie haben beispielsweise Datensätze für einzelne Länder, aber eine Karte der Bundesländer.
- Sie können die Methoden für das Aggregationsauswertungsmaß für stetige und ordinale Felder angeben. Nominale Felder werden auf der Basis des Modalwerts aggregiert.

Stetig Bei stetigen (metrischen) Feldern kann das Auswertungsmaß **Mittelwert**, **Median** oder **Summe** sein.

Ordinal
Bei ordinalen Feldern kann das Auswertungsmaß **Mittelwert**, **Modus**, **Größter** oder **Kleinster** sein.

Anzahl der Klassen
Legt die maximale Anzahl der Klassen für stetige (metrische) Felder fest. Stetige Felder werden immer in Wertebereichen gruppiert oder "klassiert". Beispiel: kleiner-gleich 5, größer als 5 und kleiner-gleich 10 oder größer als 10.

Karte aggregieren
Die Aggregation wird sowohl auf Daten als auch auf Karten angewendet.

Benutzerdefinierte Einstellungen für bestimmte Felder
Sie können das Standardauswertungsmaß und die Anzahl der Klassen für bestimmte Felder überschreiben.

- Klicken Sie auf das Symbol, um das Dialogfeld **Feldauswahl** zu öffnen und der Liste ein Feld hinzuzufügen.
- Wählen Sie in der Spalte **Aggregation** ein Auswertungsmaß aus.
- Klicken Sie bei stetigen Feldern auf die Schaltfläche in der Spalte **Klassen**, um eine benutzerdefinierte Anzahl von Klassen für das Feld im Dialogfeld **Klassen** anzugeben.

Räumlich-temporale Vorhersage

Bei räumlich-temporalen Vorhersagen bleiben nach der Definition von Karten und Datenquellen folgende Schritte im Assistenten:

- Geben Sie das Zielfeld, Zeitfelder und optionale Prädiktoren an.
- Definieren Sie Zeitintervalle oder zyklische Perioden für Zeitfelder.

Außerdem sind die folgenden Optionen verfügbar:

- Wählen Sie andere Ausgabeoptionen aus.
- Passen Sie die Modellerstellungsparameter an.
- Passen Sie die Aggregationseinstellungen an.
- Speichern Sie vorhergesagte Werte in einem Dataset in der aktuellen Sitzung oder in einer Datendatei im IBM SPSS Statistics-Format.

Auswählen von Feldern

Die Liste verfügbarer Felder umfasst Felder von den ausgewählten Datenquellen. Sie können die Liste der angezeigten Felder steuern, indem Sie eine Datenquelle in der Liste **Datenquellen** auswählen.

Ziel Es ist ein Zielfeld erforderlich. Das Ziel ist das Feld, für das Werte vorhergesagt werden.

- Das Zielfeld muss ein stetiges (metrisches) numerisches Feld sein.

- Wenn es zwei Datenquellen gibt, enthält das Ziel Kerndichteschätzungen und als Zielname wird "Dichte" angezeigt. Diese Auswahl kann nicht geändert werden.

Prädiktoren
Es können ein Prädiktorfeld oder mehrere Prädiktorfelder angegeben werden. Diese Einstellung ist optional.

Zeitfelder
Sie müssen mindestens ein Feld auswählen, das Zeiträume darstellt, oder **Zyklische Periode** auswählen.

- Bei zwei Datenquellen müssen Sie Zeitfelder aus beiden Datenquellen auswählen. Beide Zeitfelder müssen dasselbe Intervall darstellen.
- Bei zyklischen Perioden müssen Sie die Felder angeben, die die Periodizitätszyklen im Fenster **Zeitintervalle** des Assistenten definieren.

Zeitintervalle

Die Optionen in diesem Fenster basieren auf der Auswahl für **Zeitfelder** oder **Zyklische Periode** im Schritt zur Auswahl von Feldern.

Zeitfelder

Ausgewählte Zeitfelder. Wenn Sie im Schritt zur Auswahl von Feldern ein oder mehrere Zeitfelder auswählen, werden diese Felder in dieser Liste angezeigt.

Zeitintervall. Wählen Sie das entsprechende Zeitintervall aus der Liste aus. Je nach Zeitintervall können Sie auch andere Einstellungen angeben, beispielsweise das Intervall zwischen Beobachtungen (Inkrement) und den Anfangswert. Dieses Zeitintervall wird für alle ausgewählten Zeitfelder verwendet.

- Die Prozedur setzt voraus, dass alle Fälle (Datensätze) Intervalle mit gleichen Abständen darstellen.
- Auf der Basis des ausgewählten Zeitintervalls kann die Prozedur fehlende Beobachtungen oder mehrfache Beobachtungen in demselben Zeitintervall ermitteln, die aggregiert werden müssen. Wenn das Zeitintervall beispielsweise **Tage** ist und nach dem Datum **2014-10-27** das Datum **2014-10-29** folgt, fehlt die Beobachtung für das Datum **2014-10-28**. Wenn das Zeitintervall **Monat** ist, werden mehrere Datumsangaben in demselben Monat aggregiert.
- Bei einigen Zeitintervallen kann die zusätzliche Einstellung Pausen in den normalen Zeitintervallen mit gleichen Abständen definieren. Wenn das Zeitintervall beispielsweise **Tage** ist, aber nur Wochentage gültig sind, können Sie angeben, dass die Woche fünf Tage umfasst und am Montag beginnt.
- Wenn das ausgewählte Zeitfeld kein Datumsformat- oder Zeitformatfeld ist, wird das Zeitintervall automatisch auf **Perioden** gesetzt und kann nicht geändert werden.

Zyklusfelder

Wenn Sie im Schritt zur Auswahl von Feldern **Zyklische Periode** auswählen, müssen Sie die Felder angeben, die die zyklischen Perioden definieren. Eine zyklische Periode gibt eine sich wiederholende zyklische Schwankung an, beispielsweise bei der Anzahl der Monate in einem Jahr oder bei der Anzahl der Tage in einer Woche.

- Sie können bis zu drei Felder angeben, die zyklische Perioden definieren.
- Das erste Zyklusfeld stellt die höchste Ebene des Zyklus dar. Bei einer zyklischen Schwankung bei **Jahr**, **Quartal** und **Monat** ist beispielsweise das Feld, das das Jahr darstellt, das erste Zyklusfeld.
- Die Zykluslänge für das erste und zweite Zyklusfeld ist die Periodizität auf der nachfolgenden Ebene. Wenn die Zyklusfelder beispielsweise **Jahr**, **Quartal** und **Monat** sind, beträgt die erste Zykluslänge 4 und die zweite Zykluslänge 3.
- Der Anfangswert für das zweite und dritte Zyklusfeld ist der erste Wert in jeder dieser zyklischen Perioden.
- Die Zykluslänge und die Anfangswerte müssen positive ganze Zahlen sein.

Aggregation

- Wenn Sie im Schritt zur Auswahl von Feldern **Prädiktoren** auswählen, können Sie die Aggregationsauswertungsmethode für die Prädiktoren auswählen.
- Die Aggregation ist erforderlich, wenn ein definiertes Zeitintervall mehrere Datensätze aufweist. Beim Zeitintervall **Monat** werden mehrere Datumsangaben in demselben Monat aggregiert.
- Sie können die Methode für das Aggregationsauswertungsmaß für stetige und ordinale Felder angeben. Nominale Felder werden auf der Basis des Modalwerts aggregiert.

Stetig Bei stetigen (metrischen) Feldern kann das Auswertungsmaß **Mittelwert**, **Median** oder **Summe** sein.

Ordinal
Bei ordinalen Feldern kann das Auswertungsmaß **Mittelwert**, **Modus**, **Größter** oder **Kleinster** sein.

Benutzerdefinierte Einstellungen für bestimmte Felder
Sie können das Standardaggregationsauswertungsmaß für bestimmte Prädiktoren überschreiben.

- Klicken Sie auf das Symbol, um das Dialogfeld **Feldauswahl** zu öffnen und der Liste ein Feld hinzuzufügen.
- Wählen Sie in der Spalte **Aggregation** ein Auswertungsmaß aus.

Ausgabe

Karten

Zielwerte.
Karte der Werte für das ausgewählte Zielfeld.

Korrelation
Karte der Korrelationen.

Gruppen
Karte, die Cluster von Positionen hervorhebt, die einander ähnlich sind.

Schwellenwert für Ähnlichkeit von Positionen.
Die zum Erstellen von Clustern erforderliche Ähnlichkeit. Der Wert muss eine Zahl größer als 0 und kleiner als 1 sein.

Maximale Anzahl der Cluster angeben.
Die maximale Anzahl der anzuzeigenden Cluster.

Modellevaluierungstabellen

Modellspezifikationen.
Zusammenfassung der zum Ausführen der Analyse verwendeten Spezifikationen, einschließlich der Felder **Ziel**, **Eingabe** und **Position**.

Temporale Informationszusammenfassung.
Gibt die im Modell verwendeten Zeitfelder und Zeitintervalle an.

Test der Auswirkungen auf die Mittelwertstruktur.
Die Ausgabe umfasst Teststatistikwert, Freiheitsgrade und Signifikanzniveau für das Modell und alle Effekte.

Mittelwertstruktur der Modellkoeffizienten.
Die Ausgabe umfasst Koeffizientenwert, Standardfehler, Teststatistikwert, Signifikanzniveau und Konfidenzintervalle für alle Modellterme.

Autoregressive Koeffizienten.
Die Ausgabe umfasst Koeffizientenwert, Standardfehler, Teststatistikwert, Signifikanzniveau und Konfidenzintervalle für jeden Lag.

Tests der räumlichen Kovarianz.
Bei variogrammbasierten parametrischen Modellen werden die Ergebnisse des Tests zur Anpassungsgüte für die räumliche Kovarianzstruktur angezeigt. Anhand der Testergebnisse kann festgelegt werden, ob die räumliche Kovarianzstruktur parametrisch modelliert wird oder ein nicht parametrisches Modell verwendet wird.

Parametrische räumliche Kovarianz.
Bei variogrammbasierten parametrischen Modellen werden Parameterschätzungen für parametrische räumliche Kovarianz angezeigt.

Modelloptionen

Modelleinstellungen

Automatisch konstanten Term einschließen
Der konstante Term wird im Modell eingeschlossen.

Maximale automatische Verzögerung bei Regressionen
Die maximale automatische Verzögerung bei Regressionen. Der Wert muss eine ganze Zahl zwischen 1 und 5 sein.

Räumliche Kovarianz
Gibt die Schätzmethode für die räumliche Kovarianz an.

Parametrisch
Die Schätzmethode ist parametrisch. Die Methode kann **Gauß**, **Exponentiell** oder **Potenzgesetz** sein. Bei **Potenzgesetz** können Sie den Wert von **Potenz** angeben.

Nicht parametrisch
Die Schätzmethode ist nicht parametrisch.

Speichern

Karten- und Kontextdaten als Kartenspezifikation speichern
Speichern Sie die Kartenspezifikation in einer externen Datei (.mplan). Sie können diese Kartenspezifikationsdatei für spätere Analysen in den Assistenten laden. Sie können die Kartenspezifikationsdatei auch mit dem Befehl `SPATIAL TEMPORAL PREDICTION` verwenden.

Karte und Kontextdatendateien in die Spezifikation kopieren
Daten aus Kartenshapedateien, externen Datendateien und in der Kartenspezifikation verwendeten Datasets werden in der Kartenspezifikationsdatei gespeichert.

Scoring
Speichert die vorhergesagten Werte, die Varianz sowie die obere und untere Konfidenzgrenze für das Zielfeld in der ausgewählten Datendatei.

- Sie können vorhergesagte Werte in einem geöffneten Dataset in der aktuellen Sitzung oder in einer Datendatei im IBM SPSS Statistics-Format speichern.
- Die Datendatei kann keine im Modell verwendete Datenquelle sein.
- Die Datendatei muss alle Zeitfelder und Prädiktoren enthalten, die im Modell verwendet werden.
- Die Zeitwerte müssen größer als die Zeitwerte sein, die im Modell verwendet werden.

Erweitert

Max. Fälle mit fehlenden Werten (%)
Der maximale Prozentsatz der Fälle mit fehlenden Werten.

Signifikanzniveau
Das Signifikanzniveau zum Feststellen, ob ein variogrammbasiertes parametrisches Modell geeignet ist. Der Wert muss größer als 0 und kleiner als 1 sein. Der Standardwert ist 0,05. Das Signifi-

kanzniveau wird im Test zur Anpassungsgüte für die räumliche Kovarianzstruktur verwendet. Mithilfe der Statistik für Anpassungsgüte wird bestimmt, ob ein parametrisches oder ein nicht parametrisches Modell verwendet werden soll.

Unsicherheitsfaktor (%)
Der Unsicherheitsfaktor ist ein Prozentwert, der die Zunahme der Unsicherheit für zukünftige Vorhersagen darstellt. Die Ober- und Untergrenze der Vorhersageunsicherheit nimmt bei jedem in die Zukunft reichenden Schritt um den angegebenen Prozentsatz zu.

Fertigstellen

Im letzten Schritt des Geomodellierungsassistenten können Sie das Modell entweder ausführen oder die erstellte Befehlssyntax in ein Syntaxfenster einfügen. Sie können die erstellte Syntax ändern und für die spätere Verwendung speichern.

Bemerkungen

Die vorliegenden Informationen wurden für Produkte und Services entwickelt, die auf dem deutschen Markt angeboten werden.

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
France

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und anderen Programmen (einschließlich des vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
USA

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Alle in diesem Dokument enthaltenen Leistungsdaten stammen aus einer kontrollierten Umgebung. Die Ergebnisse, die in anderen Betriebsumgebungen erzielt werden, können daher erheblich von den hier erzielten Ergebnissen abweichen. Einige Daten stammen möglicherweise von Systemen, deren Entwicklung noch nicht abgeschlossen ist. Eine Gewährleistung, dass diese Daten auch in allgemein verfügbaren Systemen erzielt werden, kann nicht gegeben werden. Darüber hinaus wurden einige Daten unter Umständen durch Extrapolation berechnet. Die tatsächlichen Ergebnisse können davon abweichen. Benutzer dieses Dokuments sollten die entsprechenden Daten in ihrer spezifischen Umgebung prüfen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

COPYRIGHTLIZENZ:

Diese Veröffentlichung enthält Beispielanwendungsprogramme, die in Quellensprache geschrieben sind und Programmiertechniken in verschiedenen Betriebsumgebungen veranschaulichen. Sie dürfen diese Beispielprogramme kostenlos kopieren, ändern und verteilen, wenn dies zu dem Zweck geschieht, Anwendungsprogramme zu entwickeln, zu verwenden, zu vermarkten oder zu verteilen, die mit der Anwendungsprogrammierschnittstelle für die Betriebsumgebung konform sind, für die diese Beispielprogramme geschrieben werden. Diese Beispiele wurden nicht unter allen denkbaren Bedingungen getestet. Daher kann IBM die Zuverlässigkeit, Wartungsfreundlichkeit oder Funktion dieser Programme weder zusagen noch gewährleisten. Die Beispielprogramme werden ohne Wartung (auf "as-is"-Basis) und ohne jegliche Gewährleistung zur Verfügung gestellt. IBM übernimmt keine Haftung für Schäden, die durch die Verwendung der Beispielprogramme entstehen.

Kopien oder Teile der Beispielprogramme bzw. daraus abgeleiteter Code müssen folgenden Copyrightvermerk beinhalten:

© (Name Ihrer Firma) (Jahr). Teile des vorliegenden Codes wurden aus Beispielprogrammen der IBM Corporation abgeleitet.

Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corp in den USA und/oder anderen Ländern. Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie auf der Webseite „Copyright and trademark information" unter www.ibm.com/legal/copytrade.shtml.

Adobe, das Adobe-Logo, PostScript und das PostScript-Logo sind Marken oder eingetragene Marken der Adobe Systems Incorporated in den USA und/oder anderen Ländern.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Markten oder eingetragene Marken der Intel Corporation oder der Tochtergesellschaften des Unternehmens in den USA und anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA, anderen Ländern oder beidem.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Java und alle auf Java basierenden Marken und Logos sind Marken oder eingetragene Marken der Oracle Corporation und/oder ihrer verbundenen Unternehmen.

Index

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z