G Wichtige Neuerungen bei Version 8

Abgesehen von der strengeren Lizenzierung fallen beim Start von Version 8 erst einmal keine großen Neuerungen auf. Einzig vielleicht, dass es nun ein eigenständiges Data-Mining-Menü gibt. Entsprechend wurde das Statistik-Menü um diese Einträge bereinigt und insgesamt etwas übersichtlicher gestaltet. Kurz gesagt 'drohen' Benutzern früherer Versionen von STATISTICA erst einmal keine großen Umgewöhnungen, Änderungen gibt es im Detail dann aber doch. Und genau diese sollen im Folgenden kurz besprochen werden. Dabei sollen zuerst Änderungen vorgestellt werden, die in irgendeiner Weise Einfluss auf die Verwaltung von Daten haben, siehe Abschnitt G.1, anschließend solche, die für die Datenanalyse relevant sind, siehe Abschnitt G.2. Abschnitt G.3 stellt neue Funktionalitäten vor, mit denen man ausgegebene Resultate bearbeiten kann, und Abschnitt G.4 kehrt nochmals zurück zum Thema Datenanalyse, indem speziell Neuerungen vorgestellt werden, welche die industrielle Statistik betreffen.

G.1 Datenverwaltung

Der *Import aus Textdateien*, siehe Kapitel 1, wurde völlig neu gestaltet. Nachdem man bei *Datei* → *Öffnen* eine Textdatei ausgewählt hat, wird man im zuerst erscheinenden Dialog gefragt, ob die Textdatei als Tabelle oder Bericht importiert werden soll. Betrachten wir den Fall *Import als Tabelle: Frei*. Dann erscheint anschließend ein sehr umfangreicher Dialog, mit dessen Hilfe man eine Reihe von Einstellungen bzgl. Trennzeichen, Zeilenzahl, u. Ä., treffen kann. Dabei wird dem Benutzer automatisch ein Vorschlag gemacht, den man auf Wunsch aber auch abändern kann: Im obersten Teil des Dialogfensters kann man das Trennzeichen auswählen, und im Feld der Import-Optionen festlegen, bei welchem Fall der Import beginnen/enden soll (indem man die Zahl der zu Beginn zu überspringenden Fälle festlegt und die anschließend zu importierende Fallzahl), welches Dezimaltrennzeichen verwendet wird, u. Ä. Im unteren Teil des Fensters gibt es eine Vorschau, die sich auch automatisch aktualisiert, falls man ein Häkchen bei der entsprechenden Option ganz unten gesetzt hat. Nach Klick auf *OK* startet dann der Import. Für das Beispiel der Pisa-Daten aus Kapitel 1 wären folgende Einstellungen passend: *Trennzeichen Variable: Tabulator* und Häkchen

im Feld *Optionen Import* genau bei *Variablennamen aus erster Zeile* und *Anzahl ... überspringen: 3*. Ferner müsste in der Vorschau die erste Spalte (mit den Namen der Bundesländer) markiert und dann im Feld *Variable 1-Optionen* der *Datentyp: Fallname* gewählt werden.

Bei EXCEL-Dateien ist dagegen ein Import, vgl. Abschnitt 2.2, nicht mehr zwangsweise nötig, dank der neuen *Office-Integration*, der wohl augenscheinlichsten Neuerung bei Version 8. Die erweiterte Office-Integration (vgl. auch Abschnitt 2.2) betrifft dabei ausschließlich das Produkt Microsoft Office, und dort die beiden Programme WORD und EXCEL. Sie kommt jedoch nur zur Geltung, falls die beiden Programme auch auf dem Rechner des Benutzers installiert sind. In diesem Fall wird immer dann, wenn einer der beiden Dateitypen in STATISTICA aktiv ist, das jeweilige Office-Menü in das STATISTICA-Menü integriert, zusammen mit den individuellen Einstellungen des Benutzers. Wechselt man dann in STATISTICA wieder zu einem STATISTICA-eigenen Dateityp, wird wieder das gewöhnliche STATISTICA-Menü gezeigt.

Betrachten wir vorerst also den Fall einer EXCEL-Datei; auf WORD-Dateien kommen wir in Abschnitt G.3 zu sprechen. Wählt man über *Datei* → *Öffnen* eine EXCEL-Datei aus, so wird man im ersten Dialog gefragt, ob man diese wie bisher in eine Arbeitsmappe oder Tabelle importieren möchte, oder, und diesen Fall betrachten wir im Folgenden, ob man die Datei als EXCEL-Arbeitsmappe öffnen möchte. Dann erscheint die EXCEL-Datei in einem eigenen Fenster und das Menü passt sich, wie oben beschrieben, dem EXCEL-Menü an. Auch über *Datei* → *Neu*, Karte *Office Dokument*, kann man eine EXCEL-Datei erzeugen. Diese kann man, wie von EXCEL gewohnt, direkt bearbeiten, oder eben mit STATISTICA analysieren. Zu Beginn der ersten Analyse muss man allerdings noch den zu analysierenden Bereich spezifizieren, wobei der entsprechende Dialog dem Importdialog recht ähnlich gestaltet ist, und kann anschließend noch den Variablentyp prüfen. Wurde dies einmal gemacht, können anschließend alle weiteren Analysen ohne Umweg ausgeführt werden.

Neu sind auch verbesserte Möglichkeiten zur *Datenaufbereitung*, verfügbar über den Menüpunkt *Daten* → *Daten filtern/umkodieren*. So kann man etwa im Untermenü *Ausreißer umkodieren* auf die gewählten Variablen und Fälle des Datenblatts individuell Ausreißertests anwenden, etwa die von Grubbs oder Tukey, die gefundenen Ausreißer umkodieren, und das gesamte Prozedere auch iterativ wiederholen lassen. Ferner können *Dubletten*, also identische Fälle, aus den Daten herausgefiltert werden. Auch wird angeboten, Variablen und Fälle zu entfernen, die einen vorgebbaren Anteil an Fehldaten überschreiten, bzw. die Fehldaten z. B. durch das arithmetische Mittel der jeweiligen Datenreihe ersetzen zu lassen.

Eine andere, sehr nützliche Neuerung stellen die neuen *Funktionen für Datentabellen* dar, siehe auch Abschnitt 3.2. Zu diesen zählen etwa die Funktion `Cusum`, mit der man die Werte einer Variable kumulativ aufsummieren kann, eine Reihe von Datumsfunktionen, neue Rundungsfunktionen, die Johnson-Verteilungsfamilie, vor allem aber die Funktion `Data(vx;n)`: Mit deren Hilfe kann man nun endlich auch einzelne Zellen ansprechen, im Beispiel den Wert der Variablen Nummer x, Zeile n.

Über das Menü *Daten* wird neuerdings die *Box-Cox-Transformation* angeboten, deren Anwendung auf eine Varianzstabilisierung und Normalisierung der Daten abzielt.

Hintergrund G.1.1

Die *Box-Cox-Transformationen* können als eine Verallgemeinerung der elementaren Logarithmus- und Wurzeltransformation aufgefasst werden, die ebenfalls zum Zwecke der Varianzstabilisierung und Normalisierung der Daten eingesetzt werden. In ihrer allgemeinsten Form ist die Transformationsfunktion gegeben zu

$$f(x) \;:=\; \begin{cases} ((x+\alpha)^{\lambda}-1)/\lambda, & \lambda \neq 0, \\ \ln(x+\alpha), & \lambda = 0. \end{cases}$$

α ist hierbei ein vom Benutzer vorzugebender Shiftparameter (Voreinstellung: $\alpha = 0$), der Transformationsparameter λ dagegen ist so zu wählen, dass die Daten bestmöglichst normalisiert werden. Dies wird durch einen Maximum-Likelihood-Ansatz realisiert, bzw. durch ein äquivalentes Fehlerminimierungskriterium. ◇

Die Fehlerminimierung wird dabei von STATISTICA automatisch durchgeführt: Im Dialog *Daten* → *Box-Cox-Transformation* wählt der Benutzer die zu transformierenden Variablen aus, kann die Voreinstellungen des Suchalgorithmus bei Bedarf modifizieren, wobei λ maximal im Bereich $[-5;5]$ gesucht wird, und ggf. einen Shiftparameter voreinstellen. Nach *OK* wird für eine jede Variable ein optimales λ bestimmt, was der Benutzer über den Knopf *Zusammenfassung* einsehen kann: Die erste ausgegebene Tabelle enthält jeweils die Original- und die transformierten Daten, die zweite den optimalen Parameterwert samt Konfidenzintervall *und* der fertigen Transformationsformel in der letzten Spalte, so dass der Benutzer die Transformation auch manuell ausführen kann, vgl. Abschnitt 3.2. Über den Knopf *Histogramme ...* kann man sich von der Güte der Normalisierung überzeugen. Vorsicht ist geboten, wenn das gefundene λ auf dem Rand des Suchraums liegt; ideal ist es dagegen, wenn die Grafiken, die man sich über *Plots Suchhistorie* ausgeben lassen kann, das gefundene λ als klares, lokales Minimum ausweisen. Schließlich bietet der Ergebnisdialog auch die Möglichkeit, die gefundenen transformierten Daten in die Originaltabelle schreiben zu lassen.

Zu guter Letzt sei noch erwähnt, dass berechnete *Statistiken für Blockdaten* nun nicht mehr an die Originaltabelle angehängt, sondern in einer eigenen, frei stehenden Tabelle ausgegeben werden. Somit wird verhindert, dass diese Statistiken ungewollt in andere Berechnungen einfließen.

G.2 Datenanalyse

Mit Version 8 wurde ein völlig neuer Dateityp eingeführt, die *Projektdatei* mit der Endung `.spf`. Diese erlaubt es, die aktuelle Arbeitsumgebung zu bewahren, selbst noch nicht gespeicherte Daten sowie aktive Analysen werden gesichert und später wiederhergestellt. Nachdem ein Projekt erneut geöffnet wurde, werden neue Resultate wieder in die ehemals aktive Arbeitsmappe eingefügt.

Positiv bemerkenswert ist, dass die *gruppenweisen Analysen* nun in die jeweiligen Dialoge integriert wurden. Dort gibt es jetzt jeweils einen *Gruppen...*-Knopf, der nach Anklicken in einen Dialog führt, bei dem die Gruppierungsvariablen ausgewählt werden müssen. Ferner kann man festlegen, ob die sich ergebenden Gruppen sortiert werden sollen, und ob die Ausgabe z. B. in lauter einzelne Ordner erfolgen soll. Trotzdem wurde die Lösung aus Version 7, ein zentraler Dialog *Batch-Analyse für Gruppen*, aus Kompatibilitätsgründen zusätzlich beibehalten, siehe auch Abschnitt 5.1, S. 76ff.

Auch eine Reihe neuer *Statistiken und Konfidenzintervalle* werden nun angeboten. Im Menü der deskriptiven Statistik, siehe Abschnitt 5.1, gibt es eine Karte *Robust*, welche das getrimmte und das winsorisierte Mittel anbietet, sowie den Grubbs-Test zum Auffinden von Ausreißern. Die Karte *Details* wurde um ein Konfidenzintervall für die Standardabweichung S_n (*KI für Stichpr.stdabw.*) sowie um den *Variationskoeffizienten* $S_n/\bar{X}_n$ ergänzt.

Hintergrund G.2.1

α-getrimmtes bzw. winsorisiertes Mittel können als Kompromiss zwischen Mittelwert und Median aufgefasst werden, wobei der Anteil α vom Benutzer vorgegeben werden muss. Beim α-getrimmten Mittel werden die jeweils $\alpha \cdot 100\,\%$ kleinsten und größten Datenwerte (welche potentielle Ausreißer enthalten können) aus dem Datensatz entfernt, und dann vom Rest das arithmetische Mittel berechnet. Beim winsorisierten Mittel werden diese Randwerte nicht entfernt, sondern entsprechend oft durch den nächstkleinsten bzw. -größten Wert ersetzt, und dann gemittelt.

Das Konfidenzintervall für die Standardabweichung baut auf der Annahme auf, dass der Datensatz $X_1, \ldots, X_n$ i.i.d. normalverteilt ist gemäß $N(\mu, \sigma^2)$; in diesem Fall ist nämlich $(n-1) \cdot S_n^2/\sigma^2$ χ^2-verteilt mit $n-1$ Freiheitsgraden, vgl. Hintergrund 12.1.3.12. Unter dieser Annahme ergibt sich also

ein Konfidenzintervall für σ zum Vertrauensniveau $1-\alpha$ als

$$\left[S_n \cdot \sqrt{(n-1)/\chi^{-1}_{n-1}(1-\tfrac{\alpha}{2})}\,;\, S_n \cdot \sqrt{(n-1)/\chi^{-1}_{n-1}(\tfrac{\alpha}{2})}\right],$$

wobei $\chi^{-1}_{n-1}(\beta)$ das β-Quantil der χ^2_{n-1}-Verteilung bezeichnet. ◇

Auch der Grubbs-Test baut auf der genannten Normalverteilungsannahme auf und kann pro Durchlauf einen Ausreißer anzeigen. Ferner wird nun auch ein Knopf *Grafiken* angeboten, der für alle Variablen eine Grafikkombination aus einem Histogramm mit Anpassung Normalverteilung, einem Normalverteilungsplot, einem Boxplot und einem Feld mit den berechneten Werten der Statistiken erzeugt.

In puncto Grafiken sind der *2D-Scatterplot mit Fehlerbalken*, der *Bagplot* und die *Wafergrafiken* zu erwähnen. Die ersten beiden Grafiktypen sind bei den *2D-Grafiken* angesiedelt, die dritte bei den *3D XYZ-Grafiken*. Ein 2D-Scatterplot mit Fehlerbalken ist ein Scatterplot, bei dem Fehlerbalken für all jene Werte auf der X-Achse erzeugt werden, die mehr als einen zugehörigen Y-Wert besitzen. Per Voreinstellung ergeben sich die Fehlerbalken als 95 %-Konfidenzintervall um den Mittelwert der jeweiligen Y-Werte herum, das Ganze ist aber auch nachträglich über die Grafikoptionen änderbar, vgl. Kapitel 4. Der Bagplot stellt eine zweidimensionale Verallgemeinerung der Boxplots dar, siehe auch Abschnitt 5.3: Im Zentrum des Bagplots findet sich eine zweidimensionale Variante des Medians wieder, darum herum eine 'Tasche', die 50 % der Daten umfasst. Um diese herum wird der 'Zaun' (engl.: fence) angelegt, der durch Ausdehnung der Tasche um den Faktor 1,5 entsteht. Beide Vorgaben lassen sich auf der Karte *Details* auch ändern, indem man die Werte bei *Bag-Koeff.* bzw. *Schranken-Koeff.* anpasst. Punkte, die außerhalb des Zaunes liegen, sind als ausreißerverdächtig einzustufen. Die *Wafergrafik* dagegen ist eine zweidimensionale Grafik, die entsteht, wenn man auf den 3D-Flächenplot mit Anpassung Wafer aufblickt, vgl. Durchführung 11.2.1.

Schließlich gab es im Bereich der *industriellen Statistik*, vgl. Abschnitt 12, einige bedeutende Ergänzungen. Da diese Neuerungen sehr umfassend ausgefallen sind, werden wir sie erst in Abschnitt G.4 in der gebotenen Ausführlichkeit besprechen.

Wer statische Verfahren benötigt, die nicht in STATISTICA verfügbar sind, wird möglicherweise bei R fündig. *R* ist eine gemäß der GNU General Public License frei erhältliche Softwareumgebung für statistische Berechnungen und Grafiken. Bei dieser Software handelt es sich um eine Kommandozeilenumgebung, welche die gleichnamige Sprache R verwendet, und welche insbesondere im Hochschulbereich zunehmende Verbreitung erfährt. Neben dem eigentlichen Grundprogramm, welches

bereits elementare Funktionalitäten anbietet, gibt es eine rasant wachsende Zahl ebenfalls frei verfügbarer Pakete zu Themen wie Biostatistik, Data Mining oder Statistischer Prozesskontrolle (SPC), mit denen R nach Bedarf erweitert werden kann. Und derartige R-Pakete kann man auch mit STATISTICA nutzen. Seit Version 6 erlauben es die COM-Schnittstelle und STATISTICA Visual Basic, recht komfortabel auf Funktionalitäten von R zuzugreifen. Dazu muss jedoch neben R und den gewünschten Paketen auch der *R (D)COM Server* von Thomas Baier und Erich Neuwirth installiert sein, welcher unter

`http://cran.r-project.org/contrib/extra/dcom/`

bzw. `http://sunsite.univie.ac.at/rcom/`

verfügbar ist. Seit dem Upgrade MR3 der Version 8 ist diese Integration der Sprache R nochmals erleichtert worden, indem nun

1. der R (D)COM Server automatisch bei der Installation von STATISTICA mitinstalliert wird, und
2. zahlreiche neue Befehle zur Einbettung in den Visual Basic- und/oder den R-Code geschaffen wurden, welche die Kommunikation mit R erheblich erleichtern.

Vertiefende Informationen hierzu findet man im Hilfe-Menü von STATISTICA.

G.3 Analyseresultate

Zu erstellten Analyseresultaten wird nun ein Makro mit in die Arbeitsmappe abgelegt, so dass man einmal gemachte *Analysen mit neuen/geänderten Daten wiederholen oder fortsetzen* kann. Erkennbar ist dies am roten Pfeil der Arbeitsmappenordner, den man mit der rechten Maustaste anklickt. Wählt man im sich öffnenden PopUp-Menü den Punkt *Analyse wiederholen*, so wird die Analyse, die zu den gemachten Ausgaben führte, wiederholt, wobei man optional auch einen anderen Datensatz wählen kann. Ferner kann man wählen, dass die alten Resultate durch die neuen ersetzt werden. Die Wahl von *Analyse fortsetzen* erlaubt es einem dagegen, die Analyse mit geänderten Einstellungen zu wiederholen.

Arbeitsmappen insgesamt kann man nun auch *als HTML speichern*. Bisher war dies nur für die einzelnen Elemente einer Arbeitsmappe möglich, siehe Abschnitt 2.3. Das Resultat dieses Exports ist exakt genauso in Baumstruktur organisiert wie die originale Arbeitsmappe. Hierbei werden die Tabellen in einzelne HTML-Dateien exportiert, die Grafiken in PNG- bzw. JPG-Dateien (je nach Wahl in den *Optionen* auf der

Karte *Arbeitsmappen*), und Logos in GIF-Dateien. Dabei wird man vor Durchführung des Exports gefragt, für welche HTML-Version das Ganze erzeugt werden soll. Es empfiehlt sich die Wahl von *Internet Explorer 7/Firefox/Opera*.

Wie in Abschnitt G.1 bereits erwähnt, ist die Office-Integration eine der bedeutensten Neuerungen in Version 8. Insbesondere lässt sich nun auch *WORD integrieren*. Dabei gibt es drei Situationen, in welchen ein WORD-Dokument zum aktiven Dokument in STATISTICA wird: Man erstellt in STATISTICA ein WORD-Dokument über *Datei* → *Neu*, man öffnet eine bestehende WORD-Datei über *Datei* → *Öffnen*, oder man lässt Analyseresultate in eine WORD-Datei ablegen. Letzteres spiegelt zugleich die Hauptanwendung wider, WORD-Dateien als Alternative zu den STATISTICA-eigenen Berichtsdateien. Um ein Resultat einzufügen, muss dieses als frei stehendes Fenster existieren, d. h. notfalls muss es aus einer Arbeitsmappe extrahiert werden, vgl. Durchführung 2.3.1. Ist dieses dann aktiv, wählt man den Punkt *Zur MS-Word-Datei*. Nun wird das Resultat entweder in eine neu zu erzeugende WORD-Datei eingefügt, oder in eine bereits geöffnete. Alternativ zum eben beschriebenen Vorgehen gibt es natürlich auch den Weg über Kopieren und Einfügen.

Neu ist auch, dass man gleichartige *Grafiken überlagern* kann. Dazu klickt man in die Zielgrafik mit der rechten Maustaste hinein, wählt im sich öffnenden PopUp-Menü den Punkt *Grafiken zusammenfügen...*, und dann die zu überlagernde Grafik. Alternativ kann man die Zielgrafik auch einfach nur markieren und wählt dann *Extras* → *Grafiken zusammenfügen*. Zudem wird nun auch angeboten, Grafiken als separate Objekte in andere Grafiken hinein zu kopieren. Dazu klickt man nach dem Kopieren der Quellgrafik in die Zielgrafik hinein und wählt dann Einfügen (*Strg*+*V*). Evtl. wird man dabei gefragt, ob die Grafik als separates Objekt eingefügt oder überlagert werden soll.

Weitere Neuerungen betreffen die Grafiktypen Quantil- und Probabilitytyplot, welche bisher nur für einzelne Variablen erzeugt werden konnten. Nun wurde eine Option ergänzt, dass bei Auswahl mehrerer Variablen alle Plots in eine Grafik gezeichnet werden. Zudem kann man beim Export aller Grafiktypen in eine Rastergrafik die Auflösung frei wählen, zwischen 72 und 7200 dpi.

G.4 Industrielle Statistik

STATISTICAs Fähigkeit zur statistischen Qualitätskontrolle, siehe Abschnitt 12, wurde mit Version 8 in zwei Punkten wesentlich erweitert: Zu den bisherigen univariaten Kontrollkarten und der etwas versteckten Hotelling-Karte für multivariate Prozesse, siehe Abschnitt 12.1, wird nun ein eigenständiges Menü zur multivariaten Prozesskontrolle angeboten,

mit Karten wie MEWMA oder MCUSUM. Darauf wollen wir in Abschnitt G.4.1 ausführlicher eingehen. Daneben wurde auch das Menü zur Prozessfähigkeitsanalyse, siehe Abschnitt 12.1.8, entscheident erweitert: Das neue Menü zur Prozessfähigkeit gemäß ISO 21747/DIN 55319 setzt das in diesen Normen beschriebene Verfahren, mit den dortigen Bezeichnungsweisen, um, so dass sich ein normgerechtes Vorgehen mit ein paar Mausklicks bewältigen lässt. Mehr hierzu in Abschnitt G.4.2. Abschließend sei bemerkt, dass das letztgenannte Menü auch in Hinblick auf Attributdaten (Präzision, Übereinstimmung) erweitert wurde.

G.4.1 Multivariate Kontrollkarten

Neben dem bisherigen Menü zu univariaten Kontrollkarten bietet STATISTICA nun auch eines zu multivariaten Karten an, über *Statistik* → *Industrielle Statistik & Six Sigma* → *Multivariate Qualitätsregelkarten*. Dieses bietet *Hotellings T^2-Karte* für Einzelwerte oder Mittelwerte an, wobei die Karte im zweiten Fall um die *Karte für verallgemeinerte Varianzen (GV)* ergänzt wird. Auch die *MEWMA-Karte* wird für die zwei genannten Situationen angeboten, ferner die *Multivariate CUSUM-Karte* für Einzelwerte.

Hintergrund G.4.1.1

Sei $\boldsymbol{X}_{t,1}, \ldots, \boldsymbol{X}_{t,n}$, mit Zielbereich $\mathbb{R}^p$, die zur Zeit t erhobene Teilstichprobe gemäß Voraussetzung 12.1.6.1, dann ist auch das Stichprobenmittel $\bar{\boldsymbol{X}}_t$ multivariat normalverteilt. *Hotellings T^2-Karte* wird über das neue Menü konstruiert wie in Hintergrund 12.1.6.2 beschrieben, allerdings mit zwei Unterschieden: Erstens wird bei Einzelwerten ($n = 1$) in Phase I eine Betaverteilung mit $p/2$ und $2(m-1)^2/(3m-4)$ Freiheitsgraden zu Grunde gelegt, zweitens verwendet STATISTICA im neuen Menü, auch bei den anderen Einzelwertkarten, einen anderen Varianzschätzer. Wurde für das in Durchführung 12.1.6.3 beschriebene Vorgehen zur Erstellung von Hotellings T^2-Karte die gewöhnliche empirische Kovarianzmatrix verwendet, setzt STATISTICA nun den auf den *sukzessiven Differenzen* $\boldsymbol{D}_t := \boldsymbol{X}_t - \boldsymbol{X}_{t-1}$, $t = 2, \ldots, m$, beruhenden Schätzer

$$\mathbf{S}_{sd} \;:=\; \tfrac{1}{2(m-1)} \cdot \mathbf{D}^T\mathbf{D}, \qquad \mathbf{D}^T \;:=\; (\boldsymbol{D}_2, \ldots, \boldsymbol{D}_m),$$

ein (Montgomery, 2005, S. 502).

Die *verallgemeinerte Varianz* A_t, verfügbar nur für $n > 1$, berechnet sich als die Determinante der Kovarianzmatrizen $\mathbf{S}_t$ aus Hintergrund 12.1.6.2: $A_t := \det \mathbf{S}_t$. Erwartungswert und Varianz dieser Statistik ergeben sich laut Montgomery (2005), S. 511f, zu

$$E[A_t] \;=\; b_1 \cdot \det \boldsymbol{\Sigma} \qquad \text{und} \qquad V[A_t] \;=\; b_2 \cdot \det{}^2\boldsymbol{\Sigma},$$

wobei $\boldsymbol{\Sigma}$ die wahre Kovarianzmatrix ist. Die Faktoren b_1 und b_2 sind hierbei

gegeben durch die recht komplexen Ausdrücke

$$b_1 \;=\; \textstyle\prod_{i=1}^{p} (n-i) \;/\; (n-1)^p,$$
$$b_2 \;=\; \textstyle\prod_{i=1}^{p} (n-i) \cdot \left(\prod_{j=1}^{p} (n-j+2) - \prod_{j=1}^{p} (n-j) \right) \;/\; (n-1)^{2p}.$$

STATISTICA trägt nun die Statistiken A_t/b_1 auf eine Kontrollkarte mit 3-σ-Grenzen auf, wobei im Falle von Phase I die wahre Matrix $\mathbf{\Sigma}_0$ im Kontrollzustand durch $\mathbf{S}$ ersetzt wird.

Bei der MEWMA-Karte mit geg. Parameter λ werden zuerst die transformierten Vektoren

$$\boldsymbol{Z}_t \;=\; \lambda \cdot \bar{\boldsymbol{X}}_t \;+\; (1-\lambda) \cdot \boldsymbol{Z}_{t-1}$$

berechnet, vgl. Hintergrund 11.4.1.8 und 12.1.5.1, und davon anschließend die gewöhnliche Hotelling-Statistik. Hierbei muss berücksichtigt werden, dass die Kovarianzmatrix der Statistik $\boldsymbol{Z}_t$ gegeben ist zu

$$\mathbf{\Sigma}_{\boldsymbol{Z}_t} \;=\; \tfrac{\lambda}{2-\lambda} \cdot (1 - (1-\lambda)^{2t}) \cdot \tfrac{1}{n} \mathbf{\Sigma}.$$

Bei der MCUSUM-Karte für Einzelwerte ($n = 1$), die in der Literatur leider nicht einheitlich definiert ist, hat STATISTICA den Vorschlag $MC1$ von Pignatiello & Runger (1990)[1] implementiert. Betrachtet wird die kumulierte Summe

$$\boldsymbol{C}_t \;:=\; \textstyle\sum_{t-n_t+1}^{t} (\boldsymbol{X}_t - \boldsymbol{\mu}_0) \quad \text{bzw. deren Norm } \|\boldsymbol{C}_t\|^2 \;:=\; \boldsymbol{C}_t^T \mathbf{\Sigma}_0^{-1} \boldsymbol{C}_t,$$

aus der man die Statistik $MC1_t := \max\{\|\boldsymbol{C}_t\| - k \cdot n_t, 0\}$ berechnet, wobei n_t die Zahl an Beobachtungen seit dem letzten Nulldurchgang bezeichnet, d. h.

$$n_t \;=\; n_{t-1} + 1 \quad \text{für } MC1_{t-1} > 0, \qquad \text{und } 1 \quad \text{sonst.}$$

Für k wählt STATISTICA den Wert 0,5, und in Phase I wird statt $\mathbf{\Sigma}_0$ wieder $\mathbf{S}_{sd}$ verwendet. ◇

Die Berechnung all dieser Karten geht dabei jeweils ähnlich vonstatten: Nach Auswahl des Kartentyps bestimmt man im ersten Dialog die zu betrachtenden Variablen, gibt ggf. die Stichprobenvariable an bzw. legt den Stichprobenumfang fest, und drückt *OK*. Es wird dann eine Karte mit den standardmäßigen Voreinstellungen erstellt, wobei die Grenzen aus den Daten heraus geschätzt werden (Phase I, siehe Abschnitt 12.1.2). Das Menü kann aber erneut aktiviert und Änderungen an den Voreinstellungen vorgenommen werden. Auf der Karte *X(Multivariat)* kann man für Phase II die Werte für den Mittelwertsvektor und die Kovarianzmatrix auch vorgeben, ferner die Zahl m der Stichproben, auf denen diese Schätzung beruht, vgl. Hintergrund 12.1.6.2 und G.4.1.1.

[1] Pignatiello, J.J., Jr., Runger, G.C.: *Comparisons of multivariate CUSUM Charts.* Journal of Quality Technology 22(3), S. 173-186, 1990.

Besonders positiv zu bemerken ist, dass im Falle der MEWMA- und MCUSUM-Karte, bei denen das Design sinnvollerweise auf *ARL*-Betrachtungen beruhen sollte, siehe Abschnitt 12.1.5, die notwendigen *ARL*-Berechnungen fertig implementiert wurden: Der Benutzer gibt nach Klick auf *UCL* die gewünschte ARL_0 im Kontrollzustand vor, z. B. den in der Literatur beliebten Wert 200, STATISTICA berechnet daraufhin die passenden Kontrollgrenzen.

Zum Schluss sollte noch erwähnt werden, dass drei Kombinationen von Multiple-Stream-Karten implementiert wurden. Diese dienen der Überwachung gleichartiger, parallel ablaufender Prozesse. Aus Gründen der Übersicht werden dabei aber immer nur der jeweilige Maximal- und Minimalwert zur Zeit t aufgetragen.

G.4.2 Prozessfähigkeitsuntersuchungen

Mit Version 8.0 wurde bei STATISTICA ein neues Menü zur Prozessfähigkeitsanalyse gemäß ISO 21747 oder DIN 55319 geschaffen, erreichbar über *Statistik → Industrielle Statistik & Six Sigma → Prozessanalyse → Prozessfähigkeit ISO/DIN (Verteilungszeitmodelle)*. Bei diesem neuen Menü können die in der jeweiligen Norm festgelegten Verteilungszeitmodelle und Berechnungsmethoden, in dortiger Bezeichnungsweise, gewählt werden. Dabei muss sich der Benutzer zu Beginn für eine der Normen entscheiden, nämlich im Feld *Auswahl Norm*. Je nach Wahl wird eine leicht modifizierte Auswahl an Karten angeboten, auch passt sich die Bezeichnungsweise auf den Karten der jeweiligen Norm an. Wir wollen im Folgenden den Fall der Norm DIN 55319 besprechen. Für über diesen Abschnitt hinausgehende Informationen sei dabei auf Dietrich & Schulze (2003) verwiesen.

Hintergrund G.4.2.1

Gemäß Abschnitt 12.1 ist ein Prozess *unter Kontrolle*, wenn er *stationär* verläuft, d. h. wenn die Prozessverteilung (und mit ihr so wesentliche Charakteristika wie Erwartungswert und Varianz) über die Zeit hinweg konstant bleiben. Bei der Behandlung der Prozessfähigkeitsindizes C_p, C_{pk}, etc., in Abschnitt 12.1.8 hatten wir betont, dass eine Schätzung der Indizes auch nur bei solch kontrollierten Prozessen vorgenommen werden sollte. Nichtsdestotrotz verfolgt man in der Praxis einen etwas liberaleren/pragmatischeren Kurs und nimmt Prozessbeurteilungen auch bei, im statistischen Sinne, nicht beherrschten Prozessen vor. Obwohl dieses Vorgehen kritisch hinterfragt werden sollte, vgl. auch Hintergrund 12.1.8.1, wollen wir hier nüchtern die Sachlage beschreiben.

Die Norm DIN 55319, siehe auch Dietrich & Schulze (2003), unterscheidet acht verschiedene Prozessmodelle, die sog. *Verteilungszeitmodelle* A1, A2, B, C1 bis C4, und D, welche insgesamt ein so breites Spektrum an Prozessfiguren abdecken, dass sich ein realer Prozess in den meisten Fällen

näherungsweise einem dieser Modelle zuordnen lässt. Die Modelle werden dabei immer über fünf Charakteristika definiert: Lage, Streuung, Schiefe, Exzess und Momentan-/Randverteilung. Ferner wird die sog. 'resultierende Verteilung' betrachtet, also die Mischverteilung über die gesamte Zeit hinweg. Im Einzelnen machen diese Modelle folgende Voraussetzungen:

A1 und A2: Beide Modelle fordern, dass die fünf Charakteristika über die Zeit konstant bleiben, somit im Wesentlichen einen stationären Prozess. Dabei verlangt A1 Normalverteiltheit (was also insgesamt der Voraussetzung 12.1.3.1 entspricht), während A2 eine nichtnormale Randverteilung annimmt. Somit entspricht Modell A2 in etwa der Voraussetzung 12.1.2.1. Die resultierende Verteilung entspricht der Randverteilung des Prozesses.

B: Stellt die Forderungen aus A1, erlaubt jedoch, dass die Streuung zufällig variiert. Die resultierende Verteilung ist somit eine 'echte' Mischverteilung, mit nur einem Modus.

C1 bis C4: Stellen die Forderungen aus A1, erlauben jedoch, dass die Lage variiert: Bei C1 ist die Lage (Erwartungswert) eine normalverteilte Zufallsvariable, bei C2 ist sie zufällig, aber nicht normalverteilt. In beiden Fällen resultiert erneut eine unimodale Mischverteilung, im ersten Fall sogar eine Normalverteilung. Bei C3 wird eine rein systematische Änderung der Lage angenommen (substantielle Variation, z. B. Trend), bei C4 kann sich die Lage systematisch und zufällig ändern. Über die Mischverteilung lässt sich keine allgemeingültige Aussage treffen.

D: Der 'worst case', alle fünf Charakteristika können sich systematisch oder zufällig ändern.

Zu jedem Modell wird eine Auswahl der Berechnungsmethoden $M1_1$ bis $M1_4$, M2 bis M6 angeboten, für deren ausführliche Beschreibung auf die STATISTICA-Hilfe, Dietrich & Schulze (2003) oder die DIN-Norm selbst verwiesen sei. Erwähnt sei an dieser Stelle, dass das in Abschnitt 12.1.8 empfohlene Schätzverfahren über $\bar{\bar{X}}$ und $\bar{S}$ der Methode $M1_2$ entspricht, anwendbar nur für den Prozesstyp A1. Falls nur die Normalverteilungsannahme verletzt ist (Modell A2), so empfahl Abschnitt 12.1.8 die Verwendung quantilbasierter Schätzer. Dies entspricht der Methode M4. Diese ist laut DIN, zusammen mit der Methode M3, welche die Prozessstreuung aus der Spannweite schätzt, als universell anwendbar empfohlen. Der Leser hinterfrage jedoch kritisch, inwiefern eine Schätzung für einen nichtstationären Prozess (insbesondere vom Schlage C oder D) überhaupt sinnfällig ist. ◇

Bemerkung zu Hintergrund G.4.2.1.

Falls der Leser an einer Prozessfähigkeitsanalyse gemäß ISO 21747 interessiert ist, so sind nur geringe Modifikationen gegenüber DIN 55319 zu berücksichtigen. Beide Normen werden in der STATISTICA-Hilfe, Karte *Inhalt*, miteinander verglichen, unter dem Pfad *Statistik* → *Analysen* → *Industrielle*

Statistiken & Six Sigma → Prozessanalyse → Prozessanalyse – Überblick → Prozessfähigkeit ISO/DIN (Verteilungszeitmodell). Der Eintrag *Modelle* bietet einen tabellarischen Überblick über die Charakteristika der Verteilungszeitmodelle A1 bis D in beiden Normen, der Eintrag *Berechnungsmethoden ...* zeigt Entsprechungen bei den Berechnungsmethoden M1 bis M6 (DIN) bzw. M1 bis M4 (ISO) auf. □

Bei geg. Prozessdaten gilt es nun, zuerst das passende Verteilungszeitmodell zu identifizieren. Anschließend wird, basierend auf diesem Modell bzw. dessen resultierender Mischverteilung, eine angemessene Berechnungsmethode ausgewählt, mit deren Hilfe man die Prozessfähigkeit schätzen kann. Für den erstgenannten Schritt führt STATISTICA eine Reihe von Tests aus, um die genannten Voraussetzungen zu prüfen. Durch Kombination der Resultate versucht STATISTICA eines der Modelle eindeutig zu identifizieren, siehe auch Kapitel 4 bei Dietrich & Schulze (2003). Um etwa eine systematische Änderung festzustellen, wird versucht, ein lineares Regressionsmodell anzupassen.

Durchführung G.4.2.2

Um eine Prozessfähigkeitsanalyse gemäß der Empfehlungen der gewählten Norm durchzuführen, wechselt man zuerst auf die Karte *Gruppierung*. Dort muss die Datenvariable sowie die Stichprobengröße, bzw. die Variable mit den Stichprobencodes, ausgewählt werden. Bei *Toleranzen* müssen ferner der *Sollwert* sowie die Spezifikationsgrenzen LSL und USL bestimmt werden. Die Berechnungen starten dann nach einem Klick auf *OK*. Zuvor kann der Leser jedoch auf den zahlreich vorhandenen Karten eine Reihe von Einstellungen treffen, die etwa die durchzuführenden Testverfahren festlegen oder die exakten Schätzverfahren zu den einzelnen Berechnungsmethoden.

Im anschließenden Dialog sieht man auf der Karte *Methoden*, welches Verteilungszeitmodell als zutreffend erkannt wurde. Entsprechend kann man zwischen bestimmten Berechnungsmethoden wählen. Auf der Karte *Standard* kann man die Schätzwerte für die Indizes und die durchgeführten Tests einsehen. •

Nach einem Klick auf *Zus.fsg.* wird eine Tabelle mit den geschätzten Indizes ausgegeben, ferner eine Grafik mit $\bar{X}$- und R-Karte, einem Normalverteilungsplot, einem Histogramm, und einem Feld mit den geschätzten Indizes und einigen deskriptiven Statistiken. Über *Deskriptive Stat. & Tests* auf der Karte *Standard* lassen sich eine Reihe von Tabellen ausgeben, von denen besonders die Tabellen *Deskriptive Testergebnisse* und *Normalverteilung (Momentanverteilung)* interessant sind. Diese geben nämlich die Ergebnisse aller durchgeführten Testverfahren wieder, lassen somit erkennen, wie STATISTICA zu seiner Entscheidung gelangt

ist. Bei letztgenannter Tabelle, letzte Zeile, lässt sich der im Tipp kritisch erwähnte Test auf normale Randverteilung nachvollziehen.

Vorsicht ist bzgl. der Untersuchung auf normale Randverteilung geboten: STATISTICA erwartet, dass die Daten in Stichproben der Größe ≥ 3 gegeben sind. Die Daten werden nun innerhalb der einzelnen Stichproben studentisiert, und diese studentisierten Daten werden dem gewählten Normalverteilungstest unterzogen. *Aber:* Falls der Prozess tatsächlich normalverteilt ist, so sind diese studentisierten Variablen t-verteilt! Somit ist eine übermäßig häufige, fälschliche Ablehnung der Normalverteiltheit zu erwarten. Evtl. sollte der Leser also bei Durchführung G.4.2.2 auf der Karte *Normalverteilung* die Grenze für den kritischen p-Wert auf einen Wert kleiner als die vorgegebenen 0,05 absenken, z. B. 0,001.

Literaturverzeichnis

Dietrich, E., Schulze, A.: *Statistische Verfahren zur Maschinen- und Prozessqualifikation*. 4. Auflage, Carl Hanser Verlag, 2003.

Montgomery, D.C.: *Introduction to statistical quality control*. 5^{th} edition, John Wiley & Sons, Inc., 2005.

H Wichtige Neuerungen bei Version 9

Ähnlich wie bei der Version 8 fallen die mit Version 9 eingeführten Neuerungen (mit einer Ausnahme) nicht gleich ins Auge, so etwa die dank neuer Compiler verbesserte Performance und die 64-Bit-Unterstützung. Somit sind auch hier keine großen Umgewöhnungen nötig. Tatsächlich wurden aber doch eine Reihe von Änderungen vorgenommen, welche im Folgenden vorgestellt werden. Und bei der einzigen augenscheinlichen Neuerung handelt es sich um die seit Office 2007 bekannten Multifunktionsleisten („Ribbons“). Diese kann man seit Version 9 im Menü *Ansicht* auswählen. Um von diesen Multifunktionsleisten wieder zurück zur klassischen Menüstruktur zu gelangen, wählt man entweder ganz oben links den Punkt *Menüs* oder deaktiviert die Multifunktionsleisten ganz oben rechts unter *Optionen*.

H.1 Datenverwaltung

Seit Version 9 kann man eine Bilddatei punktweise in eine Tabelle importieren. Wählt man über *Datei* → *Bild importieren* eine Rastergrafik mit $m \times n$ Bildpunkten aus, so wird eine $m \times n$-Tabelle angelegt, in deren Zellen dann die numerischen RGB-Farbwerte (optional auch nur in *Graustufen*) stehen.

Möchte man die komplette Bildinformation (zeilenweise aneinandergehängt) in einem einzigen Fall ablegen, etwa um mehrere (kleine) Bilder einer Klassifikation zu unterziehen, kann man die beim eben beschriebenen Import erzeugte Tabelle auf folgende Weise transformieren:

Zunächst hängt man eine Spalte mit Laufindex (z. B. über die Formel „=v0“) an die Tabelle an. Dann öffnet man das Menü *Daten* → *Unstacking/Stacking*, Karte *Unstacking*, und wählt bei *Variablen* die eben neu erzeugte Variable als *Code-Variable* aus, und die restlichen als *Unstack-Variablen*. Anschließend bestätigt man mit *OK*, muss zuvor ggf. noch die *Maximale Anzahl der Variablen ...* anpassen.

Ferner gab es eine kleine Umbenennung, die den in Abschnitt 2.2 besprochenen Import von ActiveX-Dokumenten in Arbeitsmappen betrifft:

Statt *ActiveX-Dokumentobjekt* heißt es nun im betreffenden PopUp-Menü schlicht *Aus anderer Anwendung*.

H.2 Datenanalyse

Bei den in Abschnitt 5.4 besprochenen Histogrammen konnte man bis einschließlich Version 8 maximal 255 Kategorien anlegen. Da dies bei extrem großen Datensätzen oft nicht ausreichend war, kann man seit Version 9 immerhin maximal 999 Kategorien anlegen.

Eine Neuerung betrifft auch die in Abschnitt 5.6 besprochenen, weiteren grafischen Darstellungen. Seit Version 9 kann man nämlich ein Paretodiagramm (allerdings ohne empirische Verteilung) auch über den Umweg eines normalen Histogramms erzeugen. Hat man ein Histogramm erstellt, vgl. Abschnitt 5.4, so aktiviert man anschließend die Grafikoptionen, vgl. Abschnitt 4.1, und wechselt zur Rubrik *Plot: Histogramm*. Hier findet sich im Feld *Typ* die Option *Paretodiagramm*, nach deren Aktivierung und Klick auf *OK* das Histogramm in ein Paretodiagramm umgewandelt wird. An gleicher Stelle des Optionendialogs befände sich übrigens auch die analoge Option *Kreisdiagramm*.

Seit Version 8 bietet STATISTICA in der *Deskriptiven Statistik*, vgl. Durchführung 5.1.1, auf den Karten *Standard* und *Details* die Schaltfläche *Grafiken* an bzw. seit Version 9 sogar die zwei Schaltflächen *Grafiken 1* und *Grafiken 2*. Nach deren Betätigung wird für alle ausgewählten Variablen jeweils eine zusammengesetzte Grafik erzeugt, in der sich neben einer Auswahl der in Abschnitt 5.1 besprochenen Kenngrößen auch verschiedene Grafiken (Histogramm, Boxplot, etc.) befinden. Ebenfalls seit Version 9 wird auf der Karte *Standard* noch der Schalter *Zusf.: Vergleichsgrafiken* angeboten, bei dem eine zusammengesetzte Grafik für alle gewählten Variablen zugleich ausgegeben wird.

Bei den gemäß Abschnitt 6.1 erzeugten Korrelationsmatrizen werden seit Version 8 automatisch auch die Mittelwerte und Standardabweichungen mit ausgegeben; um dies zu unterdrücken, deaktiviert man auf der Karte *Optionen* die Option *Mw. und Std.abw. ...* Schließlich kann man sich seit Version 9 die Korrelationsmatrix auch noch farbig hinterlegen lassen (nützlich vor allem bei sehr vielen Variablen), indem man diese auf der Karte *Farbgebung* über *Matrix mit farbigen Zellen erstellen* erzeugt. In der Titelleiste der ausgegebenen Tabelle wird dann das Farbschema angezeigt; markiert man dieses mit der Maus, so erkennt man die zugeordneten Werte.

Teilweise seit Version 8 und stets seit Version 9 kann man sich den

Wechsel ins Menü *Poweranalyse* sparen, wenn man an Konfidenzintervallen für die verschiedenen t-Tests interessiert ist, vgl. Kapitel 9. Denn seither findet man auf der Karte *Optionen* die Option *Konfidenzintervall* bzw. *Konfidenzgrenzen*. Nach deren Aktivierung wird dann ein Konfidenzintervall für den Erwartungswert (Einstichproben-t-Test) bzw. die Differenz der Erwartungswerte (sonst) mit ausgegeben.

Zu den größten Neuerungen im Bereich Datenanalyse zählen die zwei neuen Module: Im *Data-Mining*-Menü gibt es nun ein Modul zur allgemeinen Optimierung von Funktionen, und im *Statistik*-Menü eines zur Verteilungsanalyse und Simulation von Daten. Auf Letzteres werden wir ausführlich in Abschnitt H.4 eingehen.

H.3 Analyseresultate

Hat man mit STATISTICA eine Grafik erzeugt, so kann man, wie in Abschnitt 4.1 beschrieben, den zu dieser Grafik gehörigen Optionendialog aufrufen und dort die Grafik den individuellen Bedürfnissen anpassen. Mit Version 9 wurde dieser Dialog neu gestaltet, die einzelnen Kategorien sind nun nicht mehr auf Karten organisiert, sondern in eine Baumstruktur integriert, siehe Abbildung H.1. Auch wenn wir nur gezielt auf ein Element der Grafik klicken, erscheint nun, im Gegensatz zu früheren Versionen, der Gesamtdialog, aber nur der für das angeklickte Element relevante Zweig ist geöffnet. Erscheint ein reduzierter Dialog, obwohl wir doch lieber den vollständigen Dialog gewünscht hätten, so kann man auf die '+'-Zeichen in der Baumstruktur klicken.

Das gleiche Design, wie man es beim Grafikoptionendialog findet, wurde mit Version 9 übrigens auch auf den in Abschnitt 3.5 behandelten Dialog der Programmoptionen angewandt.

Eine weitere, äußerst nützliche Funktion im Hinblick auf die Anpassung von Grafiken wurde mit Version 9 eingeführt, nämlich die in Abbildung H.1 ganz unten sichtbare Option *Makro*, vgl. Kapitel 13. Aktiviert man diese Option und führt dann eine Anpassung der Grafik durch, so wird nach finalem Klick auf *OK* ein Makro erstellt, siehe auch Abschnitt 13.2, das alle gemachten Änderungen protokolliert. Erzeugt man nun eine weitere Grafik gleichen Typs (also z. B. eine Balkengrafik) und führt das Makro aus, siehe Abschnitt 13.2, wenn diese Grafik in der Arbeitsmappe zuoberst ist, dann werden alle protokollierten Anpassungen auf diese Grafik angewendet.

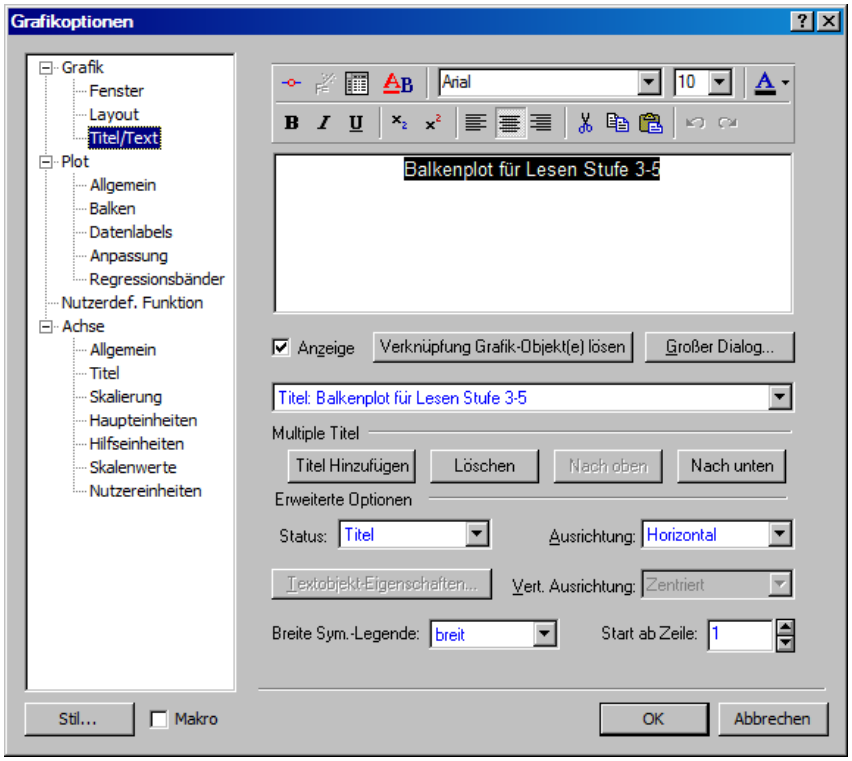

Abb. H.1: *Alle Optionen einer Grafik, Dialog seit Version 9.*

H.4 Verteilungsanalyse

Seit Version 9 verfügt STATISTICA über das Menü *Statistik* → *Verteilungen & Simulationen*, über welches man auf viele der in Kapitel 8 besprochenen Verfahren zur Verteilungsanalyse für eine Vielzahl von Verteilungen zentral zugreifen kann. Ferner bietet dieses Menü eine Reihe neuer Funktionen, von denen wir einige im Folgenden behandeln wollen.

Ist man an der Schätzung von Verteilungsparametern interessiert, siehe Abschnitt 8.1, kann man seit Version 9 auch das Menü *Verteilungen & Simulationen* → *Verteilung anpassen* wählen. Dort gibt man unter *Variablen* Selbige an, unterschieden nach *stetig* oder *diskret*. Nach *OK* wechselt man auf die Karte *Anpassung speichern* und kann dort bereits die Parameterschätzwerte einsehen. Über *Zusf.: Verteilungsstatistiken* kann man sich diese Schätzwerte auch in einer Tabelle ausgeben lassen. Die Verteilungen sind dabei nach ihrer Anpassungsgüte geordnet.

Ferner wird seit Version 9 ein neues grafisches Werkzeug zur Verteilungsanalyse angeboten, siehe auch Abschnitt 8.2, nämlich die nun zu besprechende *empirische Verteilungsfunktion*.

Hintergrund H.4.1

Im Gegensatz zum Histogramm ($\approx$ empirische Dichte) ist für die empirische Verteilungsfunktion keine Kategorisierung der Daten nötig, stattdessen ist sie definiert als

$$\hat{F}_n(x) \;:=\; \tfrac{1}{n} \cdot \big|\{1 \le i \le n \mid X_i \le x\}\big|, \tag{H.1}$$

d. h. $\hat{F}_n(x)$ ist $1/n$ mal die Anzahl der Datenwerte $\le x$. ◇

Im Rahmen der Verteilungsanalyse ist $\hat{F}_n(x)$ zu vergleichen mit einer hypothetischen Verteilungsfunktion $F_0(x)$. Dieser Vergleich wird in STATISTICA durch Ausgabe des zu $\hat{F}_n(x)$ gehörenden 95 %-Konfidenzintervalls erleichtert, d. h. $F_0(x)$ sollte weitestgehend innerhalb dieses Konfidenzbereichs verlaufen.

Durchführung H.4.2

Im Menü *Verteilungen & Simulationen* → *Verteilung anpassen* wählen wir die zu analysierenden *Variablen* aus, wobei nach *stetig* oder *diskret* zu unterscheiden ist. Nach *OK* wählen wir auf der Karte *Standard* die einzelnen *Variablen* und *Verteilungen* aus (genau genommen nur die Verteilungsfamilie, es werden die geschätzten Parameter verwendet) und klicken dann auf *Empirische Verteilungsfunktion*, um Selbige zu erstellen. •

Ein Beispiel ist in Abbildung H.2 zu sehen, wo von den in Abschnitt 8.4.3 untersuchten Frauen-Daten eine empirische Verteilungsfunktion erzeugt wurde. Die ebenfalls gezeichnete Verteilungsfunktion der Normalverteilung verläuft nahe der empirischen Verteilungsfunktion, zwischen den Konfidenzbändern.

Der in Abschnitt 8.3 behandelte χ^2-*Anpassungstest* wird seit Version 9 auch im Menü *Verteilungen & Simulationen* angeboten, siehe Durchführung H.4.3. Dort findet man auch den *Kolmogorov-Smirnov-Test* sowie den *Anderson-Darling-Test*, und vor allem viele weitere Verteilungsfamilien, die im ursprünglichen Menü nicht vorhanden sind.

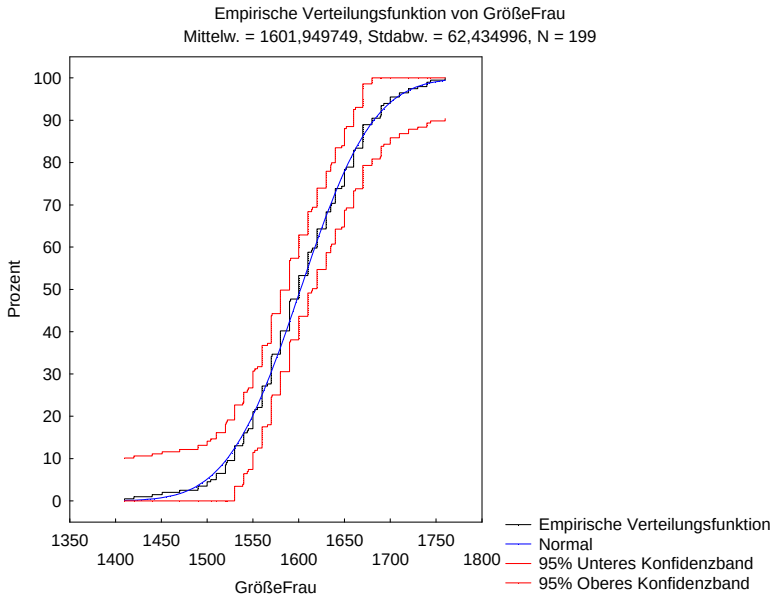

Abb. H.2: *Empirische Verteilungsfunktion der Frauen-Daten.*

Durchführung H.4.3

Seit Version 9 kann man zur Verteilungsanalyse auch das Menü *Verteilungen & Simulationen* → *Verteilung anpassen* wählen und gibt dort unter *Variablen* Selbige an, unterschieden nach *stetig* oder *diskret*. Man beachte den Plural: Bei mehreren ausgewählten Variablen werden alle zugleich berücksichtigt und auch die wechselseitigen Korrelationen geschätzt.

Nach *OK* kann man auf der Karte *Standard* alle *Variablen* und *Verteilungen* über die entsprechenden Schaltflächen einzeln durchlaufen und via *Zusf.: Verteilungsstatistiken* die jeweiligen Ergebnisse aller relevanten Verfahren tabellarisch ausgeben lassen. Für einen Gesamtüberblick wechselt man auf die Karte *Anpassung speichern* und kann dort pro Variable über *Zusf.: Verteilungsstatistiken* eine Tabelle mit allen Verfahren für alle Verteilungen erzeugen. •

Auf der Karte *Standard* werden übrigens auch alle anderen in Kapitel 8 besprochenen Verfahren der Verteilungsanalyse angeboten (wenn auch teilweise nicht alle dortigen Verteilungsfamilien), so dass man ab Ver-

sion 9 eine solche Verteilungsanalyse in vielen Fällen vollständig über *Verteilungen & Simulationen* durchführen kann.

Seit Version 9 bietet STATISTICA noch ein weiteres, wesentlich mächtigeres Werkzeug zur Simulation von Zufallszahlen an: das Menü *Statistik* → *Verteilungen & Simulationen*. Idee ist hierbei, dass man eingangs einen Datensatz vorliegen hat, den man wie in Durchführung H.4.3 besprochen einer Verteilungsanalyse unterzieht. Von der auf der Karte *Anpassung speichern* vorgeschlagenen Verteilung mit bester Anpassung kann man nun weitere Zufallszahlen erzeugen, indem man *Simulation ausführen* wählt, im nächsten Dialog ein Verfahren sowie die Anzahl zu erzeugender Zufallszahlen auswählt und schließlich auf *Simulation* klickt. Je nach gewähltem Verfahren kann man dabei auch die Korrelation zwischen den Variablen berücksichtigen.

Auf der Karte *Anpassung speichern* bestände auch die Möglichkeit, die empirische Korrelationsmatrix (samt Mittelwerten und Standardabweichungen) via *Verteilungsanpassung speichern* abzuspeichern (als `.smx`-Datei). Für den Anwender nicht unmittelbar[1] sichtbar werden in dieser Matrix Informationen zu den angepassten Verteilungen hinterlegt. Öffnet man diese Datei zu einem späteren Zeitpunkt im Menü *Statistik* → *Verteilungen & Simulationen* → *Simulation ausführen*, so kann man, wie eben beschrieben, Zufallszahlen/-vektoren der zugehörigen Verteilung[2] erzeugen.

[1] Diese Informationen kann man im Dialog der Variablenspezifikationen (vgl. Abschnitt 3.1.2) einsehen und bearbeiten; man klickt dort auf die Schaltfläche *Eigenschaften*.

[2] Wie in voriger Fußnote beschrieben, kann man die Verteilungen und Korrelationen ggf. anpassen.