

Abb. 1.1: Der Startbildschirm von STATISTICA.

# 1 Erste Schritte in STATISTICA

Das Programm STATISTICA umfasst eine sehr große Sammlung statistischer Verfahren und grafischer Methoden, ist aber trotzdem ein sehr überschaubares und leicht bedienbares Programm geblieben. Dazu trägt auch die benutzerfreundliche Oberfläche bei, die den meisten Computeranwendern ohnehin durch EXCEL oder ähnliche Tabellenkalkulationsprogramme vertraut erscheinen dürfte.

Wie in Abbildung 1.1 angedeutet, verfügt das Tabellenblatt von STATISTICA über ähnliche Funktionen wie das von EXCEL. So kann man beispielsweise auch hier eine Reihe von Zahlen mittels Mauszeiger fortsetzen.

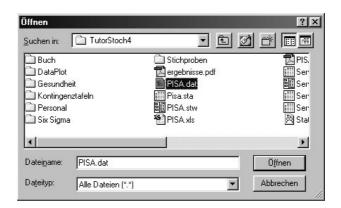


Abb. 1.2: Öffnen der Datendatei.

Abweichend von EXCEL ist allerdings die Funktion von Zeilen und Spalten. Während bei EXCEL Zeilen und Spalten prinzipiell gleichwertig sind, sind in STATISTICA, wie bei einer Datenbank, die Spalten den Zeilen übergeordnet. Jedoch erlaubt STATISTICA, bei Bedarf ein Tabellenblatt zu transponieren.

Die Spalten, hier *Variablen* genannt, bezeichnen verschiedene Merkmale, in den Zeilen, den *Fällen*, werden die einzelnen konkreten Messwerte gesammelt. Ein Beispiel:

Beim innerdeutschen Vergleich der PISA-Studie (OECD PISA 2000, zusammengefasst vom Berliner Max-Planck-Institut für Bildungsforschung 2002) wurde u. a. die Fähigkeit im Lesen, in der Mathematik und den Naturwissenschaften von Schülern der Jahrgangsstufe 9 in 14 deutschen Bundesländern (ohne Hamburg und Berlin) gemessen. In diesem Fall wären Lesefähigkeit, Math. Fähigkeit und Naturwiss. Fähigkeit die drei Variablen, die Fälle würden mit den einzelnen Bundesländern bezeichnet und die jeweils erzielten Werte enthalten.

Nehmen wir dieses Beispiel, um uns ein wenig mit STATISTICA vertraut zu machen:

Die konkreten Daten befinden sich in der (Text-)Datei PISA.dat und müssen erst in ein STATISTICA-Tabellenblatt importiert werden. Dazu wählt man wie gewohnt den Menüpunkt  $Datei \rightarrow \ddot{O}ffnen$ , worauf ein Dialogfenster wie in Abbildung 1.2 erscheint. Nachdem man bei Dateityp Alle Dateien eingestellt hat, wird die gesuchte Datei angezeigt und kann ausgewählt werden.

<sup>&</sup>lt;sup>1</sup>Es sei an dieser Stelle nochmals erwähnt, dass alle in diesem Text besprochenen Datensätze von der im Vorwort angegebenen Seite heruntergeladen werden können.





Abb. 1.3: Die Datendatei als Tabelle importieren und Dateityp Auto wählen.

Im nächsten Dialog, vergleiche Abbildung 1.3, wählt man Als Tabelle importieren und belässt es anschließend beim Dateityp Auto.

Nun öffnet sich das Fenster Textdatei öffnen aus Abbildung 1.4. Hier müssen einige Einstellungen gemacht werden. Ein Blick in die Datei PISA.dat zeigt, dass sich über dem eigentlichen Datensatz ein paar Zeilen mit einer kleinen Beschreibung befinden. Da der beschreibende Text nicht im Tabellenblatt erscheinen soll, beginnen wir den Import bei Zeile 4, welche die Bezeichnungen der späteren Variablen enthält. Wir machen einen Haken bei Variablennamen aus erster Zeile. Ebenso verfahren wir bei Fallnamen aus erster Spalte, damit die Fälle die Namen der einzelnen Bundesländer tragen. Schließlich darf bei Feldtrennung(en) lediglich Tabulator markiert sein. Eine Auswahl von Komma würde dagegen dazu führen, dass die Kommazahlen in mehrere Spalten zerlegt würden.

Falls der Computer auf deutsche Sprache eingestellt ist, verwendet STATISTICA die deutsche Schreibweise, insbesondere also das Komma als Dezimaltrennzeichen. In der Praxis hat man jedoch häufig Datensätze aus dem angloamerikanischen Raum vorliegen, die Punkte verwenden. Damit solche Daten importiert werden können, müssen bei den Versionen bis inkl. 6.1 vorher in einem Editor die Punkte durch Kommata ersetzt werden. Alternativ, aber wohl problematischer, kann in der Systemsteuerung unter Ländereinstellungen die Sprache geändert werden. Seit Version 7 werden auch Punkte als Dezimaltrennzeichen erkannt.



Neu ab Version 7 sind leicht modifizierte Dialoge zum Import von Dateien, verglichen mit denen aus den Abbildungen 1.3 und 1.4. Nützlich sind dabei insbesondere einige weitere Optionen im Dialog Textdatei öffnen aus Abbildung 1.4, wie etwa Voranst. Leerzeichen abschneiden oder Leere Zeilen ignorieren. Ferner erlaubt ein Häkchen bei Ansicht/Bearbeiten von Spaltentypen vor Import eine Vorschau auf die von STATISTICA erstellten Variablen.





Abb. 1.4: Unter Textdatei öffnen müssen eine Reihe wichtiger Einstellungen gemacht werden.

Nun ist es geschafft, und durch Klick in das weiße Feld links oben in der Tabelle kann diese markiert werden. Durch erneuten Klick mit der rechten Maustaste ins Tabellenblatt öffnet sich ein PopUp-Menü. Dort kann unter Format der Befehl Optimale Höhe/Breite betätigt werden, damit die Daten und Randbeschriftungen tatsächlich gut lesbar sind, vergleiche hierzu Abbildung 1.5.

Betrachten wir die Daten genauer: In den Disziplinen Lesen, Mathematik und Naturwissenschaften wurde jeweils der Anteil von Schülern an den Stufen 3 bis 5 gemessen, konkret: Wenn wir unter der Rubrik Lesen bei Baden-Württemberg eine 63,4 finden, so bedeutet dies, dass 63,4 % aller dortigen Schüler der 9. Klasse über mittlere bis sehr gute Lesekenntnisse verfügen. Wir wollen uns die Werte der Variablen Lesen grafisch veranschaulichen, und zwar mit Hilfe eines Balkendiagramms. Dazu wählen wir wie in Abbildung 1.6 den Menüpunkt  $Grafik \rightarrow 2D$ - $Grafiken \rightarrow Balkenplots$ . Anschließend stellen wir wie in Abbildung 1.7  $Lesen\ Stufe\ 3-5$  als Variable und horizontale Ausrichtung ein, und betätigen dann den OK-Knopf. Nun wird eine  $Arbeitsmappe\ (Workbook)$  erstellt und die gewünschte Grafik in Selbige eingefügt. Das Ergebnis ist in Abbildung 1.8 zu sehen.

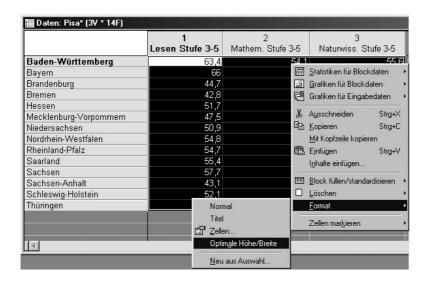


Abb. 1.5: Die gewünschten Daten wurden nun als Tabellenblatt importiert.

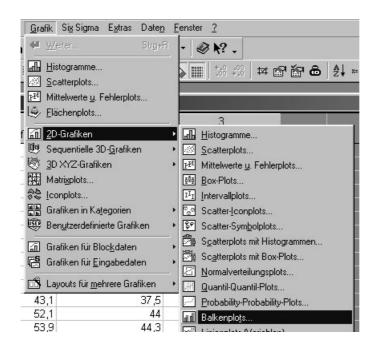


Abb. 1.6: Wir erstellen ein Balkendiagramm der Daten.

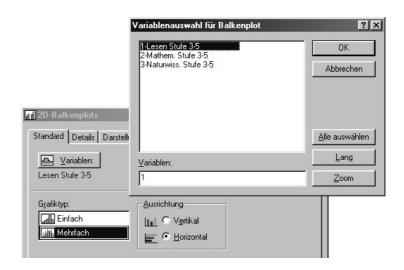


Abb. 1.7: Als Variable wählen wir Lesen Stufe 3-5.

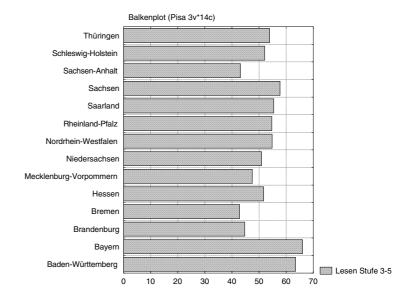


Abb. 1.8: Ein Balkendiagramm der Lesedaten.



Abb. 1.9: Aktives Grafikmenü – Reaktivierung per Knopfdruck.

Links unten im STATISTICA-Fenster befindet sich nun ein Knopf, mit dem das Balkendiagramm-Menü mit den getroffenen Einstellungen wieder aktiviert werden kann, siehe Abbildung 1.9. Nach erneuter Aktivierung können weitere Diagramme, etwa alle drei Disziplinen in einem mehrfachen Balkendiagramm, erstellt werden können. Dies sei dem Leser zur Übung empfohlen. Parallel dazu können auch beliebige weitere Analysen gestartet werden, auch hier sei an die Experimentierfreude des Lesers appelliert.

Bis zu dem Zeitpunkt, an dem wir die erstellte Arbeitsmappe zum ersten Mal schließen, werden alle weiteren Analyseresultate, auch wenn sie von anderen Datenblättern herrühren sollten, in dieser Arbeitsmappe abgelegt. Man erkennt dies zum Beispiel im Hintergrund von Abbildung 2.2 auf Seite 15. Beim Abspeichern der Arbeitsmappe bekommt diese die Dateiendung .stw.

Setzt man zu einem späteren Zeitpunkt Analysen fort, die man gern in einer früheren Arbeitsmappe ablegen möchte, gibt es zwei Möglichkeiten: Entweder man analysiert wie gewohnt, öffnet am Ende die gewünschte Arbeitsmappe und zieht mit gedrückter Maustaste die Resultate in die Zielarbeitsmappe. Oder man wählt vor Beginn der Analysen den Punkt  $Datei \rightarrow Ausgabemanager$ , woraufhin der Optionen-Dialog geöffnet wird, mit der Karte Ausgabemanager in Front. Dort wählt man Alle Ergebnisse  $(\dots)$   $in \rightarrow Arbeitsmappe$   $\rightarrow$  Bestehende Arbeitsmappe, und in letzterem Feld die gewünschte Zielarbeitsmappe. Nach Bestätigung mit OK werden alle weiteren Analyseresultate dort eingefügt.

**Achtung!** STATISTICA merkt sich diese Einstellung. Solange Sie daran nichts ändern, werden von nun an immer Resultate in genau dieser Arbeitsmappe abgelegt.

Neu ab Version 7 ist die Möglichkeit, sog. Variablenbündel zu definieren, und zwar immer vom Variablendialog einer Analyse aus. Bei aktuellen STATISTICA-Versionen verfügt dieser nämlich, im Gegensatz zu Abbildung 1.7, noch über den zusätzlichen Knopf [Bündel]..., nach dessen Betätigung sich der Variablenbündel-Manager öffnet. Hier kann man bestehende Bündel verändern, oder durch Klick auf Neu... ein neues definieren. Im letztgenannten Fall muss man zuerst einen Namen





für das Bündel angeben, anschließend erscheint ein Variablenauswahldialog wie in Abbildung 1.7. Hier wählt man die Variablen für das Bündel aus und bestätigt.

Zukünftig werden, wenn die Tabelle im .sta-Format ab Version 7 gespeichert wird, im Variablendialog neben den Variablen selbst auch die definierten Bündel angezeigt. Klickt man auf ein solches, so werden auf einen Schlag genau jene Variablen ausgewählt, die im Bündel vermerkt sind. Gerade bei Dateien mit vielen Variablen, bei denen immer wieder die gleiche Auswahl einer Analyse unterzogen wird, ist dies eine große Erleichterung.

# 9.7 Kruskal-Wallis- und Friedman-Test

Die im vorigen Abschnitt 9.6 betrachtete ANOVA hat den Nachteil, dass sie an die vorliegenden Daten die Forderung nach Normalverteilung stellt. In vielen Fällen ist diese Forderung jedoch nicht einmal näherungsweise erfüllt. Sind dann auch noch die Stichprobenumfänge klein, so müssen wir uns in einem solchen Fall nach einer Alternative umsehen.

Eine nichtparametrische Alternative ist der Kruskal-Wallis-Test. Dieser betrachtet nicht die konkreten Realisierungen  $x_{i,j}$  selbst, sondern nur ihre jeweiligen  $R\ddot{a}nge\ r_{i,j}$ .



#### Hintergrund 9.7.1

Seien  $y_1, \ldots, y_N$  die realisierten Werte einer Stichprobe, und bezeichne  $y_{(i,N)}$  die *i*-größte Realisation, vgl. Anhang A.1.1 zu Ordnungsstatistiken. Nehmen alle  $y_i$  verschiedene Werte an, d. h. gilt

$$y_{(1,N)} < y_{(2,N)} < \dots < y_{(i,N)} < \dots < y_{(N,N)}$$

so ist der Rang von  $y_{(i,N)}$  gerade i.

Gibt es dagegen übereinstimmende Werte im Datensatz, und sind die Daten angeordnet gemäß

$$y_{(1,N)} \leq \ldots < y_{(i,N)} = \ldots = y_{(k,N)} < \ldots \leq y_{(N,N)},$$

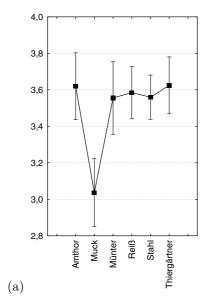
so ordnet man einem jedem der  $y_{(i,N)},\dots,y_{(k,N)}$  den  $\textit{mittleren Rang}\ \frac{i+k}{2}$ zu.  $\diamondsuit$ 



#### Voraussetzung 9.7.2

Die Zufallsvariablen  $X_{i,j}, j=1,\ldots,n_i$ , der *i*-ten Stichprobe besitzen die gleiche *stetige* Verteilung,  $i=1,\ldots,I$ . Es sei  $N:=n_1+\ldots+n_I$ . Ferner sind alle Zufallsvariablen unabhängig voneinander.

Es wird nun die Hypothese  $H_0$  untersucht, dass sogar *alle* Zufallsvariablen ein und derselben Verteilung unterliegen, womit insbesondere auch ihre Erwartungswerte übereinstimmen würden.



	Bonferroni T Homogene ( Fehler: MQ(	Gruppen, A	Alpha	= ,0	
	Prüfer	Urteil	1	2	
Zelle Nr.		Mittel			
2	Muck	3,036429		****	
3	Münter	3,555333	****		
2 3 5	Stahl	3,559800	****		
4	Reiß	3,584667	****		
1	Amthor	3,619714	****		
6	Thiergärtner	3,625000	****		

	Simultane	ni Test; Vari e Konfidenz IQ(Zwischei	intervalle	` ,	2,00	
	versus	Mittel	Standard	р	-95,00%	+95,00%
Zelle Nr.	Zelle Nr	Differ.	Fehler		LSD KG	LSD KG
1	2	0,583286	0,091792	0,000000	0,310912	0,855660
	3	0,064381	0,090076	1,000000	-0,202902	0,331664
	4	0,035048	0,081593	1,000000	-0,207063	0,277158
	5	0,059914	0,079788	1,000000	-0,176841	0,296669
	6	-0,005286	0,083794	1,000000	-0,253928	0,243357
2	3	-0,518905	0,095131	0,000002	-0,801187	-0,236623
	4	-0,548238	0,087141	0,000000	-0,806812	-0,289664
	5	-0,523371	0,085454	0,000000	-0,776938	-0,269805
	6	-0,588571	0,089206	0,000000	-0,853272	-0,323871
3	4	-0,029333	0,085332	1,000000	-0,282538	0,223872
	5	-0,004467	0,083608	1,000000	-0,252556	0,243623
	6	-0,069667	0,087439	1,000000	-0,329125	0,189791
4	5	0,024867	0,074390	1,000000	-0,195872	0,245606
	6	-0,040333	0,078672	1,000000	-0,273777	0,193110
5	6	-0,065200	0,076799	1,000000	-0,293085	0,162685

(b)

 ${\bf Abb.~9.11:~} Mittelwert vergleiche~der~Weinpr\"{u}ferdaten~aus~Beispiel~9.6.8.$ 



#### Hintergrund 9.7.3

Bezeichnet  $R_{i,j}$  den zufälligen (mittleren) Rang von  $X_{i,j}$  in der Gesamtstichprobe, so definieren wir in Analogie zu Hintergrund 9.6.2 das Rang-Gesamtmittel  $\bar{R}_{\bullet \bullet} = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{n_i} R_{i,j} = \frac{N+1}{2}$  und das *i*-te Rang-Gruppenmittel  $\bar{R}_{i \bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{i,j}$ . Die Zwischengruppen-Quadratsumme  $SS_A$  aus (9.13) ersetzen wir durch

$$SRS_A = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{R}_{i\bullet} - \bar{R}_{\bullet\bullet})^2, \qquad (9.17)$$

und man kann zeigen:

**Kruskal-Wallis-Test:** Es gelte Voraussetzung 9.7.2. Ist die Hypothese  $H_0$  erfüllt, so ist

$$\frac{12}{N(N+1)} \cdot SRS_A = \sum_{i=1}^{I} \sum_{i=1}^{n_i} \bar{R}_{i\bullet}^2 - 3 \cdot (N+1)$$

approximativ  $\chi^2$ -verteilt mit I-1 Freiheitsgraden.



#### Bemerkung zu Hintergrund 9.7.3.

Die beim Kruskal-Wallis-Test gemachte Approximation durch die  $\chi^2$ -Verteilung ist für praktische Bedürfnisse als hinreichend zu erachten, wenn gilt:

$$I=2, \quad n_1, n_2 \ge 4, \ N \ge 20, \qquad \text{oder} \quad I=3, \quad n_1, n_2, n_3 \ge 5,$$
 
$$\text{oder} \quad I \ge 4, \quad n_1, \dots, n_I \ge 4.$$



# Durchführung 9.7.4

Den eben besprochenen Kruskal-Wallis-Test findet man im Menü  $Statistik \rightarrow Nichtparametrische Verfahren \rightarrow Vergleich für mehrere unabhängige Stichproben (Gruppen). Dabei müssen die Daten in folgender Form vorliegen: In einer Variablen befinden sich alle Daten, in einer anderen die jeweiligen Gruppenzugehörigkeiten. Erstere Variable wählt man als <math>Abhängige$  Variable, zweitere als Gruppierungsvariable. Anschließend drückt man auf Zusammenfassung.



#### Beispiel 9.7.5

Wenden wir dieses Verfahren nun ebenfalls auf die Weinprüferdaten aus Beispiel 9.6.4 an. Auch der Kruskal-Wallis-Test aus Abbildung 9.12 zeigt, dass die Hypothese, die Prüfer würden die Weinsorten gleichartig bewerten, abzulehnen ist. Nach unseren Überlegungen aus Beispiel 9.6.4 sollten wir auch hier erneut das Testverfahren

	Unabh	ängige (G	rupp.) Vari	iesamturteil (Wein.sta) iable: Prüfer N= 228) =39,87527 p =,0000
Abh.:	Code	Gültige	Rang-	
Gesamturteil		N	summe	
Amthor	101	35	4557,000	
Muck	102	28	1239,000	
Münter	103	30	3503,500	
Reiß	104	45	5555,000	
Stahl	105	50	6003,000	
Thiergärtner	106	40	5248,500	

Abb. 9.12: Kruskal-Wallis ANOVA der sechs Weinprüfer.

	Unabh	ängige (G	rupp.) Vari	Gesamturteil (Wein.sta) iable: Prüfer N= 200) =1,701151 p =,7905
Abh.:	Code	Gültige	Rang-	
Gesamturteil		N	summe	
Amthor	101	35	3722,500	
Münter	103	30	2807,500	
Reiß	104	45	4478,500	
Stahl	105	50	4811,000	
Thiergärtner	106	40	4280,500	

Abb. 9.13: Kruskal-Wallis ANOVA von nur fünf der sechs Weinprüfer.

auf die fünf Prüfer ohne Muck anwenden (Abbildung 9.13). Das Resultat zeigt, dass der Kruskal-Wallis-Test unserer Vermutung nicht widerspricht, dass wenigstens die fünf übrigen Prüfer die von ihnen getesteten Weine auf ähnlichem Niveau bewerten.  $\hfill \Box$ 

In den Abschnitten 9.6 und 9.7 gingen wir bisher davon aus, dass die einzelnen Stichproben, die es zu vergleichen gilt, voneinander unabhängig sind. In den Abschnitten 9.1 und 9.2 hatten wir jedoch bereits eine Situation kennengelernt, bei der zwar zwei Stichproben vorlagen mit interner Unabhängigkeit, bei der jedoch zwischen den Stichproben eine Abhängigkeit bestand. Diese wurde durch zusammengehörige Paare (matched pairs) verursacht, und es ist naheliegend, diese Situation auf mehr als zwei Stichproben zu verallgemeinern, so dass man allgemeiner von zusammengehörigen Tupeln sprechen könnte. Üblicher ist es jedoch, dies als Messwiederholungen zu bezeichnen, da eine solche Abhängigkeit häufig durch mehrfache Messungen an jeweils denselben Objekten entsteht. Beispielsweise werden mehrere Probanden, welche jeweils unabhängig voneinander sind, wiederholt einer bestimmten Analyse unterzogen, etwa im Rahmen einer Langzeitstudie.



#### Voraussetzung 9.7.6

Die Zufallsvariablen  $X_{i,j}, j=1,\ldots,n$ , der i-ten Stichprobe besitzen die gleiche stetige Verteilung,  $i=1,\ldots,I$ . Es sei  $N:=I\cdot n$ . Ferner sind alle Zufallsvektoren  $(X_{1,j},\ldots,X_{I,j})$  unabhängig voneinander. Die Komponenten eines Vektors untereinander dagegen dürfen Abhängigkeiten aufweisen.

Es wird nun die Hypothese  $H_0$  untersucht, dass sogar *alle* Zufallsvariablen ein und derselben Verteilung unterliegen.



#### Hintergrund 9.7.7

Bezeichnet  $R_{i,j}$  den zufälligen (mittleren) Rang von  $X_{i,j}$  im j-ten Tupel  $(X_{1,j},\ldots,X_{I,j})$ , so definieren wir die Rang-Gesamtsumme  $R_{\bullet \bullet} = \sum_{j=1}^n \sum_{i=1}^I R_{i,j} = \frac{nI(I+1)}{2}$  und die i-te Rang-Gruppensumme  $R_{i \bullet} = \sum_{j=1}^n R_{i,j}$ . Die mittlere Quadratsumme

$$MRS_A = \frac{1}{I} \sum_{i=1}^{I} (R_{i\bullet} - \frac{1}{I} R_{\bullet \bullet})^2$$
(9.18)

 $\Diamond$ 

der Abweichungen zwischen den Messungen wird zur Definition einer Teststatistik verwendet:

**Friedman-Test:** Es gelte Voraussetzung 9.7.6. Ist die Hypothese  $H_0$  erfüllt, so ist

$$\frac{12}{n(I+1)} \cdot MRS_A = \frac{12}{nI(I+1)} \sum_{i=1}^{I} R_{i\bullet}^2 - 3n(I+1)$$

approximativ $\chi^2\text{-verteilt}$ mit I-1Freiheitsgraden.

Auch hier ist die Approximation durch die  $\chi^2$ -Verteilung wieder nur dann zulässig, wenn die Teilstichprobenumfänge groß genug sind, etwa  $n \geq 10$  für I=3 und  $n \geq 5$  für  $I \geq 4$ .



#### Durchführung 9.7.8

Der Friedman-Test ist analog zum Kruskal-Wallis-Test aus Durchführung 9.7.4 implementiert, nur wählt man hier Vergleich für mehrere abhängige Stichproben (Variablen). Dabei müssen die einzelnen Stichproben in einzelnen Variablen vorliegen, so dass jede Zeile ein zusammengehöriges Tupel beschreibt. Man wählt dann einfach alle Variablen aus und klickt auf Zusammenfassung.

					lls Konkor			
			ANOVA Chi <sup>2</sup> (N = 35, FG = 3) = 63,41143 p = ,0000 Konkordanz-Koeffizient = ,60392 Mittl. Rang r = ,592					
		Mittl.	Rang-	Mittelw.	Stdabw.			
	Variable	Rang	summe					
	Trunk1	2,057143	72,0000	2,197609	0,116506			
	Trunk2	2,057143	72,0000	2,189411	0,090837			
	Trunk3	4,000000	140,0000	2,503286	0,103429			
(a)	Trunk4	1,885714	66,0000	2,187582	0,089164			
(4)								
					lls Konkor			
		ANOVA C	hi <sup>2</sup> (N = 35	, FG = 2)	= ,6857143	3 p = ,70974		
		ANOVA C	hi <sup>2</sup> (N = 35	, FG = 2)	= ,6857143			
		ANOVA C	hi <sup>2</sup> (N = 35	, FG = 2)	= ,6857143	3 p = ,70974		
	Variable	ANOVA C Konkorda	hi² (N = 35 nz-Koeffizi	o, FG = 2) ent = ,009	= ,6857143 80 Mittl. Ra	3 p = ,70974		
	Variable Trunk1	ANOVA C Konkorda Mittl. Rang	hi² (N = 35 nz-Koeffizi Rang- summe	o, FG = 2) ent = ,009	= ,6857143 80 Mittl. Ra Stdabw.	3 p = ,70974		
		ANOVA C Konkorda Mittl. Rang 2,057143	hi² (N = 35 nz-Koeffizi Rang- summe 72,00000	5, FG = 2) ent = ,009 Mittelw.	= ,6857143 80 Mittl. Ra Stdabw. 0,116506	3 p = ,70974		
(c)	Trunk1	ANOVA C Konkorda Mittl. Rang 2,057143 2,057143	hi² (N = 35 nz-Koeffizi Rang- summe 72,00000 72,00000	6, FG = 2) ent = ,009 Mittelw. 2,197609	= ,6857143 80 Mittl. Ra Stdabw. 0,116506 0,090837	3 p = ,70974		
(c)	Trunk1 Trunk2	ANOVA C Konkorda Mittl. Rang 2,057143 2,057143	hi² (N = 35 nz-Koeffizi Rang- summe 72,00000 72,00000	6, FG = 2) ent = ,009 Mittelw. 2,197609 2,189411	= ,6857143 80 Mittl. Ra Stdabw. 0,116506 0,090837	3 p = ,70974		

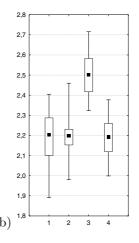


Abb. 9.14: Friedman-Test und Boxplots zu Beispiel 9.7.9.

#### Beispiel 9.7.9

Die Datei Erfrischung.sta enthält (fiktive) Daten zu 35 Personen, welche jeweils vier verschiedene Erfrischungsgetränke testeten. Deren erfrischende Wirkung schlägt sich u. a. in der kurzfristigen Steigerung eines bestimmten Blutwertes nieder. Gibt es Unterschiede in der Wirksamkeit der Getränke?

Um dies zu untersuchen, wird der Friedman-Test durchgeführt, welcher die Hypothese  $H_0$ : 'Alle Getränke wirken gleich' prüft. Das Resultat aus Abbildung 9.14 (a) zeigt, dass diese Hypothese abzulehnen ist. Ein Blick auf die Boxplots aus Abbildung 9.14 (b) macht deutlich, dass dies wohl auf eine stärkere Wirkung von Getränk 3 zurückzuführen ist. Zwischen der Wirksamkeit der Getränke 1, 2 und 4 dagegen lässt sich kein signifikanter Unterschiede feststellen, siehe Abbildung 9.14 (c).



# 12.1.5 Kontrollkarten für Messdaten: Komplexere Ansätze

In Abschnitt 12.1.3 beschäftigten wir uns mit den elementarsten Arten von Kontrollkarten für Messdaten, den sog. Shewhartkarten. Charakteristisch für diese ist, dass die nacheinander aufgetragenen Kontrollstatistiken vollkommen unabhängig voneinander sind, so dass diese Karten, wenn man keine Verlaufsregeln berücksichtigt, völlig 'gedächtnislos' sind. Deshalb wollen wir in diesem Abschnitt drei

populäre Kontrollkarten mit Gedächtnis vorstellen, d. h. Karten, deren Kontrollstatistiken auch immer frühere Prozesswerte berücksichtigen. Es sind dies die EWMA-Kontrollkarte, die Moving-Average-Kontrollkarte und die CUSUM-Karte. Auch diese Karten beschränken sich auf Prozesse, die Voraussetzung 12.1.3.1 genügen, d. h. der beobachtete Prozess  $(X_t)_T$  ist univariat, und im Zustand statistischer Kontrolle wird angenommen, dass der Prozess unabhängig und identisch verteilt (i. i. d.) ist mit Randverteilung  $N(\mu_0, \sigma_0^2)$ .

Beginnen wir mit der EWMA-Karte, wobei EWMA hier wieder als Kürzel für den in Abschnitt 11.4.1, Hintergrund 11.4.1.8, besprochenen Ansatz des einfachen exponentiell gewichteten gleitenden Durchschnitts steht. Die EWMA-Karte stellt eine Alternative zur  $\bar{X}$ -Karte aus Hintergrund 12.1.3.2 dar, d.h. sie ist geeignet, eine Änderung im Prozessniveau  $\mu_0$  festzustellen. Zur Kontrolle der Prozessstreuung dagegen müssen wir weiterhin auf S- oder Moving-Range-Karte zurückgreifen, vgl. Hintergrund 12.1.3.12.



#### Hintergrund 12.1.5.1

Sei  $X_{t,1},\ldots,X_{t,n_t}$  die zur Zeit t aufgenommene Teilstichprobe des Prozesses vom Umfang  $n_t$ . Zwar ist die EWMA-Karte in STATISTICA tatsächlich für beliebige  $n_t$  implementiert, bei unserer Beschreibung der EWMA-Karte beschränken wir uns jedoch vereinfachend auf die Situation konstanter Teilstichprobenumfänge  $n_t=n$ , wobei auch n=1 erlaubt ist. Es bezeichne  $\bar{X}_t$  das zugehörige arithmetische Mittel, welches im Falle n=1 gleich  $X_t$  ist. Im Zustand statistischer Kontrolle ist  $\bar{X}_t$  verteilt gemäß  $N(\mu_0,\frac{\sigma_0^2}{n})$ .

Als Kontrollstatistik betrachten wir nun, für gewähltes  $\lambda \in (0;1)$ ,

$$Z_t := (1 - \lambda) \cdot Z_{t-1} + \lambda \cdot \bar{X}_t, \quad t \ge 1, \qquad Z_0 := \mu_0.$$

Dann besitzt auch  $Z_t$  den Erwartungswert  $\mu_0$ , jedoch die Varianz

$$V[Z_t] = \frac{\sigma_0^2}{n} \cdot \frac{\lambda}{2-\lambda} \cdot \left(1 - (1-\lambda)^{2t}\right) \quad \underset{t \to \infty}{\longrightarrow} \quad \frac{\sigma_0^2}{n} \cdot \frac{\lambda}{2-\lambda}.$$

Entsprechend konstruiert man in Phase II, wenn die Werte für  $\mu_0$  und  $\sigma_0^2$  als bekannt angenommen werden, die EWMA-Karte mit  $k\text{-}\sigma\text{-}Grenzen$  wie folgt:

$$UCL_{t} = \mu_{0} + k \cdot \frac{\sigma_{0}}{\sqrt{n}} \cdot \sqrt{\frac{\lambda}{2-\lambda} \cdot \left(1 - (1-\lambda)^{2t}\right)},$$
  
Mittellinie =  $\mu_{0}$ ,  
$$LCL_{t} = \mu_{0} - k \cdot \frac{\sigma_{0}}{\sqrt{n}} \cdot \sqrt{\frac{\lambda}{2-\lambda} \cdot \left(1 - (1-\lambda)^{2t}\right)}.$$

In Phase I dagegen ersetzt man  $\mu_0$  und  $\sigma_0$  durch die in Hintergrund 12.1.3.2 diskutierten Schätzer.

Design einer EWMA-Karte: Da die Kontrollstatistik  $Z_t$  nicht mehr normalverteilt ist, ist die Wahl der symmetrischen k- $\sigma$ -Grenzen wie auch des konkreten Wertes von k durch andere Überlegungen zu rechtfertigen. Hierzu berechnet man approximativ ARLs für verschiedene Kombinationen von  $(\lambda, k)$ , und wählt dann solche Pärchen  $(\lambda, k)$  aus, für die sich gute ARL-Werte ergeben. Eine detaillierte Diskussion findet der Leser in Abschnitt 8-2.2 bei Montgomery (2005), dort insbesondere Tabelle 8-10. Kurz gefasst lässt sich Folgendes sagen: Generell sollte  $\lambda$  zwischen 0,05 und 0,25 gewählt werden, wobei kleine  $\lambda$  besser geeignet sind zur Entdeckung kleiner Verschiebungen des Prozessniveaus  $\mu$  und umgekehrt. Für  $\lambda > 0,1$  ist k=3 empfehlenswert, für  $\lambda \leq 0,1$  dagegen  $2,6 \leq k \leq 2,8$ . (Montgomery, 2005, S. 411)

In Hintergrund 12.1.5.1 haben wir in der Rekursion von  $Z_t$  als Parameter  $\lambda$  gewählt, während wir in Hintergrund 11.4.1.8  $\alpha$  verwendet haben. Dies geschah nicht, um den Leser zu verwirren. Leider hat STATISTICA in seiner Menüführung die Bezeichnung gewechselt. Diese Rekursion jedenfalls, die ja die gesamte Prozessvergangenheit exponentiell gewichtet berücksichtigt, ist es, welche die EWMA-Karte zu einer Karte mit Gedächtnis macht.

#### Durchführung 12.1.5.2

Eine EWMA-Karte erstellt man in völliger Analogie zu Durchführung 12.1.3.3, mit dem Unterschied, dass man nun

- $EWMA\ X$ -quer & R-Karte wählt, um diese mit einer R-Karte zu kombinieren, oder wenn n=1 ist, oder
- EWMA X-quer & S-Karte wählt, um diese mit einer S-Karte zu kombinieren, wobei  $n \geq 2$  sein muss.

Ferner muss man nun im Dialog Variablen für EWMA X-quer... im Feld Lambda Moving Average (...) den Wert für  $\lambda$  eingeben. Den Wert für k kann man erst anpassen, wenn man nachträglich das EWMA-Menü erneut aktiviert.

#### Bemerkung zu Durchführung 12.1.5.2.

Standardmäßig gibt STATISTICA auch hier wieder Histogramme für die EWMA-Karten mit aus, allerdings ist fraglich warum. Die einzelnen  $Z_t$  sind nicht identisch normalverteilt, auf Grund der seriellen Abhängigkeit von  $(Z_t)_{\mathcal{T}}$  ist seine Fähigkeit als Dichteschätzer ohnehin eingeschränkt.

Ferner bietet STATISTICA auch hier wieder die Überprüfung von Verlaufsregeln an. Dies sollte man tunlichst unterlassen, da auf Grund der seriellen Abhängigkeit von  $(Z_t)_{\mathcal{T}}$  diese ständig verletzt werden, was aber ganz normal ist.





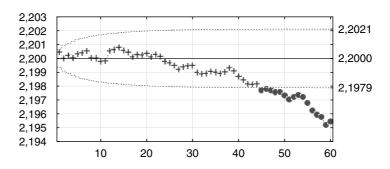


Abb. 12.13: EWMA-Karte zu Beispiel 12.1.5.3.



#### Beispiel 12.1.5.3

Setzen wir Beispiel 12.1.3.6 fort. Für die dortigen Fräsdaten (n=1) in der Datei FraesenAb.sta sind die Werte  $\mu_0=2,20$  und  $\sigma_0=0,00491$  für den Kontrollzustand bekannt. Wir erstellen nun eine EWMA-Karte mit  $\lambda=0,05$ , wobei wir entsprechend der Empfehlung von Montgomery (2005) in Hintergrund 12.1.5.1 k=2,7 wählen.

Das Resultat ist in Abbildung 12.13 wiedergegeben. Der Abwärtstrend, den wir schon in Beispiel 12.1.3.6 entdeckten, kommt hier noch deutlicher zur Geltung. Die Kontrollgrenzen werden bereits bei t=45 verletzt, im Gegensatz zur Einzelwertkarte, wo dies erst bei t=59 der Fall war. Somit scheint die EWMA-Karte klar überlegen. Berücksichtigt man bei der Einzelwertkarte jedoch auch Verlaufsregeln, siehe Beispiel 12.1.3.9, was bei der EWMA-Karte nicht möglich ist, dann relativiert sich dieser Vorteil wieder, denn bereits zwischen t=39 und t=42 wird ein erstes kritisches Muster entdeckt. Da jedoch Verlaufsregeln schwerer manuell nachzuprüfen sind, bleibt zumindest ein praktischer Vorteil auf Seiten der EWMA-Karte.

Auch die Moving-Average-Karte wird, genau wie die EWMA-Karte, durch ein Glättungsverfahren der Zeitreihenanalyse motiviert, und zwar durch den einfachen gleitenden Durchschnitt, siehe Hintergrund 11.4.1.5. Hierzu muss man die Länge w des Fensters angeben, welches über die Prozessdaten gleitet, und als Kontrollstatistik wird entsprechend der Mittelwert des Fensterinhaltes betrachtet. Somit besitzt auch diese Karte ein Gedächtnis, in diesem Fall der Länge w-1.



#### Hintergrund 12.1.5.4

Sei  $X_{t,1}, \ldots, X_{t,n_t}$  die zur Zeit t aufgenommene Teilstichprobe des Prozesses vom Umfang  $n_t$ . Zwar ist die Moving-Average-Karte in STATISTICA

tatsächlich für beliebige  $n_t$  implementiert, bei unserer Beschreibung der Moving-Average-Karte beschränken wir uns jedoch vereinfachend auf die Situation konstanter Teilstichprobenumfänge  $n_t = n$ , wobei auch n = 1 erlaubt ist. Es bezeichne  $\bar{X}_t$  das zugehörige arithmetische Mittel, welches im Falle n = 1 gleich  $X_t$  ist. Im Zustand statistischer Kontrolle ist  $\bar{X}_t$  verteilt gemäß  $N(\mu_0, \frac{\sigma_0^2}{n})$ .

Als Kontrollstatistik betrachten wir nun, für gewählte Fensterlänge  $w \in \mathbb{N}$ ,

$$Z_t := \begin{cases} \frac{1}{t} \cdot (\bar{X}_1 + \dots + \bar{X}_t), & t < w, \\ \frac{1}{w} \cdot (\bar{X}_{t-w+1} + \dots + \bar{X}_t), & t \ge w. \end{cases}$$

Dann ist  $Z_t$  selbst ebenfalls normalverteilt, und zwar gemäß  $N(\mu_0, \frac{\sigma_0^2}{n \cdot t})$ , falls t < w, und gemäß  $N(\mu_0, \frac{\sigma_0^2}{n \cdot w})$ , falls  $t \geq w$ . Allerdings sind, genau wie bei der EWMA-Karte auch, die einzelnen  $Z_t$  nun seriell abhängig, so dass eine Überprüfung von Verlaufsregeln wenig Sinn macht und auch das Histogramm mit Vorsicht zu genießen ist.

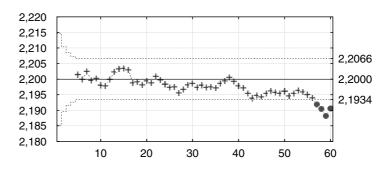
In Phase II, wenn die Werte für  $\mu_0$  und  $\sigma_0^2$  als bekannt angenommen werden, konstruiert man die *Moving-Average-Karte* mit k- $\sigma$ -Grenzen mit Mittellinie bei  $\mu_0$  und Kontrollgrenzen

$$UCL_t = \begin{cases} \mu_0 + k \cdot \frac{\sigma_0}{\sqrt{n \cdot t}}, & t < w, \\ \mu_0 + k \cdot \frac{\sigma_0}{\sqrt{n \cdot w}}, & t \ge w, \end{cases} \quad LCL_t = \begin{cases} \mu_0 - k \cdot \frac{\sigma_0}{\sqrt{n \cdot t}}, & t < w, \\ \mu_0 - k \cdot \frac{\sigma_0}{\sqrt{n \cdot w}}, & t \ge w. \end{cases}$$

In Phase I dagegen ersetzt man  $\mu_0$  und  $\sigma_0$  durch die in Hintergrund 12.1.3.2 diskutierten Schätzer.

Design einer Moving-Average-Karte: Da die Kontrollstatistik  $Z_t$  erneut normalverteilt ist, ist die Wahl der symmetrischen k-σ-Grenzen gerechtfertigt, alternativ können auch quantilbasierte Grenzen konstruiert werden. Bezüglich der Fensterlänge gilt, dass zur Entdeckung kleiner Shifts eher große Fensterlängen zu empfehlen sind, für große Shifts dagegen kleine. (Montgomery, 2005, S. 417)

Obwohl die Kontrollstatistik  $Z_t$  der Moving-Average-Karte gegenüber jener der EWMA-Karte den Vorteil hat, dass im Kontrollzustand ihre exakte Verteilung von einfacherer Gestalt ist als die der EWMA-Statistik, wird in der Praxis die EWMA-Karte bevorzugt, wohl insbesondere wegen der leichteren Implementierung. Schließlich muss bei der EWMA-Karte nur ein Wert im Arbeitsspeicher verbleiben, und nicht w-1 wie bei der Moving-Average-Karte.



**Abb. 12.14:** Moving-Average-Karte mit w = 5 zu Beispiel 12.1.5.6.



#### Durchführung 12.1.5.5

Eine Moving-Average-Karte erstellt man in völliger Analogie zu Durchführung 12.1.3.3, mit dem Unterschied, dass man nun

- MA X-quer & R-Karte wählt, um diese mit einer R-Karte zu kombinieren, oder wenn n=1 ist, oder
- MA X-quer & S-Karte wählt, um diese mit einer S-Karte zu kombinieren, wobei  $n \geq 2$  sein muss.

Ferner muss man nun im Dialog Variablen für MA X-quer... im Feld Moving Average Spanne den Wert für w eingeben. Den Wert für k kann man erst anpassen, wenn man nachträglich das Moving-Average-Menü erneut aktiviert.



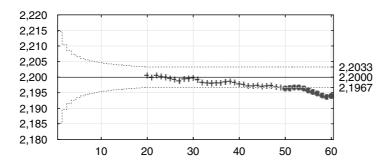
#### Bemerkung zu Durchführung 12.1.5.5.

Erstellt man, bei gewählter Fensterlänge w, eine Moving-Average-Karte gemäß Durchführung 12.1.5.5, so werden, wie in Hintergrund 12.1.5.4 beschrieben, die Kontrollgrenzen korrekt berechnet und angegeben für  $alle\ t$ . Leider jedoch werden die Kontrollstatistiken nur für  $t \geq w$  berechnet und aufgetragen. Somit ist mit den bei STATISTICA implementierten Moving-Average-Karten eine Kontrolle der ersten w-1 Werte nicht möglich.



# Beispiel 12.1.5.6

Setzen wir Beispiel 12.1.3.6 bzw. 12.1.5.3 fort; es ist n=1,  $\mu_0=2,20$  und  $\sigma_0=0,00491$ . Wir erstellen zwei Moving-Average-Karten mit diesen Vorgaben, einmal mit Fensterlänge w=5, siehe Abbildung 12.14, dann mit w=20, siehe Abbildung 12.15. Offenbar



**Abb. 12.15:** Moving-Average-Karte mit w = 20 zu Beispiel 12.1.5.6.

erkennt die Karte mit w=20 den Abwärtstrend wesentlich früher, übrigens nicht so früh wie obige EWMA-Karte, siehe Beispiel 12.1.5.3. Jedoch erkauft man sich diesen Vorteil bei STATISTICA damit, dass sich die ersten 19 Werte jeglicher Kontrolle entziehen.

Die CUSUM-Kontrollkarte<sup>9</sup> wird motiviert insbesondere durch die Schwäche der  $\bar{X}$ - bzw. Einzelwertkarte beim Erkennen kleiner Verschiebungen im Prozessniveau  $\mu$ , zumindest wenn keine Verlaufsregeln berücksichtigt werden. Die grundlegende Idee hierbei ist, dass sich viele kleine Abweichungen zu einer großen addieren. Die Anwendung der tabellarischen CUSUM-Karte, wie sie bei STATISTICA implementiert ist, beschränkt sich dabei auf einen Prozess mit Einzelwerten, d. h. wir betrachten  $(X_t)_T$  selbst, ohne dass wir Teilstichproben ziehen.

## Hintergrund 12.1.5.7

Vom Prozess  $(X_t)_T$  wird angenommen, dass er im Kontrollzustand unabhängig und identisch verteilt ist mit Randverteilung  $N(\mu_0, \sigma_0^2)$ , siehe Voraussetzung 12.1.3.1. In Phase II, wenn die Werte für  $\mu_0$  und  $\sigma_0^2$  als bekannt angenommen werden, sind die Kontrollstatistiken  $C_t^+$  und  $C_t^-$  definiert als

$$C_t^+ = \max \left(0 \; ; \; (X_t - \mu_0) - \epsilon \; + \; C_{t-1}^+ \right), \quad C_0^+ := 0,$$

$$C_t^- = \min \left(0 \; ; \; (X_t - \mu_0) + \epsilon \; + \; C_{t-1}^- \right), \quad C_0^- := 0.$$

Zur Interpretation betrachten wir die Statistik  ${C_t}^+$  etwas ausführlicher: Der Wert von  ${C_t}^+$  wächst gegenüber dem von  ${C_{t-1}}^+$  immer dann an, wenn  $X_t$  um mehr als  $\epsilon$  größer als  $\mu_0$  ist. Um hierbei den Toleranzparameter  $\epsilon$  zu bestimmen, folgt STATISTICA einer Empfehlung von Montgomery (2005) und erwartet die Vorgabe eines kritischen Shifts vom Betrage  $\delta \cdot \sigma_0$ , den es auf jeden Fall zu entdecken gilt; es ist dann  $\epsilon := \frac{\delta}{2} \cdot \sigma_0$ .



<sup>&</sup>lt;sup>9</sup>Cumulated Sum.

Die tabellarische CUSUM-Karte wird nun wie folgt konstruiert: Man wähle

Mittellinie = 0, 
$$UCL = k \cdot \sigma_0$$
,  $LCL = -k \cdot \sigma_0$ ,

und trage zur Zeit t auf die entsprechend konstruierte Karte nun jene der beiden Statistiken  $C_t^+$  und  $C_t^-$  auf, die stärker von der 0 abweicht. Per Definition sind die aufgetragenen Kontrollstatistiken wieder seriell abhängig, so dass eine Überprüfung der Verlaufsregeln nicht sinnvoll ist, auch ein Histogramm ist wenig hilfreich. In Phase I ersetzt man  $\mu_0$  und  $\sigma_0$  durch die in Hintergrund 12.1.3.2 diskutierten Schätzer.

**Design einer CUSUM-Karte:** Wenn durch äußere Gegebenheiten nicht anders erforderlich, ist bei der Wahl des kritischen Shifts ein Wert  $\delta=1$ , also  $\epsilon=\frac{\sigma_0}{2}$ , empfehlenswert. Bei den k- $\sigma$ -Grenzen sollte k zwischen 4 und 5 gewählt werden. Diese Ratschläge beruhen auf ARL-Überlegungen, wie bei der EWMA-Karte auch, siehe Hintergrund 12.1.5.1. (Montgomery, 2005, S. 395ff)



## Durchführung 12.1.5.8

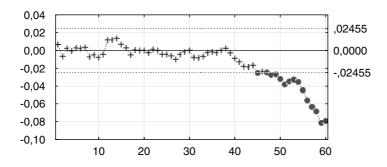
Eine CUSUM-Karte erstellt man in völliger Analogie zu Durchführung 12.1.3.3, mit dem Unterschied, dass man nun CUSUM-Karte Einzelwerte wählen muss, vorerst nur die Variable auswählt und OK klickt. Neben einer auf Schätzwerten basierenden CUSUM-Karte wird auch eine Moving-Range-Karte mit ausgegeben, siehe Hintergrund 12.1.3.12.

Die Werte für  $\delta$ ,  $\mu_0$ ,  $\sigma_0$  und k kann man erst anpassen, wenn man nachträglich das CUSUM-Menü erneut aktiviert. Ersteres gibt man auf der Karte Karten unter Entdecke Shift ab ein, die Übrigen wie gewohnt auf der Karte X (MA..) Einst.

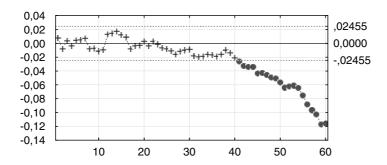


# Beispiel 12.1.5.9

Setzen wir die Beispiele 12.1.3.6, 12.1.5.3 und 12.1.5.6 fort; es ist  $\mu_0=2,20$  und  $\sigma_0=0,00491$ . In Abbildung 12.16 ist eine CUSUM-Karte mit  $\delta=1$  und k=5 zu sehen; offenbar erkennt diese ab t=45 den Abwärtstrend als kritisch. Die Sensibilität der CUSUM-Karte kann erhöht werden, indem man beispielsweise den Toleranzparameter verkleinert. Bei der CUSUM-Karte in Abbildung 12.17 wurde  $\delta=\frac{1}{2}$  und k=5 gewählt, entsprechend wird der Abwärtstrend bereits bei t=41 erkannt. Es sollte aber beachtet werden, dass man sich eine erhöhte Sensibilität auch mit erhöhten Raten falscher Alarmmeldungen erkauft.



**Abb. 12.16:** CUSUM-Karte mit  $\delta = 1$  und k = 5 zu Beispiel 12.1.5.9.



**Abb. 12.17:** CUSUM-Karte mit  $\delta = \frac{1}{2}$  und k = 5 zu Beispiel 12.1.5.9.