

4 Statistische Verfahren

Viele Aktivitäten, die in der Systemanalyse untersucht werden, haben die Eigenschaft, daß ihre Ergebnisse nicht genau vorausgesagt werden können. Es läßt sich meist nur ein Bereich der möglichen Ergebnisse angeben. Wird eine solche Aktivität mehrfach und unabhängig durchgeführt, dann werden in zufälliger Folge verschiedene Ergebnisse aus dem jeweiligen Wertebereich realisiert. Man nennt daher eine derartige Aktivität ein *Zufallsexperiment*. Simulationsexperimente mit stochastischen Modellen sind als solche Zufallsexperimente aufzufassen. Es bedarf statistischer Verfahren, um die Zufallsaspekte in ein Simulationsmodell einzubringen (4.1 und 4.2), entsprechende statistische Verteilungen für das Modell abzuleiten (4.3) und die Ergebnisse von Simulationsexperimenten exakt auszuwerten (4.4).

4.1 Zufallszahlenerzeugung

4.1.1 Die Bedeutung von Zufallszahlen für die Simulation

Wenn man das Verhalten zufallsabhängiger Prozesse an einem entsprechenden Simulationsmodell untersuchen will, kommt man nicht umhin, die real vorhandene Zufälligkeit in geeigneter Weise im Modell nachzubilden. Hierfür werden sogenannte Zufallszahlen verwendet. Beispielsweise wurden die Abstände zwischen Kundenanrufen und die Bedienungsdauer im Reservierungsmodell (2.4.1.2) jeweils durch eine Zufallszahl festgelegt. In gleicher Weise wurde im Beispiel zur Simulation ausfallanfälliger Systeme (2.4.3) die Zufallsgröße "Lebensdauer eines Maschinenteils" nachgebildet. Allgemein läßt sich feststellen:

Zufallszahlen dienen zur Realisierung von *Zufallsvariablen*. Der betreffenden Variablen wird mittels einer Zufallszahl ein Wert ihres Wertebereichs zugeordnet.

Dabei sind Zufallszahlen zunächst als Ergebnis **stochastischer** physikalischer Vorgänge aufzufassen. Als einfachstes und bekanntestes Beispiel wird dafür fast immer der Wurf eines idealen Würfels angegeben, als dessen Ergebnis eine zufällige Folge von

Zahlen aus der Menge 1, 2, ..., 6 entsteht, z. B. 5, 2, 1, 1, 6, 3, 4, Ein weiterer physikalischer Prozeß, der Zufallszahlen liefert, ist der Zerfall radioaktiver Substanzen. Als Zufallszahlen können dabei die Anzahl der je Zeiteinheit zerfallenden Teilchen registriert werden. Auf diese oder ähnliche Weise gewonnene Zufallszahlen bezeichnet man als *echte* oder *natürliche Zufallszahlen*. Damit wird angezeigt, daß sie Produkt realer physikalischer Prozesse sind. Ein Verfahren zur Erzeugung von Zufallszahlen wird *Zufallszahlengenerator* genannt.

Zufallszahlentabellen auf der Basis von Beobachtungen stochastischer physikalischer Prozesse eignen sich aber kaum für die Simulation komplexer Prozesse, die praktisch nur mit Hilfe von Rechnern durchführbar ist. Da aus statistischen Gründen jede Zufallsgröße in vielfacher Wiederholung im Modell realisiert werden muß, ist für die Nachahmung eines Prozesses mit mehreren Zufallsvariablen eine sehr große Anzahl von Zufallszahlen notwendig. Somit wären riesige Zufallszahlentabellen im Rechner zu speichern. Die Speicherung der Zufallszahlen in Dateien würde keinen effizienten Zugriff ermöglichen. Sinnvoller ist es, die benötigten Zufallszahlen erst dann zu erzeugen, wenn sie bei der Simulation benötigt werden. Das bedeutet, daß als Zufallszahlengenerator ein Algorithmus zu implementieren ist, dessen Ausgabe einen brauchbaren Ersatz für natürliche Zufallszahlen darstellt.

Es leuchtet ein, daß es bei Zufallszahlen nicht auf die Art ihrer Erzeugung ankommt. Wesentlich ist nur, daß sie in ihrer Gesamtheit Eigenschaften aufweisen, die es gestatten, Zufallsgrößen nachzubilden. Mit anderen Worten: Wichtig ist nur, daß die Zahlen in statistischer Hinsicht annähernd gleiche Eigenschaften wie die nachzuahmende reale Zufallsgröße aufweisen.

Im Sinne dieser Feststellung wurden algorithmische Verfahren zur Berechnung von Zahlen gesucht, die Zahlenfolgen mit vorgegebenen statistischen Eigenschaften liefern. Es ist klar, daß die mit Hilfe von Algorithmen berechneten Zahlen nicht das Ergebnis stochastischer, sondern nur deterministischer Prozesse sein können. Deshalb bezeichnet man solche Zufallszahlen als *Pseudo-Zufallszahlen*.

Ein Vorteil der synthetisch konstruierten Pseudo-Zufallszahlen ist ihre **Reproduzierbarkeit**. Die gleiche Zufallszahlenfolge kann beliebig oft rekonstruiert werden. Diese Eigenschaft ist für Kontrollrechnungen und für die Durchführung von Alternativsimulationen wichtig. Zufallszahlengeneratoren, die sehr schnell in praktisch endloser Folge voneinander unabhängige Zufallszahlen produzieren und sie zur Weiterverarbeitung dem Simulationsprozeß zuführen, stellen somit das Herz einer jeden stochastischen Simulation dar. Es ist offensichtlich, daß die Simulationsergebnisse ganz wesentlich von der Qualität der eingesetzten Zufallszahlengeneratoren abhängen.

4.1.2 Erzeugung gleichverteilter Zufallszahlen

In zufallsabhängigen Prozessen treten Zufallsvariablen mit sehr unterschiedlicher Charakteristik auf, ausgedrückt durch das **statistische Verteilungsgesetz**, dem sie folgen, und die zugehörigen **Verteilungsparameter** wie Mittelwert und Varianz. Zur adäquaten Nachbildung solcher Zufallsvariablen sind daher auch Zufallszahlen mit ganz unterschiedlichen wahrscheinlichkeitstheoretischen Merkmalen erforderlich.

Die dafür aufgestellten Erzeugungsalgorithmen schließen jedoch i. d. R. in einem ersten Schritt die Berechnung gleichverteilter Zufallszahlen ein, die in einem zweiten Schritt transformiert werden. Zufallszahlen X sind in einem Intervall (a, b) *gleichverteilt*, wenn sie der Verteilungsfunktion

$$F(X) = 0 \quad \text{für } X \leq a$$

$$F(X) = \frac{X-a}{b-a} \quad \text{für } a < X < b$$

$$F(X) = 1 \quad \text{für } X \geq b$$

bzw. der Dichtefunktion

$$f(X) = \frac{1}{b-a} \quad \text{für } a < X < b$$

$$f(X) = 0 \quad \text{für } X \leq a \text{ oder } X \geq b$$

folgen. Erwartungswert E und Varianz Var der Gleichverteilung bestimmen sich aus¹:

$$E(X) = \frac{a+b}{2} \quad \text{bzw.} \quad Var(X) = \frac{(b-a)^2}{12}$$

Die für stetige Gleichverteilungen gültigen Beziehungen lassen sich näherungsweise auch für diskrete Zufallszahlen verwenden. Eine Folge gleichverteilter Zahlen kann man durch Würfeln oder durch Ziehen aus einer Urne erhalten, da für jede Zahl die gleiche Wahrscheinlichkeit dafür besteht, daß sie als nächste gezogen wird. Ob dies für ein Berechnungsverfahren für (Pseudo-) Zufallszahlen auch gilt, muß jeweils geprüft werden. Die Statistik stellt dafür geeignete Prüfmethode in Form sogenannter *Anpassungstests* zur Verfügung, mit denen eine gegebene Zahlenfolge daraufhin untersucht werden kann, ob sie einem bestimmten statistischen Verteilungsgesetz genügt.

¹ Als Erwartungswert bezeichnet man allgemein den Mittelwert einer theoretischen Verteilung. Die Varianz (bzw. die Standardabweichung als deren Wurzel) beschreibt die Schwankungen einer Verteilung um deren Mittelwert.

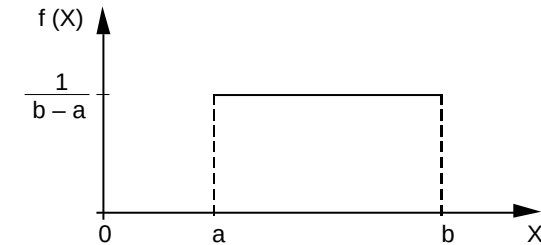


Abb. 4-1: Wahrscheinlichkeitsdichte der Gleichverteilung

Zu den bekanntesten **Methoden für die Zufallszahlenerzeugung** gehört die (multiplikative) *Kongruenz-Methode*, die von D. H. Lehmer bereits 1951 eingeführt wurde [Leh51]. Sie liefert Zahlenfolgen nach der rekursiven Rechenvorschrift:

$$z_{i+1} \equiv (A \cdot z_i) \bmod P \quad i = 0, 1, \dots \quad (4.1-1)$$

Eine weitere Zufallszahl z_{i+1} ergibt sich somit aus der vorangegangenen Zahl z_i durch Multiplikation mit A und durch Modulo-Division (Rest der ganzzahligen Division) durch P . Die Größen A und z_0 stellen festzusetzende ganze Zahlen aus dem Bereich $[1, P-1]$ dar, wobei P eine Primzahl sein sollte. Die durch das Verfahren erzielbare Zahlenfolge bewegt sich ebenfalls im Intervall $[1, P-1]$. Durch Wahl eines sehr großen Wertes für P (maximal bis zur Rechengenauigkeit des Computers) kann der Wertebereich der Zahlenfolge sehr weit ausgedehnt werden.

Die rekursive Berechnung sei an einem Beispiel erläutert (das wegen der kleinen Periodenlänge allerdings nicht realistisch ist). Mit $A = 5$, $P = 17$ und $z_0 = 5$, d. h.

$$z_{i+1} \equiv (5 \cdot z_i) \bmod 17$$

ergibt sich die in der dritten Zeile von Tabelle 4-1 angegebene Zahlenfolge. Die Zahlenfolge ist **periodisch**. Sobald ein bereits erzeugter Wert erneut entsteht – im Beispiel ab $i = 16$ – wiederholt sich die Folge.

Tabelle 4-1: Zahlenbeispiel für die multiplikative Kongruenz-Methode

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...
$A \cdot z_i$	25	40	30	65	70	10	50	80	60	45	55	20	15	75	35	5	25	...
z_{i+1}	8	6	13	14	2	10	16	12	9	11	4	3	15	7	1	5	8	...

Zufallszahlenfolgen dürfen eine solche Eigenschaft nicht aufweisen, da sonst stochastische Abhängigkeiten eingeführt werden. Man verwendet daher nur Zufallszahlen, die **innerhalb einer Periode** liegen. Die Länge der Periode läßt sich durch die Wahl von P beeinflussen. Für eine geeignete Wahl der Parameter A und P gibt es mehrere Möglichkeiten, z. B.

$$P = 2^{31} - 1$$

$$A = p^k$$

wobei p eine Primzahl als Faktor von $P-1$ und k ein positiver Exponent ist. P stellt dabei die größte binär darstellbare Primzahl dar, die in einem Maschinenwort heute üblicher Länge (32 Bit) gespeichert werden kann. Es läßt sich nachweisen, daß die bei geeigneter Wahl von A und z_0 maximal erzielbare Periodenlänge 2^{29} beträgt.

Die mit Hilfe dieses Verfahrens gebildeten Zahlen z_i werden meist noch durch P dividiert:

$$z'_i = \frac{z_i}{P} \quad (i = 1, 2, 3, \dots) \quad (4.1-2)$$

Die so erzeugten Pseudozufallszahlen liegen im Intervall $(0, 1)$ und erfüllen gut die Forderung nach Gleichverteilung. Die Umrechnung (4.1-2) entspricht einer Abbildung des Intervalls $(0, P)$ auf das Intervall $(0, 1)$ (s. Abb. 4-2).

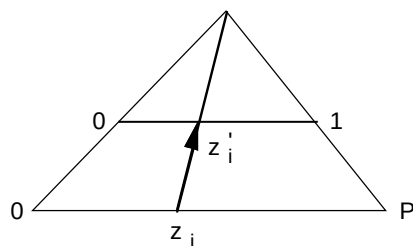


Abb. 4-2: Transformation zur Erzeugung von $(0, 1)$ -gleichverteilten Zufallszahlen

Damit ist zugleich der Weg für die Erzeugung gleichverteilter Zufallszahlen X für ein beliebiges Intervall (a, b) aufgezeigt (s. Abb. 4-3):

$$\frac{x_i - a}{b - a} = z'_i \Rightarrow x_i = a + (b - a) z'_i \quad (4.1-3)$$

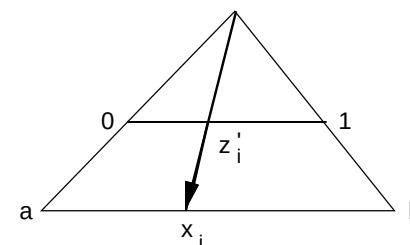


Abb. 4-3: Transformation zur Erzeugung von (a, b) -gleichverteilten Zufallszahlen

Diese Beziehung gestattet eine einfache Transformation von $(0, 1)$ -gleichverteilten Zufallszahlen Z' in gleichverteilte Zufallszahlen X eines beliebigen Intervalls (a, b) .

Um als Ersatz für echte Zufallszahlen dienen zu können, genügt es nicht allein, daß Pseudo-Zufallszahlen ein bestimmtes statistisches Verteilungsgesetz erfüllen. Sie müssen außerdem

- **stochastisch unabhängig** sein und
- **zufällig in der Folge** auftreten (nicht stellenweise gehäuft).

Für den Nachweis dieser Eigenschaften einer erzeugten Zahlenfolge werden ebenfalls statistische Testverfahren herangezogen. Einige Testverfahren für Zufallszahlen werden beispielsweise in [Bar83] behandelt. In diesem Rahmen werden auch einige aus der Literatur bekannte Zufallszahlengeneratoren exemplarisch diesen Testverfahren unterworfen und Empfehlungen für Generatoren guter statistischer Qualität abgeleitet. Gute statistische Eigenschaften besitzt nachweislich der Zufallszahlengenerator, der im folgenden Abschnitt 4.2 eingeführt wird.

Ein ebenfalls weit verbreitetes Verfahren zur Zufallszahlenerzeugung stellt die lineare Kongruenz-Methode dar, die von *M. Greenberger* als Verallgemeinerung der oben vorgestellten multiplikativen Kongruenz-Methode eingeführt wurde [Gre51], wobei hier ein weiterer Parameter B verwendet wird:

$$z_{i+1} \equiv (A \cdot z_i + B) \bmod P \quad (4.1-4)$$

$$\text{mit } i = 0, 1, \dots \text{ und } 0 \leq z_{i+1} < P$$

$$z'_i = \frac{z_i}{P-1}$$

Andere Verfahren zur Gewinnung von Pseudo-Zufallszahlen findet man z. B. bei *D. E. Knuth* [Knu69]. Die Bemühungen um effiziente Zufallsgeneratoren mit hoher statistischer Qualität können jedoch noch keineswegs als abgeschlossen angesehen werden.

4.1.3 Zufallszahlen beliebiger Verteilungen

Die in Simulationsmodellen nachzubildenden Zufallsvariablen sind natürlich nicht immer gleichverteilt, sondern können sehr verschiedenartigen theoretischen oder empirischen Verteilungsgesetzen unterliegen. Daher benötigt man auch Verfahren zur Erzeugung von Zufallszahlen, die z. B. einer Exponential-, Normal- oder einer empirischen Verteilung genügen. Solche Verfahren gehen in der Regel von gleichverteilten Pseudo-Zufallszahlen des Intervalls (0, 1) aus, die nach einer geeigneten Berechnungsvorschrift in Zufallszahlen der geforderten Verteilung **transformiert** werden.

Die Erzeugung beliebig verteilter Zufallszahlen erfolgt somit in zwei Schritten (s. Abb. 4-4). Der erste Schritt wurde im vorausgegangenen Abschnitt beschrieben. Für den 2. Schritt wird der folgende Satz (ohne Beweis) zugrundegelegt:

Im Intervall (0, 1) gleichverteilte Zufallszahlen $z_1, z_2, z_3, \dots$ können in Zufallszahlen $x_1, x_2, x_3, \dots$ transformiert werden, die nach der Verteilungsfunktion $F(X)$ verteilt sind, indem man die Umkehrfunktion von F auf z_i anwendet, also:

$$x_i = F^{-1}(z_i) \quad (4.1-5)$$

Durch **Inversion** der im Modell zu realisierenden Verteilungsfunktion $F(X)$ ergibt sich also eine Berechnungsvorschrift für entsprechend verteilte Zufallszahlen. Man spricht daher auch von *inverser Transformation*. Dieses Verfahren wird nachfolgend für jene statistischen Verteilungstypen angewendet, die in der Simulation am häufigsten benötigt werden (vgl. [Fra79], S. 58).

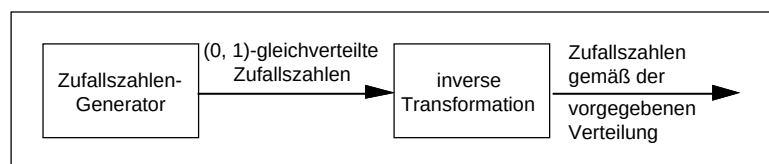


Abb. 4-4: Erzeugung von Zufallszahlen mit vorgegebener Verteilung

Exponentialverteilte Zufallszahlen

Bei Simulationsstudien sind häufig Zufallsvariablen im Modell zu realisieren, die einer Exponentialverteilung folgen (z. B. Zwischenankunftszeiten oder die Lebensdauer eines Gerätes). Die nachzubildende Zufallsvariable sei mit X bezeichnet, ihre Verteilungsfunktion mit $Z = F(X)$. Für die Exponentialverteilung lautet die (theoretische) Verteilungsfunktion:

$$Z = 1 - e^{-\lambda X} = F(X) \quad (4.1-6)$$

Die inverse Funktion F^{-1} ergibt sich durch Auflösung der Verteilungsfunktion nach X :

$$e^{-\lambda X} = 1 - Z$$

$$X = -\frac{1}{\lambda} \ln(1 - Z)$$

Z nimmt die Werte z_i in (0, 1) mit gleichmäßiger Verteilung an. Diesen wird nun

$$x_i = -\frac{1}{\lambda} \ln(1 - z_i) \quad (4.1-7)$$

zugeordnet. Erwartungswert und Standardabweichung (die Wurzel der Varianz) einer Exponentialverteilung betragen $1/\lambda$.

Die Transformation gleichverteilter Zufallszahlen in Zufallszahlen einer Exponentialverteilung über die Inverse der Verteilungsfunktion besteht, geometrisch veranschaulicht, in einer Zuordnung von Ordinatenwerten des Intervalls (0, 1) zu Abszissenwerten X über den Graphen der Verteilungsfunktion (s. Abb. 4-5).

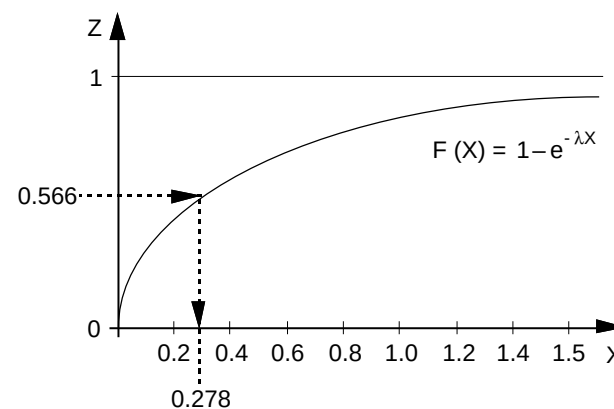


Abb. 4-5: Exponentialverteilung mit $\lambda = 3$

Erlang-verteilte Zufallszahlen

Die allgemein anwendbare Transformationsmethode kann grundsätzlich durch andere Erzeugungsverfahren ersetzt werden, wo Zufallsvariablen derart definiert sind, daß sie aus einfacher zu simulierenden Zufallsvariablen zusammengesetzt sind. Beispielsweise folgt das arithmetische Mittel aus k unabhängigen, exponentialverteilten Zufallsvariablen mit gleichem Mittelwert einer Erlang-Verteilung der Ordnung k. Dieser Verteilungstyp spielt bei Bedienungs-/Wartesystemen eine wichtige Rolle, denn in der Praxis ergeben sich oft Häufigkeitsverteilungen, die von einer exponentiellen Verteilungsform (z. B. aufgrund höherer Varianzen) abweichen. Derartige Häufigkeitsverteilungen lassen sich häufig durch eine Erlang-Verteilung mit den Parametern k und $\mu = 1/\lambda$ beschreiben. Zur Simulation einer Erlang-verteilten Zufallsvariablen benötigt man nicht die Kenntnis ihrer Verteilungsfunktion. Es genügt, deren Definition richtig anzuwenden, nämlich **k exponentialverteilte Zufallszahlen** nach dem oben genannten Verfahren zu erzeugen und ihr arithmetisches Mittel zu berechnen. In Verbindung mit der Transformationsmethode zur Erzeugung exponentialverteilter Zufallsvariablen lautet die Berechnungsvorschrift für eine Erlang-verteilte Zufallszahl x_j :

$$x_j = -\frac{1}{k \cdot \lambda} \sum_{i=1}^k \ln(1 - z_i) \tag{4.1-8}$$

Der Erwartungswert einer Erlang-Verteilung beträgt $1/\lambda$, die Varianz $1/(k \cdot \lambda^2)$.

Normalverteilte Zufallszahlen

Die explizite Inversion der Verteilungsfunktion, wie sie bei der Exponentialverteilung gezeigt wurde, ist bei der Normalverteilung – und auch bei einigen anderen theoretischen Verteilungen – mathematisch nicht möglich. Eine analytische Transformationsvorschrift kann folglich nicht abgeleitet werden. Man wendet daher Näherungsverfahren an.

Nach dem *zentralen Grenzwertsatz* der Statistik nähert sich die Summe von n identisch verteilten, unabhängigen Zufallsvariablen mit dem Mittelwert μ und der Varianz σ^2 einer Normalverteilung mit Mittelwert $n \cdot \mu$ und Varianz $n \cdot \sigma^2$ asymptotisch an. Werden gleichverteilte Zufallszahlen addiert, dann ist schon für relativ kleine n eine sehr gute Annäherung an das Normalverteilungsgesetz zu beobachten. Ein besonders günstiger Wert ist $n = 12$, da der Mittelwert einer im Einheitsintervall (0, 1) gleichförmig verteilten Zufallsvariablen $1/2$ und die Varianz $1/12$ betragen. Die Varianz der Summe von 12 dieser Variablen ist dann gerade gleich 1, ihr Mittelwert ist 6. Man erhält also mit für die Zwecke hinreichend guter Näherung standardisierte normalverteilte Zufallszahlen (d. h. mit einer Normalverteilung $N(0, 1)$), indem man 12 gleichverteilte Zufallszahlen aufaddiert und von der Summe die Zahl 6 subtrahiert:

$$z_j = \sum_{i=1}^{12} z_i - 6 \tag{4.1-9}$$

Eine Transformation der aus dieser Berechnung resultierenden normalverteilten Zufallszahlen mit dem Erwartungswert 0 und der Varianz 1 in eine beliebige Normalverteilung mit dem Erwartungswert μ und der Standardabweichung σ erfolgt nach der Formel:

$$x_j = z_j \cdot \sigma + \mu \tag{4.1-10}$$

Zufallszahlen empirischer Verteilungen

Häufig wird man es in der Simulation mit Zufallsvariablen zu tun haben, für die keine theoretischen Verteilungstypen, sondern nur empirische Verteilungen ermittelt werden können, die in Form von **Histogrammen** gegeben sind. Als Beispiel sei die Verarbeitungsdauer eines Auftrags in einem Rechensystem genannt. Mit Hilfe der Zeiterfassung werden Angaben darüber gewonnen, mit welcher Häufigkeit die Verarbeitungsdauer in gewisse Teilintervalle (Klassen) des gesamten Wertebereichs fällt.

Tabelle 4-2 zeigt ein Beispiel mit 500 Messungen: Es treten Werte zwischen 2 und 10 Minuten auf. Der Wertebereich wird in 8 Klassen geteilt. Jede Klasse wird in der Tabelle durch den Mittelwert der Klassengrenzen gekennzeichnet. Die erste Klasse entspricht also dem Intervall [2,3[usw. Es wird angenommen, daß die Werte innerhalb einer Klasse gleichverteilt sind.

Tabelle 4-2: Empirische Verteilung

Klasse	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
Anzahl der Meßwerte	16	20	86	130	79	75	39	55
Relative Häufigkeit	0.032	0.04	0.172	0.260	0.158	0.150	0.078	0.110
Kumulative Häufigkeit	0.032	0.072	0.244	0.504	0.662	0.812	0.890	1.000

Die Abbildungen 4-6 und 4-7 zeigen die relative und die kumulative Häufigkeit (Verteilungsfunktion) dieser empirischen Verteilung. Dabei ist das Transformationsverfahren bereits zu erkennen:

$$x_i = x_j + (z_i - F(x_j)) \frac{x_{j+1} - x_j}{F(x_{j+1}) - F(x_j)} \tag{4.1-11}$$

Gemäß der Gleichverteilungsannahme der Werte innerhalb einer Histogrammklasse werden die Klassengrenzen jeweils mit Geradenstücken verbunden, d. h. es wird linear interpoliert.

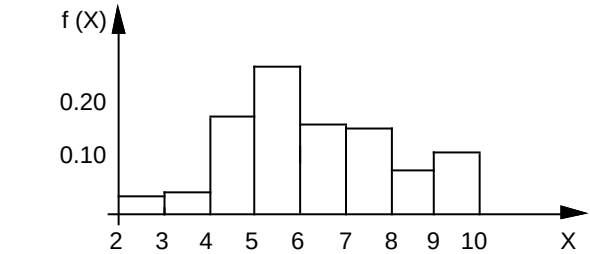


Abb. 4-6: Histogramm der empirischen Verteilung (relative Häufigkeit)

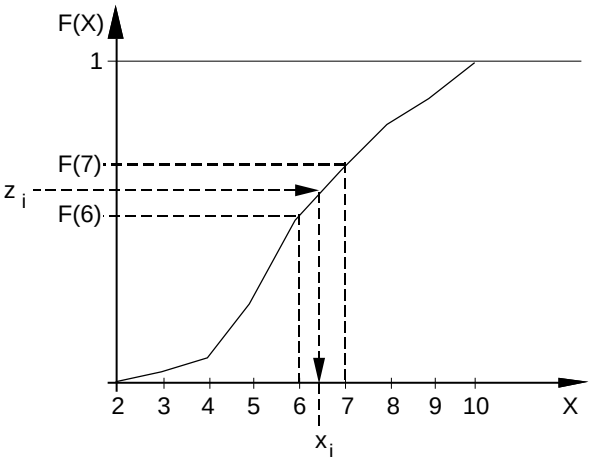


Abb. 4-7: Verteilungsfunktion der empirischen Verteilung (kumulative Häufigkeit)

Zufallszahlen diskreter Verteilungen

Eine zu simulierende Zufallsvariable (z. B. die tägliche Nachfragemenge nach einem Produkt) sei durch die Tabelle ihrer diskreten Verteilung gegeben. In Abbildung 4-8 ist die Wahrscheinlichkeitsfunktion $p_i = P(X = x_i)$, in 4-9 die Verteilungsfunktion (kumulative Häufigkeit) $F(X) = P(X \leq x_i)$ dargestellt. Um eine verteilungsgerechte Transfor-

mation gleichverteilter Zufallszahlen zu erhalten, werden diese im Verhältnis der p_i in i Intervalle auf der Ordinate unterteilt und den x_i entsprechend Tabelle 4-3 zugeordnet. Auch hier wird also die Inverse der Verteilungsfunktion zur Erzeugung von Zufallszahlen einer gegebenen Verteilung verwendet.

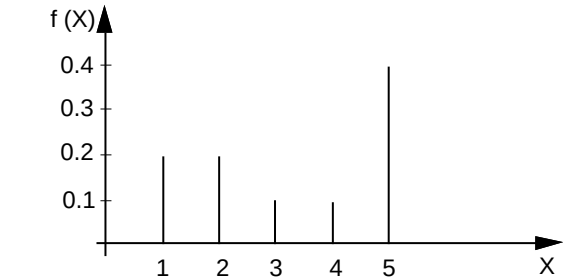


Abb. 4-8: Wahrscheinlichkeitsfunktion einer diskreten Verteilung

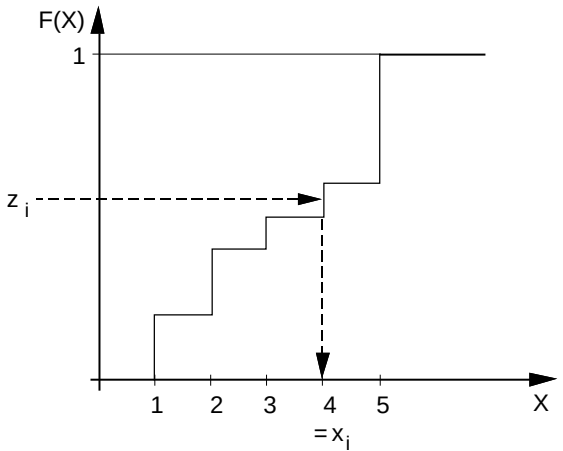


Abb. 4-9: Verteilungsfunktion der diskreten Verteilung

Tabelle 4-3: Transformation (0,1)-gleichverteilter Zufallszahlen in die diskrete Verteilung

z_i	$0 < z_i \leq 0.2$	$0.2 < z_i \leq 0.4$	$0.4 < z_i \leq 0.5$	$0.5 < z_i \leq 0.6$	$0.6 < z_i \leq 1$
x_i	1	2	3	4	5

4.2 Ein portabler Zufallszahlengenerator in Modula-2

Für die Erzeugung gleichverteilter Zufallszahlen im Intervall (0, 1) bietet sich ein spezieller Zufallszahlengenerator nach der **multiplikativen Kongruenzmethode** an, der ursprünglich von *L. Schrage* eingeführt wurde [Sch79]. *K. Marse* und *S. D. Roberts* [Mar83] haben mit Hilfe statistischer Testverfahren nachgewiesen, daß dieser Zufallszahlengenerator über ausgezeichnete statistische Eigenschaften verfügt. Darüber hinaus ist der Generator aufgrund seiner speziellen Überlaufbehandlung besonders leicht portierbar. Eine Implementation in Modula-2 wird hier vorgestellt.

Die Funktionsprozedur *Random*, die unten erläutert wird, nimmt als Eingabeparameter eine ganze Zahl – genauer: einen Zeiger auf einen *RECORD*, der eine ganzzahlige Komponente besitzt – entgegen und bestimmt daraus zunächst einen neuen ganzzahligen Zufallszahlenwert. Dieser *INTEGER*-Wert wird anschließend in eine reellwertige Zufallszahl im Intervall (0, 1) konvertiert und als Funktionswert zurückgegeben.

In der ursprünglichen Version dieses Generators in FORTRAN [Mar82] bzw. Ada [Fri85] ist nur ein einziger Zufallszahlenstrom vorgesehen. Für viele Anwendungen werden jedoch **mehrere stochastisch unabhängige Zufallszahlenströme** benötigt, etwa wenn unterschiedliche Variablen wie Bedienzeiten, Ankunftszeiten, Auftragsarten realisiert werden sollen. Deshalb wird hier ein abstrakter Datentyp "Zufallszahlenstrom" (*RandomNumberStream*) implementiert, von dem der Anwender beliebig viele Exemplare erzeugen kann (vgl. auch 3.2.6).

Der Generator arbeitet auf einem sehr großen Zahlenbereich und eignet sich für alle 32-Bit-Rechner. Er beruht auf der Formel (4.1-1), wobei die Werte $P = 2^{31} - 1$ und $A = 630\,360\,016$ gewählt wurden. Als Funktionswert wird $z_i' = z_i / P$ zurückgegeben.

Das verwendete Verfahren löst mehrere Probleme, die bei Multiplikationen und Modulo-Operationen mit großen *INTEGER*-Zahlen auftreten können. Zur Vermeidung von Überläufen ist der Algorithmus für einen 15-Bit-Multiplikator ausgelegt. Um den oben genannten Wert für *A* verwenden zu können, ohne Überläufe bei Zwischenresultaten zu erhalten, erfolgt eine Aufspaltung in die zwei Faktoren 24 112 und 26 143, die jeweils kleiner als 2^{15} sind ($24\,112 \cdot 26\,143 = 630\,360\,016$). Mit diesen beiden Faktoren wird die Berechnungsschleife zweifach durchlaufen. Bei den erforderlichen Rechenoperationen sind weitere Probleme zu beachten:

- Bei der Multiplikation von 31-Bit-*INTEGER*-Werten² mit 15-Bit-Faktoren ($A \cdot z_i$) ergeben sich maximal 46 Bit umfassende Zahlen. Um einen Überlauf zu vermeiden, wird jede Zahl z_i zunächst in der Form

$$z_i = a \cdot 2^{16} + b \quad (4.2-1)$$

² Das 32. Bit enthält das Vorzeichen.

dargestellt und das Produkt aus *A* und z_i entsprechend als

$$A \cdot z_i = A \cdot a \cdot 2^{16} + A \cdot b$$

wobei für *a* maximal 15 und für *b* maximal 16 Bit benötigt werden.

- Der Modulus *P* des Zufallszahlen-Generators nach der multiplikativen Kongruenzmethode sollte eine Primzahl sein, um die Zyklenlänge (Periodenlänge) möglichst groß zu halten. Andererseits lassen sich Modulo-Operationen einfacher mit Zweierpotenzen ($P' = 2^d$) implementieren³, da dann lediglich die höherwertigen Bits abgeschnitten werden müssen. Mit $P' = P+1$ ist

$$z_i' = (A \cdot z_i) \text{ MOD } P' = A \cdot z_i - k \cdot P' \quad (4.2-2)$$

Der ganzzahlige Wert

$$k = (A \cdot z_i) \text{ DIV } P' \quad (4.2-3)$$

stellt dann den Überlauf über die 31 Bit des Produktes dar. Durch Addition von *k* in (4.2-2) erhalten wir

$$z_i' + k = A \cdot z_i - k \cdot P \quad (4.2-4)$$

Es läßt sich zeigen, daß alle so erzeugten Werte $z_i' + k$ kleiner als $2P$ sind, wenn bestimmte Voraussetzungen über *P*, *P'* und *A* erfüllt sind.⁴ Gegebenenfalls muß *P* vom Ergebnis subtrahiert werden, damit es stets im Intervall [1, *P*-1] liegt. Somit setzt man⁵

$$z_{i+1} = \begin{cases} z_i' + k & \text{wenn } (z_i' + k) < P \\ z_i' + k - P & \text{wenn } (z_i' + k) > P \end{cases} \quad (4.2-5)$$

Für die praktische Anwendung der dargelegten Überlegungen sind folgende Schritte notwendig:

1. Zerlegung des Zufallszahlenwerts z_i gemäß (4.2-1) in $z_i = a \cdot 2^{16} + b$,
2. Berechnung von z_i' und *k* gemäß (4.2-2) und (4.2-3),
3. Berechnung von $z_{i+1} = z_i' + k$,
4. ggf. Subtraktion von *P*, um im vorgegebenen Intervall [1, *P*-1] zu bleiben,
5. Wiederholung der Schritte 2 bis 4 für den zweiten Multiplikator,
6. Abbildung der resultierenden Zufallszahl auf das reelle Intervall (0, 1).

³ In unserem Fall ist $P' = 2^{31}$.

⁴ Für den mathematisch interessierten Leser sei auf [Sch79] verwiesen.

⁵ Der Fall $z_i' + k = P$ kann nicht eintreten.

Die unten gezeigte Funktionsprozedur *Random* setzt die Definition des Typs *RandomNumberStream* und eine Prozedur zur Erzeugung neuer Zufallszahlenströme (*StreamInit*) voraus. Beides ist Bestandteil des Moduls *Distributions* (Anhang B.2.3). Die dort ebenfalls vorhandene Prozedur *Random* ist eine optimierte Version des hier vorgestellten Programms. Es werden einige Berechnungsschritte zusammengefaßt und Mehrfachauswertungen vermieden.

Die konkrete Umsetzung des geschilderten Algorithmus in Modula-2 erfordert einige zusätzliche Erläuterungen.

Anstatt direkt mit der gemäß (4.2-1) zerlegten Zufallszahl $z_i = a \cdot 2^{16} + b$ zu operieren, wird die Berechnung mit $z_i' = a + b/2^{16}$ durchgeführt. Dies gilt entsprechend für die Zwischenresultate. Die sich daraus ergebenden Divisionsreste werden bei der Rekonstruktion des Gesamtergebnisses berücksichtigt.

Im Gegensatz zum oben beschriebenen Verfahren wird der Modulus P zur Vermeidung von Überläufen in jedem Falle subtrahiert und beim Auftreten eines negativen Ergebnisses wieder addiert.

```
PROCEDURE Random (s : RandomNumberStream) : REAL;
(*-----*)
(* Portabler Zufallszahlengenerator für (0,1)-Gleichverteilung *)
(* nach der multiplikativen Kongruenz-Methode: *)
(*  $z_{i+1} = A \cdot z_i \bmod P$  *)
(* Multiplier A = 630360016, Modulus P =  $2^{31}-1 = 2147483647$  *)

CONST
  b2e15 = 32768; b2e16 = 65536; b2e24 = 16777216;
  Modulus = 2147483647; (*  $2^{31}-1$  *)

VAR
  Multiplier, HighProduct, LowProduct,
  High31, Low15, Overflow, i : INTEGER;

BEGIN
  (* Zur Vermeidung eines Überlaufes des Zwischenproduktes *)
  (* über 31 Bit hinaus muß der Multiplier kleiner als  $2^{15}$  *)
  (* sein. Daher wird die Schleife mit den beiden Faktoren *)
  (* 24112 * 26143 = 630360016 jeweils einmal durchlaufen *)

  Multiplier := 24112; (* erster Faktor *)

  FOR i := 1 TO 2 DO
    (* Zerlegung des Zufallszahlenwertes in  $a \cdot 2^{16} + b$  und ge- *)
    (* trennte Berechnung des Produktes für die oberen 15 Bit *)
    (* und die unteren 16 Bit des Zufallszahlenwertes *)
    (* ( $M \cdot a$ ) bzw. ( $M \cdot b$ ). *)
    HighProduct := (s^.value DIV b2e16) * Multiplier;
    LowProduct := (s^.value MOD b2e16) * Multiplier;

    (* Bestimmung des durch  $2^{16}$  geteilten Gesamtproduktes *)
    (* ( $M \cdot a + M \cdot b / 2^{16}$ ) aus den unteren und oberen Anteilen *)
    High31 := HighProduct + (LowProduct DIV b2e16);
```

```
(* Bestimmung des Überlaufs (k) des gesamten Produktes *)
(* über 31 Bit hinaus. Da das Produkt durch  $2^{16}$  geteilt *)
(* wurde, muss lediglich noch durch  $2^{15}$  geteilt werden. *)
Overflow := High31 DIV b2e15;

(* Rekonstruktion des Gesamtergebnisses ( $z(i+1)=z'+k$ ) *)
(* unter Berücksichtigung der bei ganzzahliger Division *)
(* aufgetretenen (gewichteten) Reste und mit vorheriger *)
(* Subtraktion des Modulus  $2^{31}-1$  zur Überlaufvermeidung *)
s^.value := (High31 MOD b2e15) * b2e16 - Modulus +
  (LowProduct MOD b2e16) + Overflow;

(* Bei negativem Wert Modulus wieder addieren *)
IF s^ < 0 THEN
  s^ := s^ + Modulus;
END;

(* Wiederholung der Berechnung für den zweiten Faktor *)
Multiplier := 26143;
END; (* FOR *)

(* Transformation in das Intervall (0,1) für 24-Bit-Mantisse *)
RETURN FLOAT (2 * (s^ DIV 256) + 1) / FLOAT (b2e24)
END Random;
```

Die Klammerung im Ausdruck für das Funktionsergebnis in der zweitletzten Zeile (*RETURN FLOAT...*) ist nur scheinbar redundant; sie dient der Überlaufvermeidung. Auf den ersten Blick mag es schwer verständlich erscheinen, weshalb die Transformation der *INTEGER*-wertigen Zufallszahlen auf die reellen Zahlen so kompliziert erfolgt. Die exakte Form dieses Ausdrucks hat jedoch entscheidenden Einfluß auf die gleichmäßige Verteilung der Ergebniswerte: Auf 32-Bit-Maschinen werden üblicherweise 24-Bit-Mantissen verwendet (*single precision*). Damit ist keine 1:1-Abbildung ohne Genauigkeitsverlust zwischen *INTEGER*- und *REAL*-Zahlen möglich. Als Ausweg bietet sich die explizite Abbildung von jeweils $2^8 = 256$ *INTEGER*-Werten auf eine einzige *REAL*-Zahl im Intervall (0, 1) an. Somit werden 2^{31} verschiedene ganze Zahlen auf 2^{23} verschiedene Gleitkommazahlen transformiert, zwischen denen jeweils eine Differenz von $1/2^{23}$ besteht.

Aus der üblichen Mantissennormalisierung mit führenden Nullen kann maschinenabhängig eine unterschiedlich dichte Abbildung der *INTEGER*- auf *REAL*-Werte am oberen und unteren Ende des (0, 1)-Intervalls resultieren, so daß dort u. U. keine Gleichverteilung gegeben ist. Um zu vermeiden, daß die ersten 255 Werte auf Null abgebildet werden, erfolgt vor der Division durch 2^{24} die Multiplikation mit 2 und anschließender Addition von 1. Durch diese Verschiebung um eine Stelle nach links und Setzen des untersten Bits wird stets eine ungerade *INTEGER*-Zahl erzeugt und weiterverarbeitet.

4.3 Testverfahren zur Bestimmung statistischer Eingabevertellungen für Simulationsmodelle

Wie die Beispiele in Kapitel 2 gezeigt haben, spielen in der Simulation Eingabedaten mit stochastischem Charakter (z. B. Zwischenankunftszeiten, Nachfragemengen, Bearbeitungszeiten) eine zentrale Rolle. Unter der Annahme, daß diese einer bestimmten statistischen Verteilung folgen, werden dann im Verlauf der Simulation entsprechend verteilte Zufallszahlen als Eingabe für das Simulationsmodell erzeugt. Dabei kann die empirische Verteilung oder aber ein **theoretischer Verteilungstyp** (Normal-, Exponential-, Erlang-Verteilung usw.) zugrundegelegt werden, der durch statistische Auswertung der empirischen Daten ermittelt wurde. Diese zweite Möglichkeit ist in der Regel vorzuziehen, da die Ergebnisse dann weniger abhängig vom konkreten Datenmaterial sind (z. B. hinsichtlich der Randwerte) und somit eine **Verallgemeinerung** der beobachteten Zufallsprozesse erreicht wurde.

Es stellt sich somit die Frage, auf welche Art der Verteilungstyp der jeweiligen Eingabevertellung bestimmt werden kann. Ein statistisches **Testverfahren** zur Bestimmung der Verteilungstypen und deren Parameter für die Eingabevertellungen eines Modells aus empirischem Datenmaterial ist der χ^2 -(*Chiquadrat*)-*Anpassungstest*, der übrigens auch als Prüfverfahren für die statistische Qualität von Zufallszahlengeneratoren (s. 4.1) eingesetzt wird.

Als Beispiel für die Durchführung eines χ^2 -Anpassungstests dient die in Tabelle 4-4 gezeigte Zufallsstichprobe der Antwortzeiten eines Rechensystems (in Sekunden) bei bestimmten Datenbankanfragen. Es ist die Annahme zu prüfen, daß diese Beobachtungswerte einer Normalverteilung folgen.

Da zwischen den beobachteten und erwarteten Häufigkeiten erwartungsgemäß Abweichungen auftreten, soll mit Hilfe des χ^2 -Anpassungstests entschieden werden, welche der beiden folgenden Hypothesen zutreffend ist:

H₀: Die Häufigkeiten der Antwortzeiten entsprechen einer Normalverteilung mit $\mu = 43.6$ und $\sigma^2 = 1.4$.

H₁: Die Häufigkeiten entsprechen nicht dieser Normalverteilung.

Um sich für eine der beiden Hypothesen entscheiden zu können, benötigt man eine **Prüfgröße**, die die Abweichung der tatsächlich beobachteten Häufigkeiten h_i in einem vorgegebenen Wertebereich (eingeteilt in äquidistante Klassen) von den theoretisch zu erwartenden Häufigkeiten e_i ausdrückt. Treten beträchtliche Unterschiede zwischen diesen Häufigkeiten auf, wird man geneigt sein, H_0 zu verwerfen. Eine häufig verwendete Prüfgröße ist:

$$\chi^2 = \sum_i \frac{(h_i - e_i)^2}{e_i} \quad (4.3-1)$$

Die Größe χ^2 folgt näherungsweise (für $n \rightarrow \infty$) selbst einer statistischen Verteilung, der von Pearson eingeführten χ^2 -Verteilung.

Die Quadrierung der Differenzen zwischen beobachteter und erwarteter Häufigkeit im Zähler verhindert, daß sich positive und negative Abweichungen aufheben. Ebenso erscheint es plausibel, diesen Ausdruck durch e_i zu normieren – die Abweichung $(1000-990)^2 = 100$ ist nicht so relevant wie $(20-10)^2 = 100$.

Die χ^2 -Verteilung besitzt einen Parameter f , die *Anzahl der Freiheitsgrade*. Die Anzahl der Freiheitsgrade entspricht der Anzahl der Intervalle (Klassen) minus 1. Sie muß jedoch weiter reduziert werden, wenn beim Vergleich der empirischen mit der theoretischen Verteilung auch Parameter der theoretischen Verteilung geschätzt werden müssen: Liegen k Intervalle vor und werden aus der Stichprobe m Parameter geschätzt, so beträgt die Anzahl der Freiheitsgrade:

$$f = k - m - 1 \quad (4.3-2)$$

Für das Beispiel ergibt sich bei Einteilung der Stichprobe in acht Intervalle ($k = 8$) und bei zwei geschätzten Parametern μ und σ die folgende Anzahl von Freiheitsgraden:

$$f = 8 - 2 - 1 = 5$$

Tabelle 4-4: Stichprobe von $n = 128$ Antwortzeiten für eine Datenbankanfrage in Sekunden

44.0	42.8	40.8	41.4	44.4	43.9	42.8	44.0	42.2	44.8
43.3	42.5	43.5	44.7	45.8	42.0	45.2	41.1	43.8	43.8
42.9	43.7	45.8	41.4	42.6	45.0	44.5	41.6	44.3	43.5
43.8	44.4	43.2	42.3	42.0	41.2	44.1	45.5	43.0	39.8
43.2	44.9	42.6	40.1	43.2	43.0	42.7	43.5	44.0	44.8
44.5	44.0	42.7	44.0	42.3	44.2	44.8	41.7	43.7	42.4
43.5	44.3	43.7	45.4	44.6	42.4	45.5	40.8	44.3	43.7
44.1	46.1	45.3	43.6	43.0	46.8	44.8	42.9	45.3	44.1
43.8	42.5	46.0	44.4	42.0	45.4	44.0	45.3	42.0	42.9
44.2	43.9	43.5	42.1	44.2	44.2	43.8	41.7	46.5	43.5
44.5	44.8	45.2	43.6	42.3	43.5	43.5	46.0	43.6	44.0
42.0	41.1	43.4	45.0	44.2	42.8	42.0	46.1	43.4	42.6
46.7	44.5	44.0	44.8	43.3	42.7	43.1	43.8		
Mittelwert $\bar{x} = 43.62$ Standardabweichung $s = 1.373$									

Bezüglich des Approximationsverhaltens der Verteilung kann i. d. R. davon ausgegangen werden, daß bei einer Stichprobe vom Umfang $n > 50$ und mit $e_i > 5$ die Approximation hinreichend gut ist.

Der Anpassungstest führt dann zur Ablehnung, wenn die Stichprobenverteilung von der vorgegebenen Verteilung zu stark abweicht, wenn also das beobachtete χ^2 einen zu großen Wert annimmt. Wenn die Nullhypothese H_0 zutrifft, ist χ^2 asymptotisch χ^2 -verteilt mit $f = 5$ Freiheitsgraden. Man bestimmt dann aus einer Tabelle der χ^2 -Verteilung⁶ den kritischen Wert, den χ^2 überschreiten muß, damit H_0 mit einer Irrtumswahrscheinlichkeit von α (Signifikanzniveau) verworfen werden kann. Der kritische Wert c ist gekennzeichnet durch die Eigenschaft:

$$P(\chi^2 > c) = \alpha$$

Wenn der Wert der Prüfgröße größer ist als c , wird H_0 verworfen und H_1 angenommen. Das Ergebnis des Tests würde in diesem Falle lauten, daß die empirischen Daten *nicht* normalverteilt sind. Wären wir an diesem Ausgang des Tests interessiert, so wäre das Signifikanzniveau α möglichst klein anzusetzen. Bei $\alpha = 0.05$ beispielsweise beträgt die Wahrscheinlichkeit, einen Wert für $\chi^2 > c$ zu erhalten, falls die vorliegende Stichprobe dennoch aus einer normalverteilten Grundgesamtheit stammt, nur 5%. So hoch wäre somit das Risiko, H_0 *irrtümlich zu verwerfen*. Man spricht bei einer Irrtumswahrscheinlichkeit von 5% von einem signifikanten Ergebnis; bei einem Signifikanzniveau von 1% nennt man das Ergebnis sehr signifikant.

Nun sind wir aber daran interessiert, H_0 beizubehalten und stattdessen H_1 abzulehnen. Es soll ja positiv festgestellt werden, daß die Normalverteilungsannahme zutrifft. Ist die Prüfgröße χ^2 kleiner oder gleich dem kritischen Wert c , so besteht nun umgekehrt ein gewisses Risiko, daß die Nullhypothese *irrtümlich beibehalten* wird. Dieses Risiko bezeichnet man in der Statistik als β -Fehler (im Gegensatz zum oben beschriebenen α -Fehler). Für praktische Zwecke ist es nicht unbedingt erforderlich, den β -Fehler genau zu bestimmen. Vielmehr genügt es, das Signifikanzniveau α nicht zu klein anzusetzen (z. B. $\alpha = 0.3$), um den β -Fehler klein zu halten.

Im obigen Beispiel geht man nun folgendermaßen vor: Für $\alpha = 0.3$ und 5 Freiheitsgrade wird der kritische Wert c aus einer Tabelle der χ^2 -Verteilung entnommen. Dieser beträgt 6.06. Bei Werten von $\chi^2 \leq 6.06$ kann also die Normalverteilungsannahme als bestätigt angesehen werden.

Nun wird der χ^2 -Wert für das obige Beispiel schrittweise ermittelt, wie in Tabelle 4-5 gezeigt. Die Intervallgrenzen (Spalte 2) der acht Intervalle werden durch Subtraktion des empirischen Mittelwerts $\bar{x}$ und Division durch s standardisiert (Spalte 3). Einer gängigen Tabelle der Flächen unter der Standard-Normalverteilung werden dann die Flächenanteile der Intervallgrenzen entnommen (Spalte 4). Aus der Differenz der beiden Flächen ergibt sich die Intervallwahrscheinlichkeit (Spalte 5). Die Multiplikation mit dem Stichprobenumfang n ergibt die erwarteten Häufigkeiten e_i innerhalb der Intervalle (Spalte 6). Diesen werden die beobachteten Intervallhäufigkeiten h_i

⁶ Tabellen der χ^2 -Verteilung findet man in Statistik-Lehrbüchern.

gegenübergestellt (Spalte 7). In Spalte 8 werden die Terme $(h_i - e_i)^2 / e_i$ in Formel (4.3-1) berechnet und zur Prüfgröße χ^2 aufsummiert.

Da der Wert $\chi^2 = 5.167$ im *Annahmebereich* liegt, wird die Hypothese einer Normalverteilung der Antwortzeiten mit den geschätzten Parametern μ und σ akzeptiert. Dieser Verteilungstyp kann folglich als Eingabeverteilung für die entsprechende Variable in einem Simulationsmodell gewählt werden.

Tabelle 4-5: Berechnung der Prüfgröße χ^2 beim χ^2 -Anpassungstest (Beispiel)

1	2	3	4	5	6	7	8
1	$-\infty \dots 42.0$	$-\infty \dots -1.18$	$0.0000 \dots 0.1190$	0.1190	15.23	12	0.685
2	$42.0 \dots 42.5$	$-1.18 \dots -0.82$	$0.1190 \dots 0.2061$	0.0871	11.15	13	0.307
3	$42.5 \dots 43.0$	$-0.82 \dots -0.45$	$0.2061 \dots 0.3264$	0.1203	15.40	14	0.127
4	$43.0 \dots 43.5$	$-0.45 \dots -0.09$	$0.3264 \dots 0.4641$	0.1377	17.63	12	1.798
5	$43.5 \dots 44.0$	$-0.09 \dots 0.28$	$0.4641 \dots 0.6103$	0.1462	18.71	23	0.984
6	$44.0 \dots 44.5$	$0.28 \dots 0.64$	$0.6103 \dots 0.7389$	0.1286	16.46	21	1.252
7	$44.5 \dots 45.0$	$0.64 \dots 1.01$	$0.7389 \dots 0.8438$	0.1049	13.43	13	0.014
8	$45.0 \dots \infty$	$1.01 \dots \infty$	$0.8438 \dots 1.0000$	0.1562	19.99	20	0.000
				1.0000	$\chi^2 = 5.167$		

Der χ^2 -Anpassungstest eignet sich für diskrete und im allgemeinen auch für stetige Verteilungen. Für stetige Verteilungen mit *kleinen* Stichprobenumfängen bietet sich dagegen der *Kolmogorov-Smirnov Anpassungstest* an, der ebenfalls in den meisten Statistik-Lehrbüchern beschrieben ist.

4.4 Die statistische Planung und Auswertung von Simulationsexperimenten

4.4.1 Zur Notwendigkeit der Wiederholung von Simulationsexperimenten

Im Zusammenhang mit stochastischen Modellen versteht man unter einem *Simulationsexperiment* den simulierten Ablauf eines oder mehrerer stochastischer Prozesse (Zufallsprozesse) unter Verwendung von Zufallszahlen. Ein *stochastischer Prozeß* ist ein in der Zeit ablaufender zufallsabhängiger Vorgang. Diesem wird zu jedem Zeitpunkt t eine Zufallsvariable X_t zugeordnet. Typisch für einen stochastischen Prozeß ist, daß die X_t voneinander nicht stochastisch unabhängig, sondern vielmehr *autokorreliert* sind (zur Autokorrelation s. 4.4.3.1).

Es sei X_t ein zu simulierender stochastischer Prozeß (z. B. die Warteschlangenlänge zum Zeitpunkt t). Wird nun der fortlaufende Prozeß zu einem vorher festgelegten Zeitpunkt t_n "angehalten", können aus den bis dahin beobachteten Werten die zu untersuchenden Leistungsgrößen berechnet werden, z. B. die mittlere Warteschlangenlänge $\bar{x}_{t_n}$. Da t_n im folgenden fest bleibt, bezeichnen wir der Einfachheit halber $\bar{x}_{t_n}$ mit $\bar{x}$. Dann gilt: **Jeder Simulationslauf liefert genau einen Wert** für $\bar{x}$. Weil das Ergebnis des Simulationslaufs vom Zufall, genauer von den verwendeten Zufallszahlen, abhängt, wird gewöhnlich auch die daraus berechnete Größe $\bar{x}$ vom Zufall abhängen. Jeder neue Simulationslauf liefert einen neuen Wert für $\bar{x}$.

Diese Überlegung zeigt, daß das Ergebnis eines einzigen Simulationslaufs fast nichts darüber aussagt, mit welchem Wert einer interessierenden Größe man tatsächlich zu rechnen hat. Eine fundierte Simulationsstudie verlangt deshalb eine ausreichende Anzahl von **unabhängigen Wiederholungen** der Simulationsexperimente. Nur auf diese Weise kann man sich einen Überblick über die Variationsbreite der möglichen Ergebnisse verschaffen.

Als Alternative zur Wiederholung von Simulationsläufen kann auch ein **einzig**, **entsprechend längerer Simulationslauf** durchgeführt werden, der dann in mehrere Abschnitte eingeteilt wird mit dem Ziel, diese wie unabhängige Experimente zu behandeln. Hier muß allerdings sichergestellt sein, daß die stochastische Unabhängigkeit tatsächlich gegeben ist. Im allgemeinen kann man dies nicht voraussetzen, da die betrachteten Abschnitte (auch *Batches* genannt) ja Teile eines einzigen, zusammenhängenden stochastischen Prozesses sind. Auf diese Problematik wird noch ausführlich eingegangen. Der wesentliche Vorteil dieser Methode gegenüber Wiederholungsläufen besteht darin, daß die statistisch nicht verwertbare Anlaufphase der Simulation nur ein einziges Mal durchlaufen werden muß.

4.4.2 Anfangszustand und Anlaufphase

Jeder Simulationslauf geht von einem *Anfangszustand* aus, der bei der Initialisierung des Modells hergestellt wird.

In der Regel gilt das Interesse einer Simulationsstudie der stationären Zustandsverteilung eines Systems. Ein stochastischer Prozeß X_t wird als *stationär* bezeichnet, wenn die Wahrscheinlichkeit von X_t nicht von t abhängig ist. Viele stochastische Prozesse *konvergieren* gegen einen stationären Prozeß. Für die Praxis erscheint es ausreichend, solche Prozesse über eine **längere Zeitspanne** zu simulieren, um sich dem stationären Zustand des Systems hinreichend anzunähern. Nach dieser *Anlaufphase* ist die Verteilung von X_t praktisch von t und damit auch vom Anfangszustand des Modells unabhängig.

Obwohl es auch Problemstellungen gibt, bei denen die transienten, d. h. die nicht stationären Phasen, untersucht werden sollen, richtet sich das Interesse in der Regel auf die stationäre Phase. Es dürfen also die Beobachtungen, die während der Anlaufphase – vom Anfangszustand bis zum Erreichen der stationären Phase – gespeichert wurden, für die statistische Auswertung des Simulationslaufs **nicht berücksichtigt werden**. Das bedeutet aber, daß bestimmt werden muß, wann die Anfangsphase beendet ist und damit der stationäre Zustand des Systems beginnt.

Überschätzt man die Länge der Anlaufphase, gehen zu viele Beobachtungen für die statistische Auswertung verloren, und die Varianz des Schätzwertes wird erhöht, was bei vorgegebener Genauigkeit zu unnötig langen Rechenzeiten führt. Unterschätzt man dagegen die Länge der Anlaufphase, so kann eine systematische Verfälschung der Ergebnisse (*bias*) die Folge sein.

Wie J. P. C. Kleijnen feststellt, existieren keine exakten Verfahren zur Bestimmung des Beginns der stationären Phase eines stochastischen Prozesses (vgl. [Kle74], S. 70). Er verweist jedoch auf eine Reihe von heuristischen Regeln.

4.4.2.1 Heuristische Verfahren zur Bestimmung der Anlaufphase

Die meisten heuristischen Regeln für die Bestimmung der Anlaufphase sind nur auf bestimmte Problemklassen zugeschnitten, häufig auf *Bedienungs-/Wartesysteme* und *Lagerhaltungssysteme*. Zwei recht einfache Regeln wurden von R. W. Conway und von K. D. Tocher vorgeschlagen.

Nach der Regel von Conway [Con63] wird von einer längeren realisierten Folge $x_1, x_2, x_3, \dots$ eines stochastischen Prozesses X_t nacheinander für jedes x_i untersucht, ob es sich um Maximum oder Minimum der Restfolge handelt. Wenn das der Fall ist, wird x_i zur Anlaufphase gerechnet und somit weggelassen. Die Untersuchung wird fortgesetzt, bis man zu einem x_i gelangt ist, das weder ein Maximum noch ein Mini-

zum der Restfolge $x_i, x_{i+1}, x_{i+2}, \dots$ darstellt. Die Restfolge ist dann mit differenzierteren Verfahren auf Stationarität zu untersuchen.

Tochers Vorschlag [Toc63] beruht auf der Erkenntnis, daß viele stochastische Prozesse *zyklisch* ablaufen. Nach Abschluß von drei bis vier solcher Zyklen könnte man das Erreichen der stationären Phase annehmen. Beispielsweise können wir bei einem Lagerhaltungssystem einen Zyklus dann als abgeschlossen ansehen, wenn ein bestimmter Lagerbestand wieder erreicht wird.

Wenn diese Verfahren durch eine **grafische Darstellung** des Werteverlaufs während der Simulation unterstützt werden, kann der Anwender nach Augenmaß – Erfahrung vorausgesetzt – den Stationaritätsbeginn ungefähr abschätzen (vgl. Abb. 4-10). Darüber hinaus könnte man auch den Fall erkennen, daß ein transienter Prozeß vorliegt und somit eine Auswertung im hier beschriebenen Sinne gar nicht zulässig ist.

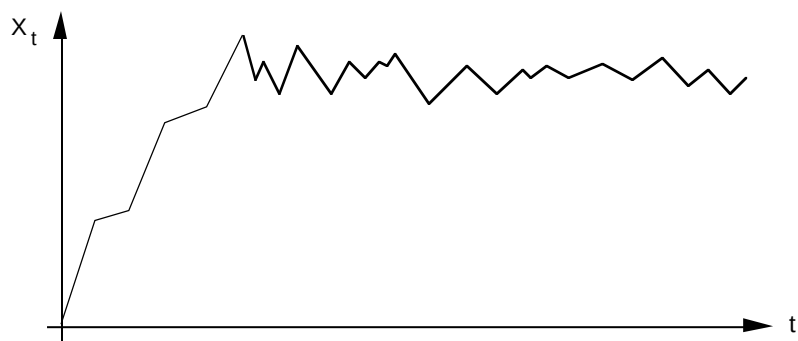


Abb. 4-10: Ein nach der Anlaufphase stationärer Prozeß

Nachteilig ist die **mangelnde Exaktheit** dieser Verfahren. Sie sind außerdem ungeeignet für Fälle, in denen ein einmal stationärer Prozeß vorübergehend wieder transient wird, wie man sich das z. B. in einer Verkehrssimulation bei Beginn der *rush hour* vorstellen kann. Nicht immer sind solche transienten Phasen so deutlich zu erkennen wie in Abbildung 4-11. Auch kann ein transientes Verhalten der **Varianz** trotz Unterstützung durch grafische Darstellungen übersehen werden (s. Abb. 4-12). Die Annahme konstanter Varianzen der Zeitreihen ist jedoch Voraussetzung für die Anwendung der hier behandelten Auswertungsmethoden.

Die heuristischen Verfahren sind dann effektiv, wenn sie eingesetzt werden, um *offensichtlich* transiente Phasen zu eliminieren. Genauere Aussagen können in einem weiteren Schritt mit entsprechenden Testverfahren gewonnen werden, die dann nur noch für die restliche Zeitreihe durchgeführt werden müssen.

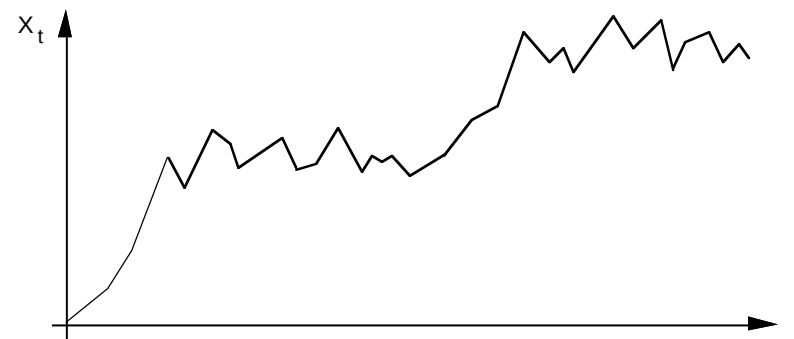


Abb. 4-11: Zeitweise transienter Prozeß

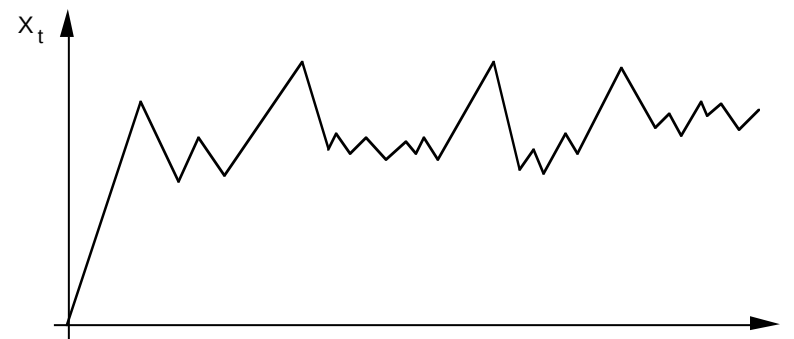


Abb. 4-12: Prozeß mit stationärem Mittelwert und transients Varianz

4.4.2.2 Statistische Testverfahren zur Bestimmung der Anlaufphase

Unabhängig davon, ob man bereits eine offensichtlich transiente Phase aus einer vorliegenden Zeitreihe eliminiert hat, kann mit statistischen Verfahren die Stationarität der Zeitreihe geprüft werden. Sollte Instationarität festgestellt werden, muß eine vom Anwender vorgegebene konstante Anzahl von Stichprobenwerten am Anfang der Sequenz "abgeschnitten" und der Test noch einmal ausgeführt werden. Nach dem *trial-and-error*-Prinzip müßte dieses Vorgehen möglicherweise sehr oft wiederholt werden, bis endlich Stationarität festgestellt würde – oder aber keine genügend große Menge von Stichproben für die Auswertung mehr vorhanden wäre. Auch mit dieser Methode kann der Beginn der stationären Phase annähernd bestimmt werden.

Voraussetzung für die Testverfahren ist die Unabhängigkeit der Stichprobenwerte, die jedoch i. a. bei den hier behandelten Zeitreihen (Simulationsergebnissen) in der Regel nicht gegeben ist. Auf welchem Wege man dennoch zu (quasi-) unabhängigen Werten gelangen kann, wird in 4.4.3 näher behandelt.

Auf der Basis dieser Unabhängigkeitsannahme lassen sich sogenannte *nicht-parametrische Testverfahren* einsetzen, um die Stationarität einer (Rest-) Folge von Stichprobenwerten zu überprüfen. Diese Testverfahren dienen der Überprüfung von nicht-parametrischen Hypothesen im Gegensatz zu den üblichen parametrischen Testverfahren, die statistische Hypothesen über die Parameter von Grundgesamtheiten (wie Mittelwert oder Varianz) untersuchen. Hier werden der *Inversionstest* und der *Run-Test* kurz dargestellt (vgl. dazu auch [Him70]).

Inversionstest

Wenn in einer Zeitreihe auf einen bestimmten Wert ein kleinerer folgt – in beliebigem Abstand –, so liegt eine sogenannte *Inversion* vor. Zum Beispiel gibt es in der Folge:

3, 5, 1, 4, 2, 1

genau zehn Inversionen ($I = 10$):

- nach 3 folgen 1, 2, 1
- nach 5 folgen 1, 4, 2, 1
- nach 4 folgen 2, 1
- nach 2 folgt 1

Bei einer Folge von n unabhängigen Stichprobenwerten (oder quasi-unabhängigen Batch-Mittelwerten, s. Abschnitt 4.4.3.3) läßt sich der theoretische Mittelwert μ der Inversionen berechnen durch:

$$\mu_I = \frac{n(n-1)}{4} \quad (4.4-1)$$

Ebenfalls läßt sich die theoretische Varianz σ^2 herleiten:

$$\sigma_I^2 = \frac{2n^3 + 3n^2 - 5n}{72} \quad (4.4-2)$$

Mit Hilfe der *Student-t-Verteilung* mit $n-1$ Freiheitsgraden als Parameter läßt sich ein Annahmebereich für die Stationaritätshypothese berechnen. Es wird also hergeleitet, in welchem Intervall sich die tatsächliche Anzahl I von Inversionen befindet, wenn die Annahme einer stationären Zeitreihe zutrifft. Zur Benutzung der t -Verteilung ist die Unabhängigkeit der Stichprobenwerte Voraussetzung. Obere und untere Grenzen OG

und UG für den **Annahmebereich** lassen sich wie folgt berechnen, wobei α für das Signifikanzniveau steht:

$$OG_{\alpha/2} = \mu_I + t_{n-1; 1-\alpha/2} \cdot \sigma_I \quad (4.4-3)$$

$$UG_{\alpha/2} = \mu_I - t_{n-1; 1-\alpha/2} \cdot \sigma_I$$

Da wir auch hier an der Annahme der Nullhypothese interessiert sind, setzen wir $\alpha = 0.2$ relativ groß an. Für die obige Zeitreihe ($n = 6$) ergeben sich Mittelwert, Varianz und Prüfgröße wie folgt:

$$\mu_I = 7.5 \quad \sigma_I^2 = 7.08 \quad t_{5; 0.9} = 1.48$$

Der t -Wert ist üblichen Tabellen für die Student'sche Verteilung aus Statistik-Lehrbüchern entnommen und dient der Berechnung der kritischen Werte für die Ablehnung der Hypothese. Die oberen und unteren kritischen Grenzen lassen sich nun wie folgt bestimmen:

$$OG_{\alpha/2 = 0.1} = 7.5 + 1.48 \sqrt{7.08} = 11.44 \approx 11$$

$$UG_{\alpha/2 = 0.1} = 7.5 - 1.48 \sqrt{7.08} = 3.56 \approx 4$$

Sollten negative Werte für die untere Grenze des Annahmebereiches auftreten, sind diese als Null (die kleinste Zahl möglicher Inversionen) zu interpretieren. Die Anzahl der Inversionen $I = 10$ liegt innerhalb dieser kritischen Grenzen, d. h. die **Hypothese der Stationarität** der Zeitreihe kann angenommen werden.

Run-Test

Zur Vorbereitung dieses Tests gehört die Bestimmung des *Medians* M . Dazu wird die ursprüngliche Meßreihe mit n Meßwerten sortiert. Der **Median** einer Stichprobe ist dann definiert durch

$$M = \begin{cases} X_{(n+1)/2} & \text{wenn } n \text{ ungerade} \\ \frac{X_{n/2} + X_{n/2+1}}{2} & \text{wenn } n \text{ gerade} \end{cases} \quad (4.4-4)$$

Der Median teilt folglich die sortierte Stichprobe in zwei gleich große Wertemengen auf.

Jeder Wert der ursprünglichen Meßreihe wird nun mit M verglichen. Falls er größer oder gleich M ist, wird ihm ein "+" zugeordnet, andernfalls ein "-". Bei der oben eingeführten Beispielfolge ist $M = 2.5$. Somit ergibt sich die Folge:

+ + - + - -

Ein *Run* ist definiert als eine ununterbrochene Folge des gleichen Zeichens (+ oder –). Als **Prüfgröße** wird die Anzahl der Runs U verwendet. Unter der Annahme, daß n gerade ist, lassen sich Erwartungswert und Varianz der Verteilung berechnen als

$$\mu_U = \frac{n}{2} + 1 \quad (4.4-5)$$

$$\sigma_U^2 = \frac{2n^2(2n^2 - n)}{n^2(n - 1)}$$

Sowohl eine zu große als auch eine zu kleine Anzahl von Runs sprechen *gegen* die Stationaritätsannahme. Es handelt sich folglich um einen zweiseitigen Test. Für die Berechnung der kritischen Grenzwerte zum Vergleich mit der Prüfgröße U kann ebenfalls die Formel (4.4-3) verwendet werden.

Deutlicher noch als beim Inversionstest wird hier die Notwendigkeit der Unabhängigkeit als Voraussetzung klar: Bei positiver Korrelation werden die Runs besonders lang, bei negativer besonders kurz.

Abschließend läßt sich festhalten: Beide genannten Testverfahren berücksichtigen alle Stichprobenwerte und können somit auch Instationaritäten nach der Anlaufphase erkennen. Keine dieser Methoden deckt jedoch eine Instationarität der Varianz bzw. Standardabweichung auf. **Stationarität der Varianz** ist aber Voraussetzung, um Konfidenzintervalle für die Simulationsergebnisse berechnen zu können. Eine gute Möglichkeit, solche Instationaritäten aufzudecken, besteht jedoch darin, die genannten Tests auf die absoluten Abweichungen der Stichprobenwerte vom Mittelwert anzuwenden.

4.4.3 Schätzgenauigkeit der Simulationsergebnisse

Viele Anwender der Simulationsmethode verzichten bei der Darstellung ihrer Ergebnisse auf die Angabe von *Konfidenzintervallen*, die eine Aussage über die Schätzgenauigkeit der Simulationsergebnisse ermöglichen. Ein Konfidenzintervall gibt an, in welchem Bereich sich der wahre Wert mit einer vorgegebenen Wahrscheinlichkeit (z.B. 95%) befindet. Modellergebnisse in Form von Stichprobenmittelwerten ohne Konfidenzintervalle sind nahezu bedeutungslos.

In den folgenden Abschnitten wird auf die Bestimmung von Konfidenzintervallen bei der Auswertung von Simulationsergebnissen näher eingegangen. Es besteht jedoch kein prinzipieller Unterschied zur der Bestimmung von Konfidenzintervallen in anderen Anwendungsgebieten der Statistik.

Die typische Problemstellung in der Simulation kann wie folgt beschrieben werden: Die Verteilungsfunktion $F(X)$ einer Zufallsvariablen X , die in einem Simulations-

experiment untersucht wird, sei unbekannt. Das eigentliche Ziel der Auswertung ist es, Aufschlüsse über die Verteilung $F(X)$ zu gewinnen. Häufig reicht es für die Fragestellung aus, nur den **Erwartungswert** μ zu schätzen. Eine Stichprobe vom Umfang n mit den Werten

$$x_1, x_2, \dots, x_n \quad (4.4-6)$$

sei das Resultat eines Simulationsexperimentes, bei dem nacheinander n Realisationen der Zufallsgröße X erzeugt wurden. $E(X)$ sei der *Erwartungswert* von X .

Nun muß auf der Grundlage der Stichprobe (4.4-6) der unbekannte Wert $E(X)$ geschätzt werden. Ein *erwartungstreuer Schätzwert* für $E(X)$ ist bekanntlich gegeben durch das Stichprobenmittel

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (4.4-7)$$

Die Notwendigkeit zur Berechnung von Konfidenzintervallen für solche Schätzwerte soll an einem Beispiel (nach [Sch82]) verdeutlicht werden:

Es wird ein Bedienung-/Wartesystem mit einer Bedienungsstelle und exponentialverteilten Zwischenankunfts- und Bedienungszeiten der Kunden untersucht. Die Warteschlangendisziplin sei FIFO, und der Quotient aus Ankunfts- und Bedienungsrate 0.85. Für den einfachen Fall dieses sog. M/M/1-Bedienungssystems ist der Erwartungswert für die mittlere Warteschlangenlänge WSL **analytisch** berechenbar. Unter den angegebenen Bedingungen gilt für den Erwartungswert $E(WSL) = 4.816$. In einem **Simulationslauf** könnte man – abhängig von den Zufallszahlen – die folgenden Ergebnisse erhalten: Für $n = 1000$ ergibt sich ein Mittelwert von $\bar{X} = 3.2$.

Man sieht, daß bei $n = 1000$ der Mittelwert um 30% vom Erwartungswert abweicht. Selbst bei $n = 20\,000$ kann der Fehler noch mehr als 10% betragen (s. Abb. 4-13).

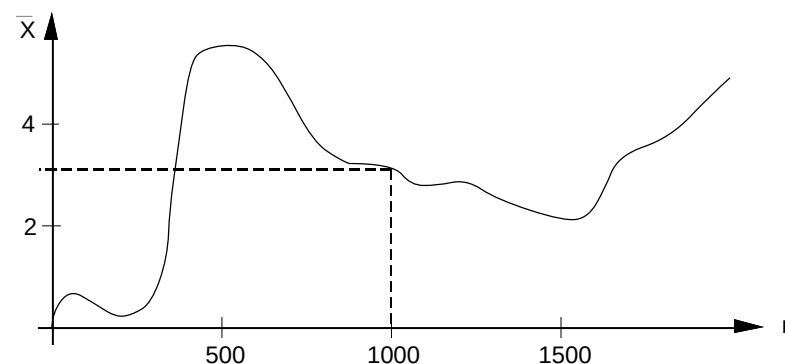


Abb. 4-13: Mittlere Warteschlangenlänge $\bar{X}$ in Abhängigkeit vom Stichprobenumfang n

Damit ist klar, daß die Ergebnisse sich nur dann sinnvoll interpretieren lassen, wenn über die Mittelwerte hinaus die zugehörigen Konfidenzintervalle angegeben werden. So darf bei $n = 20\,000$ beispielsweise das Ergebnis des Simulationslaufs nur heißen, daß der gesuchte Erwartungswert μ mit der Wahrscheinlichkeit 0.95 im (Konfidenz-) Intervall zwischen 4.56 und 6.20 liegt.

Für **unabhängige Stichprobenwerte** stellt die Berechnung von Konfidenzintervallen kein Problem dar (s. 4.4.3.2), sofern die Stichprobe hinreichend groß ist. Bei **autokorrelierten** Stichproben verkompliziert sich jedoch die Auswertung (4.4.3.3 und 4.4.3.4). Zunächst soll aber der Begriff der Autokorrelation eingeführt werden.

4.4.3.1 Autokorrelation

$x_1, x_2, \dots, x_n$ seien Glieder eines **stationären stochastischen Prozesses**. Dann ist der *Autokorrelationskoeffizient* der Ordnung j definiert als

$$\rho_j = E[(x_i - \mu)(x_{i+j} - \mu) / \sigma^2] \quad (4.4-8)$$

wobei μ der Erwartungswert der x_i und σ^2 ihre Varianz ist. Dieser Koeffizient gibt die **durchschnittliche Abhängigkeit** zweier im Abstand j aufeinanderfolgender Werte an. Er liegt allgemein zwischen -1 und $+1$. Im Falle unabhängiger – und damit nicht autokorrelierter – Stichprobenwerte sind alle

$$\rho_j = 0 \quad \text{für } j \neq 0$$

Für stationäre stochastische Prozesse ist der Autokorrelationskoeffizient allein von j und nicht vom Zeitpunkt t abhängig.

4.4.3.2 Konfidenzintervalle bei unabhängigen Stichprobenwerten

$\text{Var}(X) = \sigma^2$ sei die Varianz von X . Dann ist

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4.4-9)$$

bekanntlich Schätzwert für σ^2 . Für die Zufallsgröße

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i \quad (4.4-10)$$

gilt dann bei **unabhängigen** Stichprobenwerten $x_1, x_2, \dots, x_n$ die Beziehung

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \quad (4.4-11)$$

wofür über (4.4-9) sofort ein Schätzwert zu gewinnen ist. Bei nicht zu kleinem Stichprobenumfang ($n > 40$) ist dann ein Konfidenzintervall für $E(X)$ gegeben durch

$$\left(\bar{x} - z \frac{s}{\sqrt{n}}, \bar{x} + z \frac{s}{\sqrt{n}} \right) \quad (4.4-12)$$

wobei z abhängig vom gewählten *Konfidenzniveau* ist. Übliche Werte für z sind 1.64 (90%-iges Konfidenzniveau), 1.96 (95%-iges Konfidenzniveau) und 2.58 (99%-iges Konfidenzniveau).

Das einfachste Verfahren zur Auswertung von Simulationsergebnissen besteht in **unabhängigen Simulationsläufen** mit verschiedenen Zufallszahlen, bei denen alle Beobachtungen eines Laufes zu einem einzigen neuen Stichprobenwert x_i zusammengefaßt werden. Die x_i der einzelnen Läufe sind dann vollständig unkorreliert, und die elementare statistische Theorie zur Berechnung der Schätzgenauigkeit von $E(X)$ in (4.4-12) ist anwendbar. Der Nachteil dieses Verfahrens besteht darin, daß die Anlaufphase bei jeder Simulation erneut durchlaufen werden muß, ohne daß die Beobachtungen der Anlaufphase zur Auswertung herangezogen werden können. Besonders bei längeren Anlaufphasen wirkt sich dies nachteilig auf die erforderlichen Rechenzeiten aus.

Die in einem einzelnen Simulationslauf nacheinander beobachteten Werte einer Zufallsvariablen sind in der Regel nicht unabhängig voneinander, sondern sie sind autokorreliert, d. h. es besteht ein statistischer Zusammenhang zwischen aufeinanderfolgenden Werten. Damit ist die Berechnung von Konfidenzintervallen nach (4.4-12) unzulässig. Es gibt zwei grundsätzliche Wege, dieses Problem zu umgehen, wenn man die Ergebnisse eines einzigen (längeren) Simulationslaufs dennoch auswerten möchte. Der *erste Weg* besteht darin, aus den vorliegenden Werten neue Stichprobenwerte abzuleiten, die einem Test auf stochastische Unabhängigkeit standhalten, obwohl sie aus ursprünglich autokorrelierten Daten hervorgegangen sind. Solche Stichprobenwerte nennt man *quasi-unabhängig*. Die Berechnung der Konfidenzintervalle erfolgt dann wie oben beschrieben. Entsprechende Methoden werden im nächsten Abschnitt vorgestellt. Der *zweite Weg* beruht darauf, die bestehende Autokorrelation bei der Berechnung der Konfidenzintervalle zu berücksichtigen. Diese Berechnung muß dann allerdings nach aufwendigeren Verfahren erfolgen, die in 4.4.3.4 dargestellt werden.

4.4.3.3 Konfidenzintervalle bei quasi-unabhängigen Stichprobenwerten

Meist ist es nicht zweckmäßig, eine große Zahl von Simulationsläufen durchzuführen, um unabhängige Stichprobenwerte zu gewinnen. Besonders nachteilig ist das wiederholte Durchlaufen der Anlaufphase. Deshalb wird man es vorziehen, nur einen oder

eine geringe Anzahl von Simulationsläufen durchzuführen und diese in **Teilintervalle** zu zerlegen. Für jedes Teilintervall wird dann das Stichprobenmittel berechnet. Die Menge der Mittelwerte wird dann als neue Stichprobe betrachtet. Die Werte dieser Stichprobe sind schwächer – im Idealfall gar nicht mehr – korreliert. Der für das Intervall i berechnete Stichprobenmittelwert $\bar{x}_i$ heißt Intervallmittelwert (*batch mean*).

Es gibt mehrere gängige Verfahren, diese quasi-unabhängigen Teilintervalle zu bestimmen. Dazu gehören

- regenerative Verfahren (*independent cycles approach*)
- das Batchmittelwertverfahren (*batch means approach*)

Das von *Kleijnen* (vgl. [Kle74], S. 462) eingeführte regenerative Verfahren basiert auf der **Erneuerungseigenschaft** vieler stochastischer Prozesse, bei denen Systemzustände auftreten, die die folgenden Stichprobenwerte völlig unabhängig vom bisherigen Prozeßablauf werden lassen. Ein Beispiel dafür ist ein Bedienungs-/Wartesystem, bei dem ein unabhängiger Zyklus immer dann beginnt, wenn ein neuer Kunde in ein leeres System (Warteraum) eintritt. Wegen der Unabhängigkeit der Mittelwerte der einzelnen Zyklen ist prinzipiell die elementare statistische Theorie anwendbar. Allerdings ist eine recht komplizierte Gewichtung der Mittelwerte aufgrund der unterschiedlichen Periodenlängen erforderlich. Ein weiterer Nachteil dieses Ansatzes besteht darin, daß man am Anfang einer Simulation die Länge des Simulationslaufs nicht genau abschätzen kann. Außerdem werden sich unabhängige Zyklen nicht für alle Systeme finden lassen, denn diese Erneuerungseigenschaft kennzeichnet nur eine bestimmte Klasse von stochastischen Prozessen.

Das Batchmittelwertverfahren

Dieses Verfahren ist breiter anwendbar als die regenerativen Verfahren und wird deshalb ausführlicher besprochen. Es macht sich die Tatsache zunutze, daß die Abhängigkeit von Stichprobenwerten mit dem zeitlichen Abstand der Stichprobenwerte voneinander in aller Regel abnimmt.

Bei diesem Verfahren wird ein längerer Simulationslauf in eine Vielzahl **quasi-unabhängiger Blöcke mit fester Länge** (*batches*) aufgeteilt, deren Beobachtungen zu neuen Stichprobenwerten (Batchmittelwerten) zusammengefaßt werden, die dann nur noch wenig korreliert sind. Die elementare statistische Theorie ist wieder anwendbar, und mit (4.4-12) ist eine approximative Intervallschätzung gegeben. Eine intuitive Begründung des Verfahrens gibt *Kleijnen*:

"if the intervals are sufficiently large the averages of the subruns of intervals will effectively be uncorrelated because the effects of the correlations of early values in the interval will be averaged with the large number of values which occur later in the interval and which have no correlation with the values in the previous interval." [Kle74, S. 459]

Voraussetzung für die Anwendung des Verfahrens ist allerdings, daß die Autokorrelation mit zunehmenden Abständen zwischen den Stichprobenwerten auch tatsächlich abnimmt, was die Existenz von Periodizitäten im Modell ausschließt.

Die Hauptschwierigkeit bei dem Verfahren besteht in der **Festlegung einer angemessenen Blocklänge**. *A. M. Law* und *J. S. Carson* (vgl. [Law79] und [Law82]) haben einen Algorithmus entwickelt, um die Länge der Batches zu bestimmen. Die Darstellung dieses Algorithmus erfordert einige Vorüberlegungen.

Gegeben sei eine Meßreihe $\{x_i\}$. Ein Schätzwert für die Autokorrelation erster Ordnung (auch als *Lag-1*-Korrelation bezeichnet) ist dann gegeben durch :

$$r_1 = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{(x_i - \bar{x})(x_{i+1} - \bar{x})}{s^2} \quad (4.4-13)$$

Mittelwert und Varianz lassen sich mit $\bar{x}$ und s^2 nach den bekannten Formeln (4.4-10) und (4.4-9) ermitteln. Zur Berechnung der Autokorrelation bildet man nun q gleichlange Blöcke der Länge m (mit $m = n/q$).

Aus den m Stichproben eines Batches wird dann jeweils ein Mittelwert $\bar{x}_i$ ($i = 1, \dots, q$) gebildet. Damit entsteht eine Folge $\{\bar{x}_i\}$ von **Batchmittelwerten**, von der – wie bei der ursprünglichen Meßreihe – Varianz und Korrelation bestimmt werden können. Der Gesamtmittelwert ist natürlich der gleiche. Vom Anwender wird die maximale zulässige *Lag-1*-Autokorrelation, d. h. ein akzeptables Maß für die Annahme der (Quasi-) Unabhängigkeit, vorgegeben. Diese muß selbstverständlich bei der Ergebnisdarstellung mit angegeben werden, da sie ein wichtiges Kriterium zur Beurteilung der Güte der Simulationsergebnisse darstellt.

Ein häufig in der Literatur empfohlener Wert ist $r_1 = 0.05$. Man geht nun so vor, daß man zum Stichprobenumfang n eine Blockanzahl q bzw. Blocklänge m bestimmt, für die die Schätzung r_1 der *Lag-1*-Korrelation der Blockmittelwerte $\bar{x}_i$ diese vorgegebene Grenze nicht überschreitet. Sollte dabei aber die Anzahl der Batches q so klein werden, daß kein vernünftiges Konfidenzintervall mehr berechnet werden kann, empfiehlt sich eine verlängerte Wiederholung des Simulationsexperiments oder notfalls eine Erhöhung von r_1 .

Der Algorithmus nach *Law* und *Carson* beruht nun darauf, daß **drei Arten von Autokorrelationsverhalten** unterschieden werden. Diese sind in Abb. 4-14 bis 4-16 dargestellt, die die unterschiedlichen Beziehungen zwischen Autokorrelation und Blocklänge m veranschaulichen.

Das erste Verhalten der Autokorrelation ist unkritisch. Die Autokorrelation in Abbildung 4-14 fällt streng monoton gegen Null. Gibt man einen maximal zulässigen Autokorrelationskoeffizienten der Ordnung 1 vor, kann man die Blocklänge sukzessive erweitern, bis dieser Wert unterschritten wird. *Law* und *Carson* schlagen vor, die Anzahl relevanter Stichproben zur Mittelwertbestimmung immer zu verdoppeln, solange die Hypothese der Unabhängigkeit nicht bestätigt wird.

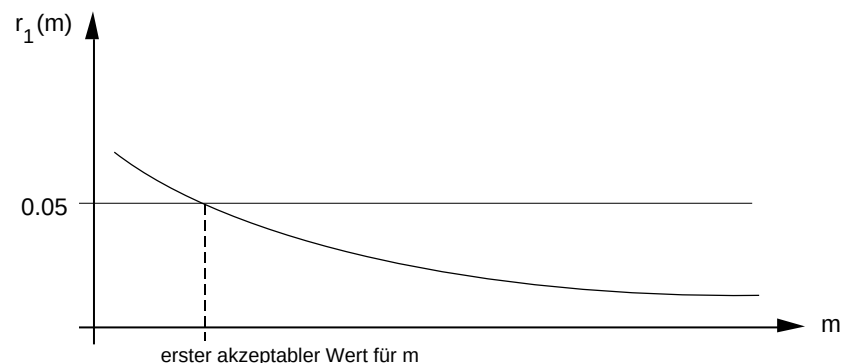


Abb. 4-14: Monotone Abnahme der Autokorrelation

Eine **negative Autokorrelation** kann ebenfalls als unkritisch angesehen werden. Wenn die Stichprobenwerte negativ korreliert sind, bedeutet das eine Oszillation um den Mittelwert der gesamten Folge. Mit anderen Worten: je größer ein bestimmter Wert ist, um so kleiner ist der folgende und umgekehrt. Dieses Verhalten impliziert eine hohe Varianz und damit ein Konfidenzintervall, das größer – und damit sicherer – ist als bei unkorrelierten Werten.

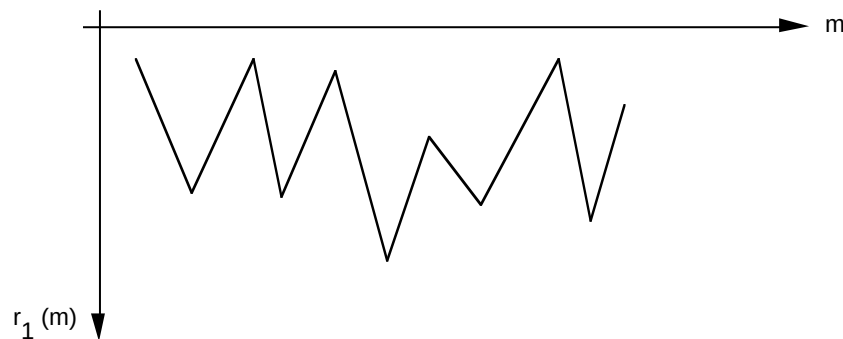


Abb. 4-15: Negative Autokorrelation

Ein **Wiederanstieg der Autokorrelation** bei Vergrößerung der Blocklänge ist als das eigentliche Problem zu betrachten. Folgt man der Vorgehensweise, die Anzahl der Batchmittelwerte durch doppelte Längen der Teilläufe zu halbieren und nach dem ersten Unterschreiten des Maximalwertes für die Autokorrelation das Verfahren abbrechen, läuft man Gefahr, versteckte Abhängigkeiten zu übersehen. Law und Carson

empfehlen deshalb, nach Unterschreiten des Maximums eine weitere Verdoppelung der Blocklänge vorzunehmen und nur dann, wenn der so ermittelte Autokorrelationskoeffizient kleiner als der vorherige ist, die Hypothese der Unabhängigkeit zu akzeptieren. Dann kann die vorletzte Intervall- bzw. Blocklänge als ausreichend lang angesehen und verwendet werden.

Das Verfahren ist auch anwendbar, um automatisch (sukzessive) die erforderliche Länge eines Simulationslaufs zu bestimmen (vgl. auch 4.4.4). Dann wird wie folgt vorgegangen: Der Benutzer des Simulationsprogramms schätzt zunächst eine gewisse Anfangsmenge n_1 von Stichprobenwerten ab, von denen er eine ausreichend genaue Berechnung eines Konfidenzintervalls erwartet. Gleichzeitig gibt er eine Blocklänge m vor, von der er annimmt, daß die resultierenden Blockmittelwerte seinen Unabhängigkeitsanforderungen genügen.

Aufgrund der Vorgaben des Benutzers erzeugt das Simulationssystem n_1 Werte und berechnet dann die Autokorrelation der n_1/m Mittelwerte. Sollte dann entweder die These der Unabhängigkeit verworfen werden oder das errechnete Konfidenzintervall eine vorgegebene Schranke des Benutzers überschreiten, werden zusätzliche Stichprobenwerte generiert und das Verfahren wird sukzessive wiederholt.

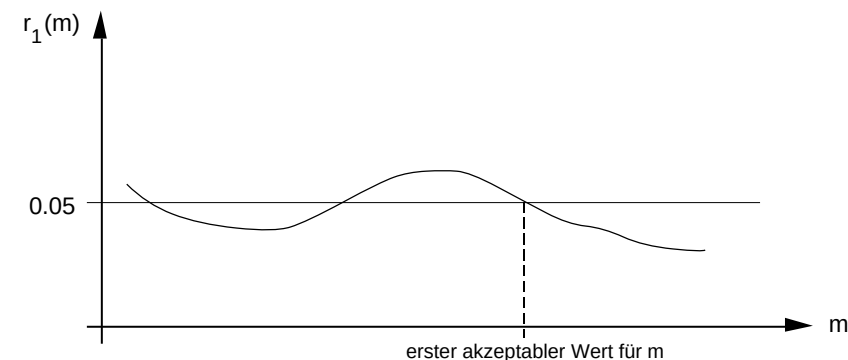


Abb. 4-16: Zeitweise wiederansteigende Autokorrelation

4.4.3.4 Konfidenzintervalle bei autokorrelierten Stichprobenwerten

Sind die Werte einer Stichprobe **autokorreliert**, was bei den Ergebniszeitreihen aus Simulationsexperimenten die Regel ist, so treten bei der Berechnung von Konfidenzintervallen einige Schwierigkeiten auf. Bei Vorliegen von Autokorrelation ist die elementare statistische Theorie nicht mehr ohne weiteres anwendbar. Würde man aus den n Stichprobenwerten $x_1, \dots, x_n$ den Mittelwert $\bar{x}$ und die Standardabweichung s berechnen, dann erhielte man bei positiver Autokorrelation aufgrund einer zu geringen

Varianzschätzung gemäß (4.4-9) mit Formel (4.4-12) ein **zu kleines Konfidenzintervall**. Damit würde eine scheinbare Genauigkeit errechnet, wie sie in Wirklichkeit keineswegs erreicht wird. Das wahre Konfidenzintervall ist größer als das auf diese Weise ermittelte. Es läßt sich zeigen, daß die Varianz bei autokorrelierten Stichproben sich berechnet zu

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \left(1 + 2 \sum_{j=1}^{n-1} \left(1 - \frac{j}{n} \right) \rho_j \right) \quad (4.4-14)$$

Diese Beziehung geht bei $\rho_j = 0$ in die bekannte Formel (4.4-11) über. Für $\rho_j > 0$ ist $\text{Var}(\bar{X})$ aber offenbar stets größer als der nach (4.4-11) berechnete Wert.

Für die Varianzschätzung im Rahmen der Berechnung eines Konfidenzintervalls für autokorrelierte Stichproben können verschiedene Verfahren wie das *direkte Einsetzen der Autokorrelationsschätzwerte* oder der sog. *Exponentialansatz* genutzt werden.

Direktes Einsetzen der Autokorrelationsschätzwerte

In Formel (4.4-14) wird die Kenntnis der σ^2 bzw. der Autokorrelationskoeffizienten ρ_j vorausgesetzt. Sie müssen jedoch erst aus der Stichprobe geschätzt werden. Für Autokorrelationen höherer Ordnung stehen nur wenige Stichprobenwerte zur Verfügung, so daß deren Schätzung sehr ungenau sein muß und man diese in (4.4-14) weglassen sollte. Dies ist an sich unkritisch, weil sie nur mit einem vergleichsweise geringen Faktor $(1-j/n)$ in die Berechnung eingehen. Um jedoch annähernd verlässliche Schätzungen für σ^2 und die ρ_j zu gewährleisten, müssen viele der Koeffizienten höherer Ordnung weggelassen werden. Das ist, ohne größere Fehler zu begehen, nur dann möglich, wenn die ρ_j mit $j \rightarrow \infty$ rasch gegen 0 streben. Unter dieser Bedingung läßt sich die Summe in (4.4-14) abkürzen und $\text{Var}(\bar{X})$ mit

$$\text{Var}(\bar{X}) = \frac{s^2}{n} \left(1 + 2 \sum_{j=1}^m \left(1 - \frac{j}{n} \right) r_j \right) \quad (4.4-15)$$

schätzen, wobei $m \ll n$ sein sollte, insbesondere so klein, daß die r_j und s^2 verlässliche Schätzungen darstellen. Die Gewichtungsfaktoren der r_j nehmen dabei zunächst linear mit der Rate $1/n$ ab, um nach $j = m$ abrupt auf Null zu fallen. Ein verbessertes Gewichtungsschema, das linear und ohne Bruch bis zum Wert 0 abfällt, erhält man, wenn man n im Quotienten durch m ersetzt (vgl. [Köc72], S. 155). Dabei soll m ungefähr 10% von n betragen.

Ein **Konfidenzintervall für $E(X)$ bei autokorrelierten Beobachtungen** ergibt sich dann zu:

$$\left(\bar{x} - z \frac{s}{\sqrt{n}} f, \bar{x} + z \frac{s}{\sqrt{n}} f \right) \quad (4.4-16)$$

Wie man sieht, ist f der Faktor, um den das Konfidenzintervall aufgrund der Autokorrelation gegenüber Formel (4.4-12) vergrößert wird. Bei dem Ansatz, der zu der Varianzschätzung in (4.4-15) führt, nimmt f den folgenden Wert an:

$$f = \sqrt{1 + 2 \sum_{j=1}^m \left(1 - \frac{j}{n} \right) r_j} \quad (4.4-17)$$

Exponentialansatz

Eine besondere Vereinfachung der Formel (4.4-15) läßt sich dann erreichen, wenn man von der Annahme ausgehen kann, daß die ρ_j in (4.4-14) **geometrisch abfallen**, so daß

$$r_j = \rho_1^j \quad \text{mit } j = 0, 1, \dots \quad (4.4-18)$$

In diesem Fall braucht nur der Autokorrelationskoeffizient **erster Ordnung** bestimmt zu werden. Formel (4.4-14) vereinfacht sich dadurch zu

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \left(1 + 2 \frac{\rho}{1 - \rho} \right) \quad (4.4-19)$$

Dieses Verfahren wird wegen der zugrundeliegenden Annahme exponentiell abfallender Autokorrelationskoeffizienten auch als *Exponentialansatz* bezeichnet [Geb74]. Es ist vor allem auf Warteschlangensysteme anwendbar, bei denen diese Annahme häufig zutrifft.

Für den Exponentialansatz mit (4.4-19) gilt bei der Berechnung des Konfidenzintervalls nach (4.4-16):

$$f = \sqrt{\frac{1 + \rho}{1 - \rho}} \quad (4.4-20)$$

4.4.3.5 Schlußbetrachtung zur Ermittlung der Schätzgenauigkeit

Die in 4.4.3.2 beschriebene Batchmittelwert-Methode veranschaulicht einen wichtigen Aspekt der statistischen Analyse von Simulationsexperimenten: die **Verwendung von Approximationen**. Wir nehmen insbesondere an, daß bei einer Verlängerung der Blöcke die (Auto-) Korrelation zwischen Paaren von Batchmittelwerten kleiner wird. Da es sich notwendigerweise um endliche Stichprobenumfänge handelt, muß man voraussetzen, daß eine endliche Anzahl von Beobachtungen je Batch so festgesetzt werden kann, daß die Abhängigkeit der Folge der Batchmittelwerte mit Hilfe eines statistischen Testes nicht mehr nachweisbar ist. Auf dieser Basis erfolgt die Schätzung der Varianz der Batchmittelwerte $\text{Var}(\bar{X}_n)$ zur Berechnung von Konfidenzintervallen.

Andere Methoden zur Schätzung der Varianz $\text{Var}(\bar{X}_n)$, die die Autokorrelation der Stichprobenwerte unmittelbar berücksichtigen, arbeiten ebenfalls mit Approximationen. Dies gilt für das direkte Einsetzen der Autokorrelationsschätzwerte und den Exponentialansatz, aber auch für andere, mathematisch komplexere Methoden wie den autoregressiven Ansatz (vgl. [Sch82], S. 122 und [Fis78], S. 247-262). Dieser betrachtet die Zeitreihe der Stichprobenwerte als autoregressiven Prozeß in Form einer Linearkombination von p Vorgängern eines Stichprobenwertes und einer Störgröße und gewinnt Varianzschätzungen über die Regressionsbeziehung.

Zur Berechnung von Konfidenzintervallen empfiehlt sich in der Praxis insbesondere die **Batchmittelwert-Methode**. Dagegen erscheinen unabhängige Wiederholungsläufe und regenerative Verfahren wegen der genannten Nachteile nur bedingt geeignet. Das direkte Einsetzen der Autokorrelationsschätzwerte ist mit hohem Rechenaufwand verbunden. Sollten die Voraussetzungen des Exponentialansatzes vorliegen, stellt dieser eine praktikable Alternative dar. Die autoregressive Methode hat wegen ihrer großen mathematischen Komplexität bisher keine größere Verbreitung in der Simulationspraxis gefunden.

4.4.4 Wahl des Stichprobenumfangs

Wie bei jeder statistischen Untersuchung muß auch bei einem Simulationsexperiment vorab der Stichprobenumfang n festgelegt werden. Einerseits muß n groß genug sein, um die Fehler bei der Berechnung des Konfidenzintervalls aufgrund der Schätzung von σ durch s und aufgrund der approximativen Annahme der Normalverteilung der Prüfgröße X unbedeutend werden zu lassen. Andererseits muß n auch ausreichend groß sein, daß eine als maximal vorgegebene Größe des Konfidenzintervalls nicht überschritten wird. Über eine festgelegte **maximale Breite des Konfidenzintervalls** bei gegebenem Konfidenzniveau läßt sich durch Umformung von (4.4-16) der erforderliche Stichprobenumfang n berechnen. Man findet n , indem man die halbe Länge des Konfidenzintervalls gleich L setzt:

$$L = z \frac{s}{\sqrt{n}} f$$

$$n = z^2 \frac{s^2}{L^2} f^2 \quad (4.4-21)$$

Es ist ersichtlich, daß zuerst eine Schätzung s für das unbekannte σ abzuleiten ist. Folglich muß man in **mehreren Schritten** vorgehen, um den notwendigen Stichprobenumfang eines Simulationsexperiments abzuschätzen. Im ersten Schritt werden durch

Simulation n_1 Stichprobenwerte generiert. Aus dem Ergebnis dieser Stichprobe vom Umfang n_1 wird σ geschätzt und ein Konfidenzintervall für den Mittelwert μ konstruiert. Liegt dieses bereits innerhalb der gewünschten Genauigkeitsschranken, so kann die Simulation abgebrochen werden. Andernfalls wird mit Hilfe der ermittelten Schätzung für σ der gesamte zur Erreichung der geforderten Genauigkeit notwendige Stichprobenumfang n nach (4.4-21) errechnet. Die Differenz $n_2 = n - n_1$ gibt dann die Anzahl der noch *zusätzlich* erforderlichen Beobachtungswerte (z. B. Batchmittelwerte) an. Da es sich hier bei den Schätzungen von σ und dem resultierenden n um Zufallsergebnisse handelt und die Schätzgenauigkeit nicht vorab garantiert werden kann, muß der Vorgang möglicherweise mehrfach wiederholt werden, bis die gewünschte Genauigkeit erreicht ist, also die vorgegebene Breite des Konfidenzintervalls nicht mehr überschritten wird.

Bei der Vorgabe einer realistischen Breite des Konfidenzintervalls muß ein Kompromiß zwischen einer noch akzeptablen Genauigkeit und den mit dem Stichprobenumfang anwachsenden Rechenzeiten gefunden werden. Pilotläufe mit vorläufigen Schätzungen von σ und n können für eine realistische Vorgabe der Genauigkeit hilfreich sein.

4.4.5 Varianzreduzierende Methoden

In den vorausgegangenen Abschnitten wurde häufig davon ausgegangen, daß Simulationsexperimente zu dem Zweck durchgeführt werden, den **Erwartungswert** einer oder mehrerer stochastischer Variablen X zu schätzen. Dabei wurde das arithmetische Mittel $\bar{X}$ als erwartungstreue Schätzfunktion für den unbekannten Erwartungswert herangezogen.

Eine erwartungstreue Schätzfunktion wird als umso besser angesehen, je geringer ihre Varianz ist (Effizienz der Schätzfunktion). Varianzreduzierende Methoden dienen zur **Reduzierung der Varianz bei gleichbleibendem Stichprobenumfang**:

"A Variance Reduction technique reduces the variance of the estimator by replacing the original sampling procedure by a new procedure that yields the same expected value with a smaller variance." [Kle74, S. 105]

Der zu erzielenden Varianzreduzierung sind jedoch bei Simulationsexperimenten die zusätzliche Rechenzeit und der erhöhte Programmieraufwand gegenüberzustellen. Zwei der zahlreichen Methoden zur Varianzreduzierung, die sich aufgrund ihrer Einfachheit bei der Implementation und wegen der kaum höheren Rechenzeiten besonders für die Simulationspraxis eignen, sind die Methode der antithetischen Variablen und die Verwendung gemeinsamer Zufallszahlen.

4.4.5.1 Antithetische Variablen

Es sei wieder der Erwartungswert einer stochastischen Variablen X in einer Simulationsstudie zu schätzen. Wenn $\bar{X}_j$ einen erwartungstreuen Schätzer für μ_j im j -ten Teilintervall (Batch) des Experimentes darstellt, dann ist der Schätzer für k solcher Wiederholungen

$$\bar{X}_k = \frac{1}{k} \sum_{j=1}^k \bar{X}_j \quad (4.4-22)$$

mit der Varianz

$$\text{Var}(\bar{X}_k) = \frac{1}{k^2} \sum_{j=1}^k (\text{Var}(\bar{X}_j)) + 2 \sum_{i=1, i < j}^k \text{Cov}(\bar{X}_i, \bar{X}_j) \quad (4.4-23)$$

Die **Kovarianz** Cov stellt ein Maß für den Zusammenhang zweier Zufallsvariablen dar. Durch Standardisierung mit den jeweiligen Varianzen der Zufallsvariablen erhält man den Korrelationskoeffizienten. Stammen die Zufallsvariablen aus dem gleichen stochastischen Prozeß – als Beobachtungen zu unterschiedlichen Zeitpunkten –, so spricht man von der Autokovarianz bzw. der Autokorrelation (vgl. 4.4.3.1). Bei unabhängigen Versuchen sind die Kovarianzen $\text{Cov}(\bar{X}_i, \bar{X}_j)$ (ebenso wie die Autokorrelationen) gleich 0. Wenn es jedoch gelingt, die Summe der Kovarianzen negativ zu machen, läßt sich eine kleinere Varianz gegenüber unabhängigen Versuchen erreichen. Durch Einführung einer negativen Korrelation zwischen den Ergebnissen verschiedener Wiederholungen können statistisch genauere Simulationsergebnisse erzeugt werden, als dies bei unabhängigen Versuchen möglich ist.

Eine **negative Korrelation** zwischen jeweils zwei Stichproben läßt sich mit Hilfe von sog. *antithetischen Zufallszahlen* in den Simulationsprozeß einbringen. Bei dieser Methode werden im Intervall $(0, 1)$ gleichverteilte Zufallszahlen $U_1, U_2, \dots, U_n$ zur Erzeugung von Ereignissen in einem Abschnitt des Simulationsprozesses verwendet und $1-U_1, 1-U_2, \dots, 1-U_n$ für die entsprechenden Ereignisse in einer zweiten Wiederholung. Man kann dann davon ausgehen, daß (abhängig von den Transformationen, die die Zufallszahlen während eines Simulationsprozesses durchlaufen) ein Teil der negativen Korrelation sich bis zur Simulationsausgabe fortpflanzt und damit zu einer Varianzreduzierung führt.

Die Ergebnisse der **antithetischen Wiederholungen** werden paarweise gemittelt, und man erhält eine neue Folge von Beobachtungswerten. Für die gleiche Anzahl von auszuwertenden Beobachtungen ist hierbei der doppelte Stichprobenumfang notwendig. Das arithmetische Mittel der auf Grund von antithetischen Zufallszahlen generierten Beobachtungswerte hat dann eine Varianz, die um den Faktor $1 + \rho$ (mit $\rho < 0$) kleiner ist als die Varianz einer gewöhnlichen Mittelwertschätzung (vgl. [Köc72], S. 191), wobei ρ den Korrelationskoeffizienten zwischen den zugehörigen antithetischen Beobachtungspaaren darstellt.

Der Einsatz von antithetischen Variablen in Simulationsstudien erfordert zusätzlichen Programmieraufwand und leicht erhöhte Rechenzeiten durch weitere Programmschritte. Bei einem Simulationsexperiment, bei dem ein einziger, sehr langer Simulationslauf beispielsweise mit quasi-unabhängigen Blöcken zur Auswertung herangezogen wird, müssen bei Erreichen der stationären Phase Systemzustand und Werte der Zufallszahlenströme gespeichert werden. Nach der Hälfte der Simulationszeit ist dann das System mit diesen Werten neu zu initialisieren und die Simulation mit den antithetischen Zufallszahlen fortzuführen. Es ist also nochmals die gleiche Anzahl von Blöcken zu simulieren. Am Ende der Simulation muß dann die **paarweise Zusammenfassung der antithetischen Beobachtungen** erfolgen. Diese gemittelten Werte der Beobachtungspaare bilden dann die neuen (für die Varianzbestimmung) auszuwertenden Beobachtungen. Diese haben meist eine geringere Varianz, als wenn alle Blockmittelwerte mit unabhängigen Zufallszahlen erzeugt worden wären, höchstens aber die gleiche Varianz. Da Ergebnisgrößen eines Simulationsmodells gewöhnlich nicht nur durch eine, sondern durch mehrere Zufallsvariablen bestimmt werden, ist eine merkliche Varianzreduzierung nicht garantiert. Daher ist der Einsatz von antithetischen Zufallszahlen nur dann sinnvoll, wenn man im Endergebnis eine negativ korrelierte Variable erhoffen kann. Gegebenenfalls muß dies vorab mit Hilfe von Pilotläufen getestet werden.

4.4.5.2 Gemeinsame Zufallszahlen

Häufig will man bei einer Simulationsstudie mehrere Systeme mit alternativen Operationsregeln (z. B. Bestellpolitiken in der Lagerhaltung) simulieren, um das beste System zu bestimmen. In diesem Fall ist man nicht so sehr an den absoluten Werten der Ausgabeparameter interessiert, sondern an den Differenzen. Daher erscheint es intuitiv plausibel, die verschiedenen Systeme unter **gleichen Versuchsbedingungen** zu simulieren. Das bedeutet neben gleichen Anfangsbedingungen auch gleiche Eingabeströme für jedes System.

Es wird nun die Varianz der Differenz zwischen X als Ergebniswert von System 1 und Y als Ergebniswert von System 2 betrachtet. Die allgemeine Beziehung für diese Varianz ist durch den folgenden Ausdruck gegeben:

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y) \quad (4.4-24)$$

Die Varianz kann für die geschätzte Differenz also reduziert werden, wenn der Kovarianzterm positiv wird. Eine positive Kovarianz ist dadurch zu erreichen, daß die gleichen Zufallszahlen für die Eingabeströme in beiden Systemen verwendet werden. Das kann in einem Simulationsmodell leicht dadurch bewerkstelligt werden, daß für jeden Eingabestrom die Zufallszahlengeneratoren mit den gleichen Anfangswerten initialisiert werden. Zusätzlicher Programmieraufwand und erhöhte Rechenzeiten fallen demnach nicht an. *Kleijnen* benennt jedoch den Nachteil dieser Methode:

"... the use of the same random numers implies that we do not have independent observations and this complicates the statistical analysis." [Kle79, S. 105]

Beim Vergleich von nur zwei Strategien läßt sich Unabhängigkeit noch dadurch erreichen, daß – wie oben beschrieben – nur die Differenz der Beobachtungen der beiden Systeme ausgewertet wird. Ein Vergleich von mehr als zwei Strategien läuft jedoch auf sog. simultane (d. h. gleichzeitige) Konfidenzintervalle für die verschiedenen Differenzen hinaus, die unabhängige Beobachtungen voraussetzen (s. [Kle79], S. 106).

4.4.6 Bestimmung des Optimums in einem Simulationsexperiment

Hängt das Ergebnis X eines Simulationsexperimentes von bestimmten Modellparametern (Steuer-, Entscheidungsparameter) in einem Simulationsmodell ab, dann interessiert man sich häufig für diejenige **Parameterkombination**, die ein in einem vorgegebenen Sinne **optimales Ergebnis** eintreten läßt. Ist X beispielsweise eine Verlustvariable, dann ist das Optimum erreicht, wenn der Mittelwert μ von X minimal wird. Umgekehrt bedeutet die Optimierung eines Systems bezüglich einer Gewinnvariablen X das Auffinden des maximalen Mittelwertes von X .

Die Bestimmung der optimalen Kombination der Steuerparameter in einem Simulationsmodell ist Gegenstand der Versuchsplanung (*experimental design*). Da dieses Gebiet einen breiten Raum in der statistischen Literatur einnimmt (s. z. B. [Mye71], [Mon76], [Box78] oder auch [Law82]), soll es hier nur angerissen werden.

Die experimentell veränderbaren Größen eines Modells werden in der Versuchsplanung üblicherweise als **Faktoren** (*factors*), das Ergebnis der Veränderungen als **Reaktion** (*response*) bezeichnet. Die funktionale Beziehung zwischen den Faktoren ($X_1, X_2, \dots, X_n$) und der Ergebnisvariablen

$$R = f(X_1, X_2, \dots, X_n) \quad (4.4-25)$$

heißt **Reaktionsoberfläche** (*response surface*). Die einzelnen quantitativen Werte bzw. qualitativen Ausprägungen eines Faktors sind die **Faktorstufen** (*levels*). Die Gesamtheit der Kombinationen aller Faktorstufen, die bei einem Experiment möglich sind, wird **Faktorraum** (*factor space*) genannt. Mit der Festlegung einer Faktorstufe bzw. einer Kombination von Faktorstufen wird somit ein Punkt im Faktorraum bestimmt. Eine Auswahl derartiger Punkte stellt einen **Versuchsplan** (*experimental design*) dar. Bei einem **Optimierungsexperiment** soll diejenige Faktorenkombination bestimmt werden, die zu einer optimalen Reaktion des Modells führt.

Es ist oft sinnvoll, mit dem Simulationsmodell **Erkundungsexperimente** durchzuführen, um die wichtigsten Faktoren und die relevanten Faktorstufen herauszufinden. Man bezeichnet diese Vorgehensweise als **Screening** (vgl. [Kle74], S. 77). Außerdem

können Vorinformationen über das Realsystem eingesetzt werden, um Faktorstufen auszuschließen, die von vornherein für die Praxis nicht relevant sind. Dennoch muß in der Regel eine Vielzahl von Faktorkombinationen analysiert werden, sofern ein **vollständiger Versuchsplan** mit sämtlichen Punkten des Faktorraumes verwendet wird. Nur so ist es möglich, das absolute Optimum mit Sicherheit aufzufinden. Ein vollständiger Versuchsplan ergibt bei m Faktoren mit k Stufen insgesamt

$$\prod_{i=1}^m k_i \quad (4.4-26)$$

Kombinationen. Es wird deutlich, daß man hier rasch an die Grenzen der praktischen Berechenbarkeit stößt.

Hier beginnt die eigentliche Aufgabe der Versuchsplanung, die die Reduzierung der zu untersuchenden Faktorkombinationen auf ein handhabbares Maß zum Ziel hat.

Es existiert eine ganze Reihe von experimentellen Methoden zur Suche nach der optimalen Faktorkombination, die den Rechenaufwand erheblich reduzieren können, jedoch das Auffinden eines globalen Optimums nicht garantieren. Vielmehr besteht die Gefahr, daß stattdessen ein lokales Optimum und damit eine *suboptimale Lösung* gefunden wird. Die gängigen Verfahren sind die *Methode des gleichförmigen Rasters*, die *stochastische Suche*, die *Methode des steilsten Anstiegs*, die *Einzelfaktor-Methoden* sowie die *Mutationsmethode* (Abb. 4-17; vgl. auch [Har74]).

Bei der **Methode des gleichförmigen Rasters** werden mit einem relativ groben Raster von Faktorkombinationen zuerst die meistversprechenden Regionen der Reaktionsoberfläche ermittelt, die dann mit einer Folge von immer kleineren Rastern weiter untersucht werden.

Bei der **stochastischen Suche** wird durch eine Zufallsstichprobe aus dem Faktorraum bestimmt, welche Punkte untersucht werden sollen. Auf diese Weise läßt sich dem Optimum beliebig nahe kommen, je nachdem, wie groß die Stichprobe ist. Diese Methode ist insbesondere vorteilhaft, wenn bei einer größeren Anzahl von Faktoren keinerlei Vorinformation für eine sinnvolle Einschränkung des Faktorraumes vorliegt.

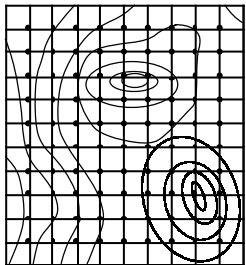
Die **Methode des steilsten Anstiegs** beginnt mit der Wahl eines geeigneten Startpunktes im Faktorraum. Für diesen Punkt und seine nähere Umgebung untersucht man den zugehörigen Ausschnitt der Reaktionsoberfläche daraufhin, in welcher Richtung er am steilsten ansteigt. In dieser Richtung wird in einem bestimmten Abstand der nächste Punkt gewählt und auf die gleiche Weise behandelt. So fortfahrend gelangt man schließlich an einen Punkt, an dem die Reaktionsoberfläche annähernd horizontal verläuft und ein weiterer Anstieg nicht mehr möglich scheint. Durch Erkundung der Umgebung stellt man dann fest, ob ein Sattelpunkt erreicht wurde, oder ob man tatsächlich am obersten Punkt angelangt ist.

Bei der **Einzelfaktor-Methode** verändert man jeweils nur einen einzigen Faktor, während die übrigen konstant gehalten werden. Dabei wird der erste Faktor so lange

variiert, bis keine Verbesserung der Zielfunktion mehr festgestellt wird. Der erste Faktor wird nun konstant gehalten, und die Veränderung des nächsten Faktors schließt sich an. Sind alle Faktoren auf diese Weise variiert worden, beginnt man erneut mit dem ersten Faktor, wobei die anderen Faktoren auf ihren günstigsten Stufen festgehalten werden. Die Suche nach dem Optimum wird abgebrochen, wenn keine weiteren Verbesserungen durch Änderung eines Faktors feststellbar sind.

Die **Mutationsmethode** schließlich beginnt mit der Wahl von zwei oder mehreren dem Startpunkt naheliegenden Punkten im Faktorraum nach dem Zufallsprinzip. Derjenige mit dem besten Zielfunktionswert wird zum neuen Ausgangspunkt gemacht, von dem aus wiederum zufällig benachbarte Faktorkombinationen ausgewählt werden. Bei einer Verschlechterung der Ergebnisse wird die Veränderung der Faktoren vom letzten Ausgangspunkt aus so lange fortgeführt, bis ein besserer Punkt gefunden ist, oder man annehmen kann, daß das Optimum erreicht wurde.

Selbst wenn man sich eines dieser Verfahren bedient, die das Auffinden eines globalen Optimums nicht garantieren, wird eine Vielzahl von Simulationsläufen benötigt. Die Simulation ist in der Tat ein sehr aufwendiges Analyseinstrument.



$\pm 1 \pm$