Sabrina Karthe

# „Anpassung von modellbasierten Verteilungsdichten an Genexpressionsdaten“

eingereicht als

DIPLOMARBEIT

an der

HOCHSCHULE MITTWEIDA
_____
UNIVERSITY OF APPLIED SCIENCES

Mathematik/Physik/Informatik

Mittweida, 2009

Erstprüfer: Prof. Dr. Egbert Lindner
Zweitprüfer: Dr. Norman Bitterlich

Vorgelegte Arbeit wurde verteidigt am: 02.02.2010

# Danksagung

An dieser Stelle möchte ich mich bei allen Personen bedanken, die mich bei der Erstellung meiner Diplomarbeit unterstützt haben.

Bei Herrn Prof. Dr. E. Lindner bedanke ich mich für die Betreuung der Arbeit seitens der Hochschule.

Besonders danken möchte ich Herrn Dr. N. Bitterlich, Medizin & Service GmbH Chemnitz, der mich durch seine engagierte Betreuung und ständige Diskussionsbereitschaft mit vielseitigen Denkanstößen bereicherte und bei der Erstellung dieser Arbeit unterstützt hat. Desweiteren danke ich Herrn Prof. Georg Hoffmann, Trillium GmbH Grafrath, für die anregenden Diskussionen sowie die Bereitstellung von Daten.

Weiterhin möchte ich meinen Eltern, meinen Geschwistern sowie meinem Freund danken, die mir durch ihre fortwährende Unterstützung das Studium und diese Arbeit ermöglichten und sie mit Anteilnahme verfolgt haben.

# Bibliographische Beschreibung

Karthe, Sabrina:
Anpassung von modellbasierten Verteilungsdichten an Genexpressionsdaten. – 2009. – 57 S.
Chemnitz, Hochschule Mittweida, Fachbereich Mathematik/Physik/Informatik, Diplomarbeit, 2009

# Kurzreferat

Ziel der Diplomarbeit ist es, ein Verfahren zur Modellanpassung für Wahrscheinlichkeitsdichtefunktionen zu entwickeln und dieses an simulierten Daten zu testen. Dabei stellt sich die Frage, ob klassische statistische Verfahren zur Prüfung von Unterschieden zwischen Dichtefunktionen ausreichen, um die Modelle zu unterscheiden. Der Kolmogorov-Smirnov-Anpassungstest liefert eine globale Bewertung der empirischen Verteilungsfunktionen. Dadurch wird schon bei kleinen Störungen der Realisierungen einer Zufallsgröße das vorliegende Modell nicht mehr als dieses erkannt. Über das Verfahren der Faltung wird eine integrale Analyse von empirischen Verteilungsfunktionen geschaffen, um somit die Robustheit gegenüber Störungen zu erhöhen und die Modellzuordnung sicherer zu gestalten. Bekannte klassische Verteilungsfunktionen werden in die Modellbetrachtungen einbezogen. Zum Schluss werden die erarbeiteten Verfahren auf reale Datensätze angewandt.

# Inhaltsverzeichnis

# Abkürzungsverzeichnis

| | |
|---|---|
| KSA | Kolmogorov-Smirnov-Anpassungstest |
| cDNA | complementary Desoxyribonukleinsäure |
| $D$ | Abbaurate $(1/min)$ |
| DNA | Desoxyribonukleinsäure |
| $f_X$ | Dichtefunktion der Zufallsgröße $X$ |
| $F_X$ | Verteilungsfunktion der Zufallsgröße $X$ |
| FG | Fließgleichgewicht |
| mRNA | messenger Ribonukleinsäure |
| $P$ | Wahrscheinlichkeitsmaß |
| $\mathbb{R}$ | Menge der reellen Zahlen |
| $S$ | Syntheserate $(1/min)$ |
| ZG | Zufallsgröße |
| ZZ | gleichverteilte Zufallszahl auf $[0; 1]$ |
| $\Sigma$ | messbare Menge |
| $\Omega$ | Menge der Elementarereignisse |

# Abbildungsverzeichnis

# Tabellenverzeichnis

# 1 Einführung

Die Simulation von Genexpressionsdaten über den Modellansatz des Fließgleichgewichts führt zu Simulationsergebnissen, die sich visuell kaum von realen Datensätzen unterscheiden. Bereits einfache Verteilungsannahmen für die Ab- und Aufbauraten ergeben realitätsnahe Datensätze.

Das Transkriptom aller bisher untersuchten Spezies weist ein rechtsschiefes Verteilungsmuster auf. Hierbei machen etwa 1% der rund 30.000 mRNA-Transkripte 20% der Gesamtkonzentration aus, 95% liegen in kaum messbaren Mengen von 0,1 bis 10 Kopien pro Zelle vor. Basierend auf Überlegungen von J. Monod et al. 1952 und Hargrove 1993 beschrieben Hoffmann et al. 2007 ein physiologisches Modell, das diese Verteilung als Folge von mRNA-Fließgleichgewichten nach der Gleichung

$$c = \frac{S}{D}$$

erklären könnte. Dabei ist $S$ die Syntheserate und $D$ die Abbaukonstante. Die Modelltypen bestehen aus zwei Funktionsästen und die Häufigkeit der Daten im oberen Zahlenbereich ist viel geringer als die im unteren Zahlenbereich (Abbildung 1). Weitere Erläuterungen zu diesem Modell werden in Kapitel 3 beschrieben.

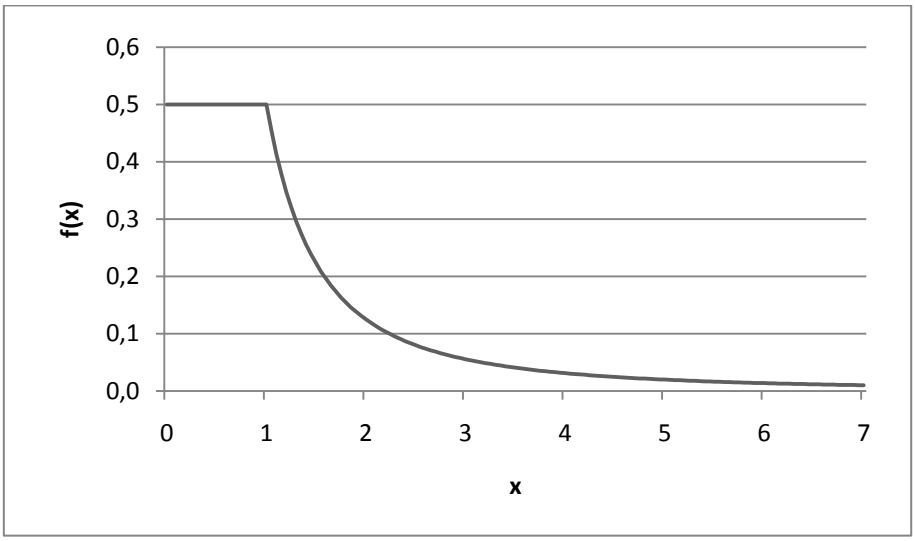


**Abbildung 1: Dichtefunktion $f(x)$ des Fließgleichgewichts**

Mithilfe statistischer Verfahren soll nun gezeigt werden, dass die Verteilung des Fließgleichgewichts am besten auf diese Daten passt. Neben dieser Verteilungsfunktion werden andere Verteilungen untersucht, die dem Fließgleichgewicht konkurrieren, unter anderem Lognormalverteilung, Weibullverteilung sowie Gammaverteilung.

Um ein statistisches Verfahren herzuleiten werden Gleich- und Normalverteilung betrachtet, da beide Verteilungsfunktionen theoretisch und praktisch gut handhabbar sind. Vergleiche dieser beiden Verteilungen mit dem Fließgleichgewicht sind nicht relevant, da die Gleichverteilung die etwas flachere rechte Flanke der Dichtefunktion nicht wiedergeben kann. Die Normalverteilung dagegen scheint ungeeignet, weil sie symmetrisch ist, die Verteilung des Fließgleichgewichts dagegen rechtsschief. Die Lognormalverteilung wurde ausgewählt, da sie in natürlichen Prozessen oft vorkommt und sich scheinbar ganz gut an die Werte anpassen lässt. Jedoch bildet sie an dem Ende der Verteilung die biologischen Werte nur grob ab. Das gleiche gilt für die Gamma- und Weibullverteilung mit geeigneter Wahl der Verteilungsparameter.
Die für die statistischen Tests verwendeten Daten werden simuliert. Die Simulation von Zufallszahlen sowie die Anpassung der Verteilungsparameter wird in Kapitel 4 beschrieben. Diese simulierten Daten finden anschließend im Kolmogorov-Smirnov-Anpassungstest Anwendung. Die Ergebnisse dieser Tests sowie die anschließende Untersuchung von Störeinflüssen auf die Daten werden in Kapitel 5 und 6 ausgewertet. In dem letzten Kapitel wird ein Verfahren vorgeschlagen, um die Modellzuordnung sicherer zu gestalten und damit die Robustheit gegenüber Störungen zu erhöhen.

# 2 Statistische Einführung

## 2.1 Kolmogorov-Smirnov-Anpassungstest

Der Kolmogorov-Smirnov-Anpassungstest ist ein statistischer Test, bei dem überprüft wird, ob eine beobachtete empirische Verteilungsfunktion an eine erwartete theoretische Verteilungsfunktion angepasst werden kann. Der Test ist sowohl einseitig als auch zweiseitig durchführbar. Da im Folgenden nur der einseitige KSA-Test benötigt wird, wird nicht weiter auf den zweiseitigen Test eingegangen. Der Test gehört zu den nichtparametrischen Tests, das bedeutet, dass keine Anforderungen an die Verteilung der zugrundeliegenden Stichprobenvariablen gestellt werden. Der KSA-Test findet bei kleinen Stichproben und ungruppiertem Datenmaterial Anwendung und ist hauptsächlich für stetige hypothetische Verteilungsfunktionen $F_0(x)$ geeignet. Er kann auch für diskrete Merkmale verwendet werden, jedoch wird in diesem Fall die Nullhypothese seltener abgelehnt als im stetigen Fall.
Die Idee des Tests besteht darin, die empirische Verteilungsfunktion $F_n(x)$ der realisierten Stichprobenvariablen mit einer vollspezifizierten hypothetischen Verteilungsfunktion $F_0(x)$ zu vergleichen.

Testablauf

Es wird ein statistisches Merkmal $X$ betrachtet, dessen Wahrscheinlichkeiten $F_0(x)$ in der Grundgesamtheit unbekannt sind.
Man stellt nun eine Nullhypothese $H_0$ sowie eine Alternativhypothese $H_1$ auf, wobei die Nullhypothese besagt, dass das Merkmal $X$ der Verteilungsfunktion $F_n(x)$ genügt.

$$H_0: F_n(x) = F_0(x) \quad \forall\, x \in \mathbb{R}$$

$$H_1: F_n(x) \neq F_0(x) \quad \exists\, x \in \mathbb{R}\,.$$

Von einer Zufallsgröße $X$ liegen $n$ Beobachtungen vor. Es wird die Rangwertfolge der realisierten Zufallsstichprobe ermittelt, das heißt die Realisierungen werden der Größe nach geordnet

$$x_1 \leq x_2 \leq \cdots \leq x_i \leq \cdots \leq x_n\,.$$

Anschließend wird die empirische Verteilungsfunktion $F_n(x)$ sowie die hypothetische Verteilungsfunktion $F_0(x)$ berechnet. Die empirische Verteilungsfunktion $F_n(x)$ berechnet sich für $i = 1;\ 2;\ \ldots;\ n-1$ wie folgt

$$F_n(x) = \begin{cases} 0, x < x_1 \\ \dfrac{i}{n}, x_i \leq x < x_{i+1} \\ 1, x \geq x_n \end{cases}.$$

Nach dem Hauptsatz der Statistik (Satz von Glivenko-Cantelli) strebt die empirische Verteilungsfunktion $F_n(x)$ gleichmäßig gegen die hypothetische Verteilungsfunktion $F_0(x)$. Man ermittelt nun die Prüfgröße $D_n$ des Testes, indem man die Abweichungen zwischen der empirischen und der hypothetischen Verteilungsfunktion berechnet.
Die maximale Differenz der beiden Verteilungsfunktionen an einer beliebigen Stelle entspricht der Prüfgröße

$$D_n = \sup_{x \in \mathbb{R}} |F_0(x) - F_n(x)|.$$

Hierbei sprechen sehr große Abweichungen gegen die Nullhypothese. Die Nullhypothese wird verworfen, wenn die Prüfgröße $D_n$ den Rückweisungspunkt $\Delta$ überschreitet. Der Schwellenwert $\Delta$ hängt von dem gewählten Signifikanzniveau $\alpha$ und vom Stichprobenumfang $n$ ab. Die Schwellenwerte liegen tabelliert vor und können entsprechender Literatur entnommen werden.

## 2.2 Klassische Verteilungsfunktionen

Die Verteilungsfunktion $F_X$ sowie die Dichtefunktion $f_X$ dienen zur Charakterisierung einer Zufallsgröße.

*Definition:*

Die Verteilungsfunktion einer reellen Zufallsvariablen $X: \Omega \to \mathbb{R}$ auf dem Wahrscheinlichkeitsraum $(\Omega, \Sigma, P)$ ist eine Funktion $F_X: \mathbb{R} \to \mathbb{R}$, die angibt, mit welcher Wahrscheinlichkeit die Zufallsvariable $X$ einen Wert kleiner oder gleich $x$ annimmt:

$$F_X(x) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), \quad x \in \mathbb{R}.$$

Weiterhin werden stetige und diskrete Zufallsvariablen unterschieden. Da beim KSA-Test nur die stetigen Zufallsvariablen Anwendung finden, wird im Folgenden nicht weiter auf diskrete Verteilungsfunktionen eingegangen.
Für stetige Verteilungsfunktionen gilt

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \ .$$

Jede stetige Verteilungsfunktion $F_X(x)$ muss folgende Eigenschaften besitzen:

1. $F_X$ ist monoton steigend
2. $F_X$ ist stetig
3. $\lim_{x \to -\infty} F_X(x) = 0$
4. $\lim_{x \to +\infty} F_X(x) = 1$ .

Mithilfe der Verteilungsfunktion $F_X$ einer stetigen Zufallsvariable $X$ wird die Wahrscheinlichkeit berechnet, mit der $X$ Werte zwischen zwei reellen Zahlen $a$ und $b$ annimmt

$$P(a < x \leq b) = \int_{a}^{b} f_X(x)dx = F_X(b) - F_X(a) .$$

### 2.2.1 Gleichverteilung

Die Gleichverteilung wird in eine diskrete und eine stetige Gleichverteilung unterschieden. Im Folgenden wird nur der stetige Fall betrachtet, da der diskrete Fall für diese Arbeit nicht von Bedeutung ist.

Es sei $X$ eine stetige Zufallsvariable. Man bezeichnet $X$ als gleichverteilt auf dem Intervall $[a; b]$, wenn für die Dichtefunktion (Abbildung 2)

$$f(x) = \begin{cases} 0, & x < a \\ \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & x > b \end{cases}$$

gilt.

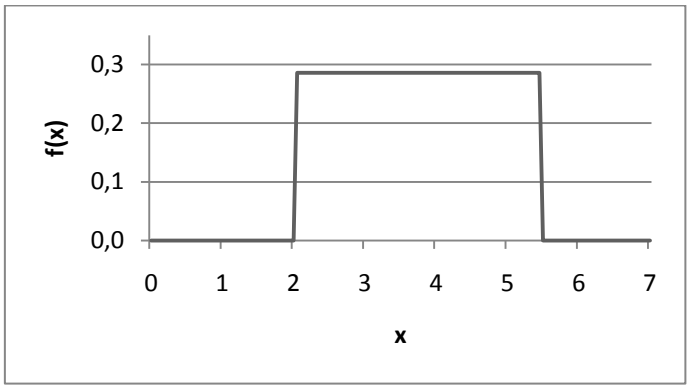


**Abbildung 2: Dichtefunktion $f(x)$ Gleichverteilung auf $[2; 5, 5]$**

Für die Verteilungsfunktion der Gleichverteilung (Abbildung 3) erhält man damit

$$F(x) = \begin{cases} 0, & x \leq a \\ \dfrac{x-a}{b-a}, & a < x < b \\ 1, & x \geq b \end{cases}.$$

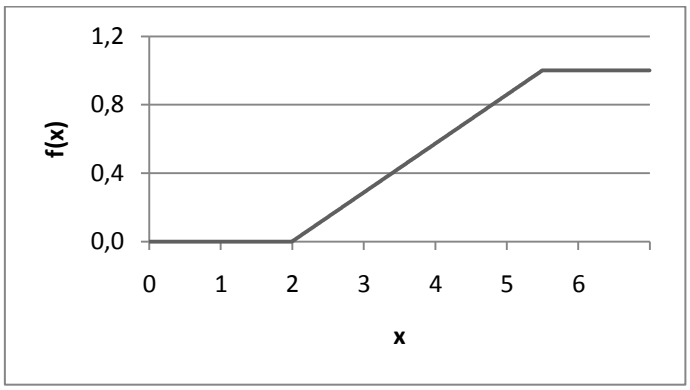


**Abbildung 3: Verteilungsfunktion $F(x)$ Gleichverteilung auf $[2; 5, 5]$**

### 2.2.2 Normalverteilung

Die Normalverteilung ist eine der wichtigsten stetigen Wahrscheinlichkeitsverteilungen. Die besondere Bedeutung beruht unter anderem auf dem zentralen Grenzwertsatz, welcher besagt, dass eine Summe von $n$ unabhängigen gleichverteilten Zufallsgrößen für $n \to \infty$ normalverteilt ist.
Die Zufallsgröße $X$ ist normalverteilt mit den Realisierungen $x \in \mathbb{R}$, falls sie folgende Dichtefunktion (Abbildung 4) besitzt

$$f(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2 \cdot \sigma^2}}$$

wobei $\mu$ der Erwartungswert und $\sigma$ die Standardabweichung ist.

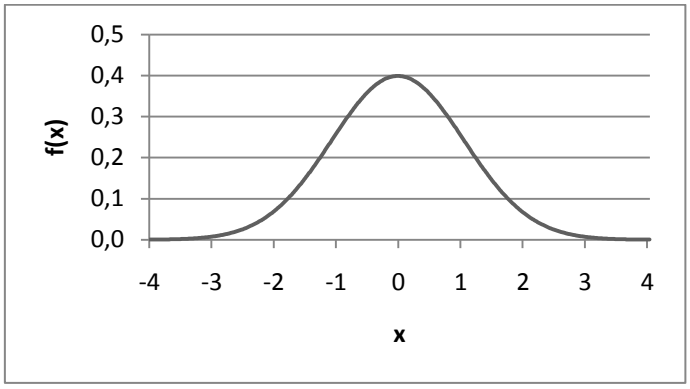


**Abbildung 4: Dichtefunktion $f(x)$ Normalverteilung mit $\mu = 0$ und $\sigma = 1$**

Die Verteilungsfunktion ist gegeben durch

$$F(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2 \cdot \sigma^2}} \, dt \, .$$

### 2.2.3 Lognormalverteilung

Die Lognormalverteilung ist eine stetige Wahrscheinlichkeitsverteilung über die Menge der positiven reellen Zahlen. Eine Zufallsgröße $X$ ist lognormalverteilt, wenn $\ln X$ normalverteilt ist.
Eine stetige Zufallsgröße $X$ ist lognormalverteilt mit den Parametern $\mu$ und $\sigma$, wenn sie folgende Wahrscheinlichkeitsdichte (Abbildung 5) besitzt

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi} \cdot \sigma x} \cdot e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

mit $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}$ und $\sigma > 0$.

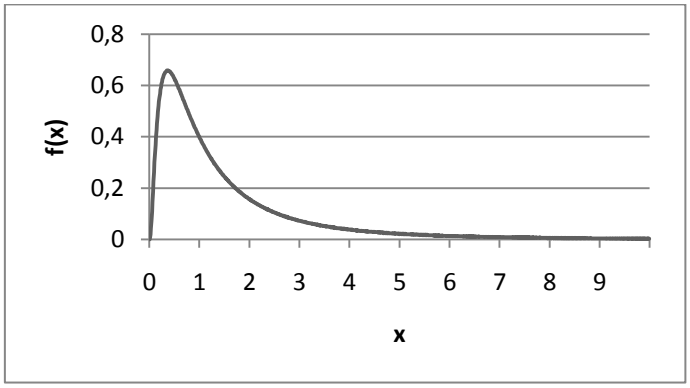


**Abbildung 5: Dichtefunktion $f(x)$ Lognormalverteilung mit $\mu = 0$ und $\sigma = 1$**

Die zugehörige Verteilungsfunktion lautet

$$F(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \int_0^x \frac{1}{t} \cdot e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}} dt \, .$$

### 2.2.4 Gammaverteilung

Eine weitere stetige Verteilung ist die Gammaverteilung. Ihr Definitionsbereich reicht über die Menge der positiven reellen Zahlen. Die Gammaverteilung ist eine Verallgemeinerung der Erlang-Verteilung für nichtganzzahlige Parameter sowie eine Verallgemeinerung für die Exponentialverteilung für $p = 1$.
Die Dichtefunktion (Abbildung 6) der Gammaverteilung $\gamma(p, b)$ ist durch

$$f(x) = \begin{cases} \frac{b^p}{\Gamma(p)} \cdot x^{p-1} \cdot e^{-bx}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

definiert, wobei $b > 0$ sowie $p > 0$ reelle Parameter sind und $\Gamma(p)$ der Funktionswert der Gammafunktion an der Stelle $p$ ist. Dabei ist die Gammafunktion definiert durch

$$\Gamma(x) = \int_0^\infty t^{x-1} \cdot e^{-t} dt \,.$$

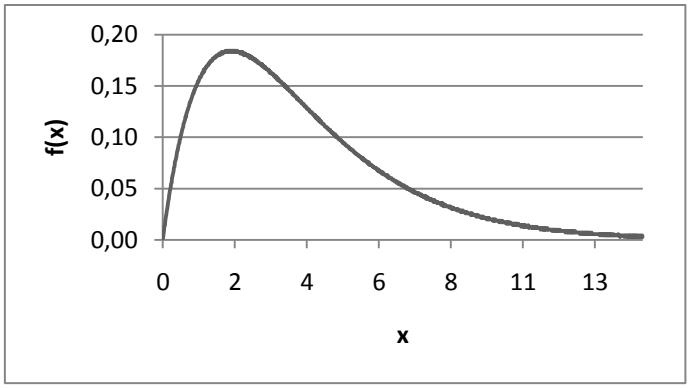


**Abbildung 6: Dichtefunktion $f(x)$ Gammaverteilung mit $b = p = 2$**

### 2.2.5 Weibullverteilung

Desweiteren wird in dieser Arbeit die Weibullverteilung als stetige Verteilung betrachtet. Sie ist neben der Exponentialverteilung die am häufigsten verwendete Lebensdauerverteilung. Die Weibullverteilung ist abhängig von den Parametern $\alpha > 0$ und $\beta > 0$. Dann besitzt sie für $x > 0$ folgende Dichtefunktion

$$f(x) = \frac{\alpha}{\beta^{\alpha}} \cdot x^{\alpha-1} \cdot e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$$

und die Verteilungsfunktion

$$F(x) = 1 - e^{-\left(\frac{x}{\beta}\right)^{\alpha}} .$$

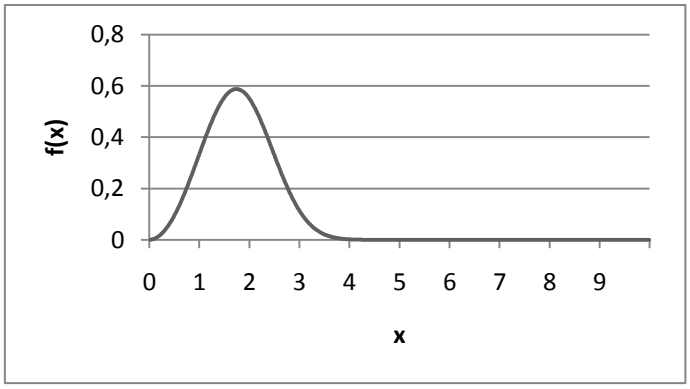


**Abbildung 7: Dichtefunktion $f(x)$ Weibullverteilung mit $\alpha = 3$ und $\beta = 2$**

# 3 Genexpression

Die Ausprägung des Genotyps zum Phänotyp einer Zelle nennt man Genexpression. Im engeren Sinne bezeichnet Genexpression die Umsetzung der genetischen Information in Proteine.
Die Lage des Gens auf dem Chromosom bestimmt seine Zugänglichkeit für die nachfolgenden Prozesse. Da die DNA gefaltet und nicht linear im Zellkern vorliegt, kann ein Genabschnitt der DNA durch die Faltung so „verdeckt“ werden, dass er für die Expression nicht zugänglich ist. Außerdem können sich dynamisch weitere Proteine an die für den Start der Transkription wichtigen und dem eigentlichen Gen vorgelagerten DNA-Abschnitte anlagern. Durch diese Transkriptionsfaktoren kann die Aktivität des Gens sowohl unterdrückt als auch verstärkt werden. Welche Aufgabe eine Zelle im Organismus übernimmt, hängt von den Genen ab, die in ihr exprimiert werden.
Die Expressionsrate eines Genes kann man zum Beispiel mit Hilfe eines Microarrays messen. Microarrays helfen dabei, mögliche Änderungen in der Aktivität bestimmter Gene in Abhängigkeit von verschiedenen Faktoren herauszufinden.
Dazu wird zunächst die mRNA aus dem zu untersuchendem Organismus extrahiert. Die Menge und Zusammensetzung der mRNA einer Zelle geben Aufschluss darüber, welche Gene wie häufig abgelesen werden.
Anschließend finden eventuelle Aufreinigungs- und Vermehrungsverfahren statt. Daraufhin wird die mRNA mit Fluoreszenzfarbstoffen markiert. Um die Genexpression von zwei verschiedenen Proben vergleichen zu können, wird die Fluoreszenz gemessen. Die Position, Intensität sowie die Farbe eines jeden Punktes auf dem Chip geben Aufschluss über die Expressionsrate eines einzelnen Genes.

## 3.1 Fließgleichgewicht

Die Signale von Microarray-Experimenten aller bisher untersuchten mRNA-Spezies in einem Zell- oder Gewebetyp weisen ein rechtsschiefes Verteilungsmuster auf. Die Häufigkeit der Daten im oberen Zahlenbereich ist viel geringer als die Häufigkeit der Daten im unteren Zahlenbereich.
Der Vergleich empirischer Verteilungsfunktionen realer Microarray-Daten zeigt, dass eine adäquate Modellbildung mittels der häufig in der Literatur diskutierten Lognormalverteilung nicht ausreichend gelingt. Der Ansatz über das Fließgleichgewicht liefert dagegen sowohl interpretierbare als auch mit der Realität qualitativ übereinstimmende Modelle.

Basierend auf zellphysiologischen Betrachtungen können Wahrscheinlichkeitsmodelle für Genexpressionsprofile auf der Grundlage von Synthese- und Abbauraten für mRNA generiert werden. Diese Teilprozesse lassen sich durch Differenzialgleichungen beschreiben, deren Zusammenführung die mRNA-Konzentration

$$c = \frac{S}{D}$$

ergibt, wobei $S$ die Synthesrate $(1/min)$ und $D$ die Abbaurate $(1/min)$ darstellt. Aus der thermodynamischen Betrachtung enzymatischer Reaktionen und aufgrund von empirischen Verteilungen realer Microarrayexperimente können die Synthese- und Abbauraten jeweils als Produkte von unabhängigen gleichverteilten Zufallsgrößen aufgefasst werden. Das bedeutet

$$S^{(m)} = S_1 \cdot S_2 \cdot \ldots \cdot S_m$$

und

$$D^{(n)} = D_1 \cdot D_2 \cdot \ldots \cdot D_n$$

für $m, n > 0$. Die daraus resultierenden Konzentrationen werden mit $c(m,n)$ bezeichnet. Für $m = n = 1$ erhält man für gleichverteilte Raten mit $S_1 \sim \mathrm{G}[0;a]$ und $D_1 \sim \mathrm{G}[0;b]$ mit $a, b > 0$ folgende Dichtefunktion $f_{c(1,1)}$

$$f_{c(1,1)}(z) = \begin{cases} \dfrac{b}{2 \cdot a}, & 0 < z < \dfrac{a}{b} \\ \dfrac{a}{2 \cdot b \cdot z^2}, & \dfrac{a}{b} \leq z \end{cases}.$$

Die Dichtefunktion des Fließgleichgewichts setzt sich somit aus zwei Funktionsästen mit einem stetigen Übergang an der Stelle $z = a/b$ zusammen (Abbildung 8). Unabhängig vom Verhältnis $a/b$ beträgt die Fläche unter jeden der beiden Funktionsäste jeweils 0,5.

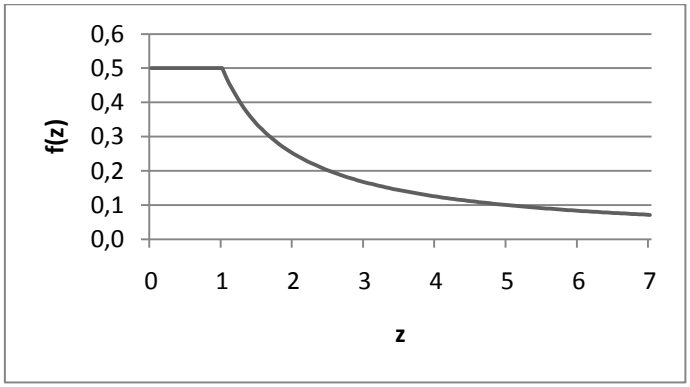


**Abbildung 8: Dichtefunktion $f_{c(1,1)}$ Fließgleichgewicht für $a = b = 1$**

Für die Verteilungsfunktion $F_{c(1,1)}(z)$ folgt für $0 < z \leq a/b$

$$F^1_{c(1,1)}(z) = \int_0^z \frac{b}{2 \cdot a}\, dt$$

$$F^1_{c(1,1)}(z) = \left[\frac{b}{2 \cdot a} \cdot t\right]_0^z$$

$$F^1_{c(1,1)}(z) = \frac{b \cdot z}{2 \cdot a}$$

und für $z \geq a/b$

$$F^2_{c(1,1)}(z) = \int_0^{\frac{a}{b}} \frac{b}{2 \cdot a}\, dt + \int_{\frac{a}{b}}^{z} \frac{a}{2 \cdot b \cdot t^2}\, dt$$

$$F^2_{c(1,1)}(z) = \left[\frac{b}{2 \cdot a} \cdot t\right]_0^{\frac{a}{b}} + \left[-\frac{a}{2 \cdot b \cdot t}\right]_{\frac{a}{b}}^{z}$$

$$F^2_{c(1,1)}(z) = 1 - \frac{a}{2 \cdot b \cdot z}.$$

Man erhält damit folgende stetige Verteilungsfunktion des Fließgleichgewichts (Abbildung 9)

$$F_{c(1,1)}(z) = \begin{cases} \frac{b \cdot z}{2 \cdot a}, & 0 < z \leq \frac{a}{b} \\ 1 - \frac{a}{2 \cdot b \cdot z}, & \frac{a}{b} \leq z \end{cases}.$$

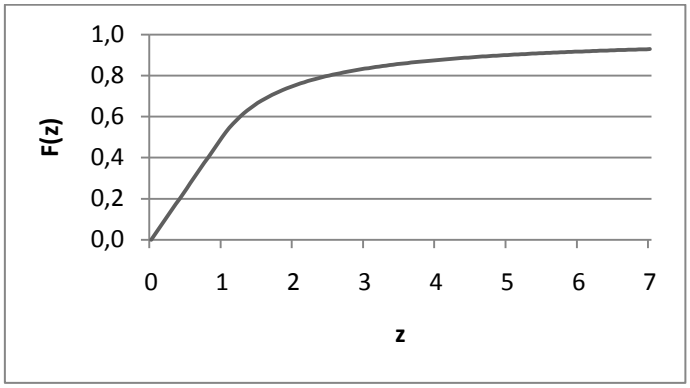


**Abbildung 9: Verteilungsfunktion $F_{c(1,1)}$ Fließgleichgewicht für $a = b = 1$**

# 4 Simulation von Zufallszahlen

## 4.1 Zufallszahlen

Mithilfe einer Simulation können verschiedene Systeme nachgebildet und analysiert werden, die für die theoretische und formelmäßige Behandlung zu kompliziert sind. Das Simulationsmodell ist eine Grundlage dafür, Erkenntnisse über das reale System zu gewinnen.

Um den KSA-Test applikationsnah anzuwenden, wurde die Simulation von Zufallszahlen angewandt. Dadurch können verschiedene Analysen durchgeführt werden, um die dadurch gewonnenen Erkenntnisse zum Schluss auf reale Daten anwenden zu können.
Für diese Arbeit wurden die Zufallszahlen mithilfe von Excel erzeugt. Die Funktion ZUFALLSZAHL() in Excel erzeugt eine Zahl zwischen 0 und 1.
Der in Excel 2003 sowie in Excel 2007 implementierte Algorithmus für die Erzeugung von Zufallszahlen wurde von B.A. Wichmann und I.D. Hill entwickelt. Die Grundidee dieses Algorithmus ist, drei Zufallszahlen auf dem Intervall $[0; 1]$ zu erzeugen, anschließend werden diese summiert. Der Bruchteil der Summe ist dann selbst eine Zufallszahl auf $[0; 1]$.
Eine Folge von Zufallszahlen wird sich jedoch irgendwann wiederholen, da die Funktion ZUFALLSZAHL() Pseudozufallszahlen erzeugt. Jedoch müssen durch die Zusammenführung von drei Zufallszahlen, wie im Algorithmus von Wichmann und Hill, mehr als $10^{13}$ Zahlen erzeugt werden, bevor eine Wiederholung auftritt.

## 4.2 Verteilungsparameter

Jede Wahrscheinlichkeitsverteilung wird durch spezifische Parameter charakterisiert. Bei der Erzeugung von Verteilungen durch Zufallszahlen sind diese Parameter bekannt. Wenn anschließend jedoch eine Störung die Verteilung beeinflusst, so ändern sich diese Parameter. Aus den simulierten Daten ist es möglich die Verteilungsparameter zu schätzen. Man kann demzufolge den Einfluss von Störgrößen prüfen sowie die Generierung der Verteilung mit den bekannten Parametern bestätigen.

### 4.2.1 Parameter der Lognormalverteilung

Um die Parameter der Lognormalverteilung zu bestimmen, muss zunächst der Mittelwert sowie die Varianz der Daten berechnet werden. Der Mittelwert ist das arithmetische Mittel einer Verteilung. Dabei werden die Daten addiert und anschließend durch die Anzahl der Daten dividiert

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \, .$$

Die Varianz gibt an, wie stark eine Zufallsgröße streut. Sie berechnet sich durch

$$\sigma^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \, .$$

Mithilfe dieser beiden Werte kann man nun zunächst die Varianz der Lognormalverteilung mit folgender Formel berechnen

$$\sigma_{LOG}^2 = \ln\left(\frac{\sigma^2}{\mu^2} + 1\right).$$

Der Erwartungswert der Lognormalverteilung lässt sich durch den Ausdruck

$$\mu_{LOG} = \ln(\mu) - \frac{\sigma_{LOG}^2}{2}$$

berechnen.

### 4.2.2 Parameter der Weibullverteilung

Die Weibullverteilung ist abhängig von den Parametern $\alpha$ und $\beta$. Hierbei gilt für $x > 0$, dass $\alpha > 0$ und $\beta > 0$. Ausgehend von der Verteilungsfunktion

$$F(x) = 1 - e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$$

werden nun folgende Umformungen durchgeführt

$$1 - F(x) = e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$$

$$\ln\big(1 - F(x)\big) = -\left(\frac{x}{\beta}\right)^{\alpha}$$

$$\ln\left(\frac{1}{1 - F(x)}\right) = \left(\frac{x}{\beta}\right)^{\alpha}$$

$$\ln\left(\ln\left(\frac{1}{1 - F(x)}\right)\right) = \alpha \cdot \ln\left(\frac{x}{\beta}\right)$$

$$\ln\left(\ln\left(\frac{1}{1 - F(x)}\right)\right) = \alpha \cdot \ln x - \alpha \cdot \ln \beta \, .$$

Vergleicht man diese Gleichung mit der Gleichung einer linearen Funktion

$$Y = mX + b$$

so stellt man fest, dass die linke Seite der Gleichung dem $Y$ in der Geradengleichung entspricht. Weiterhin ist $m = \alpha$, $X = \ln x$ und $\alpha \cdot \ln \beta = b$. Mit der Durchführung einer linearen Regression erhält man $m$ und $b$. Somit ist der Parameter $\alpha$ der Weibullverteilung gleich dem Anstieg $m$ der hergeleiteten Geraden.

Für den Parameter $\beta$ folgt aus der berechneten Gerade

$$b = -\alpha \cdot \ln \beta$$
$$-\frac{b}{\alpha} = \ln \beta$$
$$e^{-\frac{b}{\alpha}} = \beta\,.$$

### 4.2.3 Parameter der Gammaverteilung

Die Parameter $b$ und $p$ der Gammaverteilung werden mithilfe der Momentenmethode geschätzt. Mit diesem Verfahren kann man anhand einer Stichprobe die Parameter einer Verteilung schätzen. Die Parameter werden dazu in Abhängigkeit von den Momenten der Verteilung ausgedrückt.
Es wird nun wie folgt vorgegangen. Man nimmt an, dass die Grundgesamtheit gammaverteilt ist. Ausgehend von einer gegebenen Stichprobe sollen nun die Parameter $b$ und $p$ geschätzt werden. Für den Erwartungswert sowie die Varianz der Gammaverteilung gilt

$$EX = \frac{p}{b}$$

und

$$Var\,X = \frac{p}{b^2}\,.$$

Die unbekannten Parameter hängen wie folgt von den Momenten ab:

$$\Theta_1 = b = \frac{EX}{Var\,X}$$

$$\Theta_2 = p = \frac{(EX)^2}{Var\,X}\,.$$

Es sei $\bar{x}$ das arithmetische Mittel von $X$ mit

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i\,.$$

Dann erhält man mit den jeweiligen Schätzfunktionen folgende Schätzungen:

$$\hat{b} = \frac{\bar{x}}{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{p} = \frac{(\bar{x})^2}{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

# 5 Modellwahl auf Basis von Kolmogorov-Smirnov

## 5.1 Gleichverteilte Zufallszahlen

Es werden nun die im Excel simulierten Zufallszahlen auf den KSA-Test angewandt. Dabei wird auf Gleichverteilung sowie Normalverteilung getestet. Der KSA-Test wurde mit $n = 100$, $n = 1000$ sowie mit $n = 5000$ erzeugten gleichverteilten Zufallszahlen jeweils 15 mal durchgeführt. Anschließend wurde aus den 15 erhaltenen Prüfgrößen für beide Verteilungen der Mittelwert berechnet, um zufällige Fehler auszuschließen. Diese durchschnittlichen Prüfgrößen sind in Tabelle 1 für Gleichverteilung und Normalverteilung dargestellt.

| ***n*** | ***Rückweisungspunkt*** | ***Durchschnittliche Prüfgröße*** | |
|---|---|---|---|
| | | ***Gleichverteilung*** | ***Normalverteilung*** |
| 100 | 0,122 | 0,081 | 0,099 |
| 1000 | 0,039 | 0,025 | 0,068 |
| 5000 | 0,017 | 0,011 | 0,060 |

**Tabelle 1: Durchschnittliche Prüfgrößen für Gleich- und Normalverteilung des Anpassungstests bei gleichverteilten Zufallszahlen**

Aus Tabelle 1 kann man entnehmen, dass eine kleinere Prüfgröße des KSA-Tests weniger Anlass gibt, die Nullhypothese abzulehnen. Bei 100 erzeugten gleichverteilten Zufallszahlen, wird die vorliegende Verteilung durch den Test als Normal- und Gleichverteilung erkannt, obwohl gar keine Normalverteilung vorliegt. Erhöht man $n$ auf 1000 steigt die durchschnittliche Prüfgröße der Normalverteilung an und überschreitet den Rückweisungspunkt von 0,039. Die Prüfgröße der Gleichverteilung dagegen nimmt ab und bleibt kleiner als 0,039. Für $n = 5000$ ist das Testergebnis eindeutig. Die Prüfgröße der Gleichverteilung strebt immer mehr gegen Null, wogegen die der Normalverteilung sich ähnlich wie für $n = 1000$ verhält und damit größer als der Rückweisungspunkt ist.
Aus diesen Tests folgt, dass die Gleichverteilung mithilfe der in Excel erzeugten gleichverteilten Zufallszahlen umso besser dargestellt werden kann, je mehr Zufallszahlen generiert werden.

Abbildung 10 zeigt ein Beispiel für den Anpassungstest mit $n = 1000$. Es sind die hypothetischen Verteilungsfunktionen $F_0(x)$ sowie die empirische Verteilungsfunktion $F_n(x)$ dargestellt. Wie in der Tabelle erkennt man auch in diesem Diagramm, dass die erzeugten Werte sehr gut eine Gleichverteilung darstellen, da nur sehr geringe Differenzen zwischen empirischer und hypothetischer Verteilungsfunktion auftreten. Bei der Normalverteilung dagegen, sind größere Abstände zwischen empirischer und hypothetischer Verteilungs-funktion sichtbar.

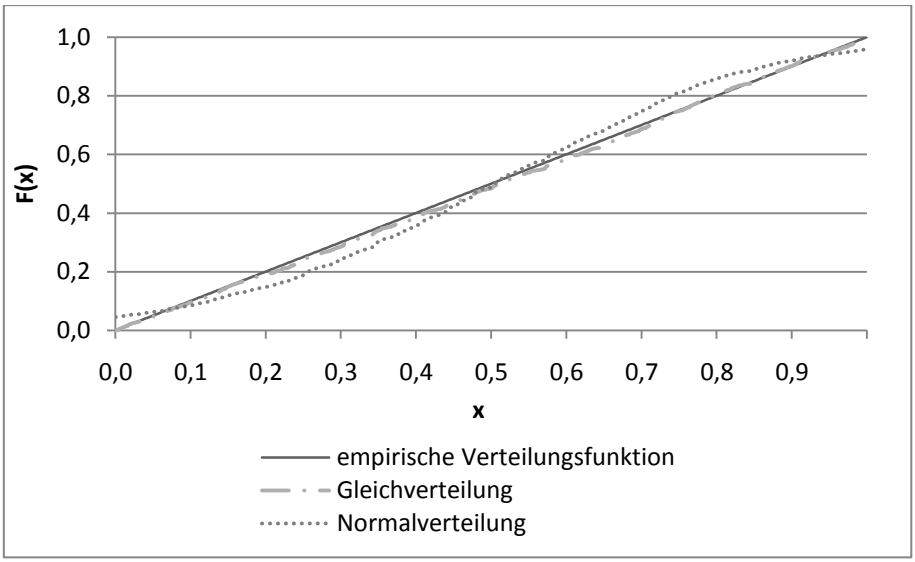


**Abbildung 10: Hypothetische und empirische Verteilungsfunktionen des Anpassungstest für gleichverteilte Zufallszahlen**

## 5.2 Summe von gleichverteilten Zufallszahlen

Es wird nun die Summe von gleichverteilten Zufallszahlen untersucht. Auch bei diesen Realisierungen wird der Test jeweils 15 mal für $n = 100$, $n = 1000$ sowie $n = 5000$ durchgeführt.

*Zentraler Grenzwertsatz nach Lindeberg und Levy:*

$X_1, X_2, X_3, \ldots$ sei eine Folge unabhängiger Zufallsvariablen, die dieselbe Verteilung mit endlichem Erwartungswert $\mu$ und endlicher Varianz $\sigma^2$ haben. Dann ist die Folge der Summen asymptotisch

$$X_n := \sum_{i=1}^{n} X_i \sim \text{Normal}\left(n\mu, \sigma\sqrt{n}\right) .$$

Demnach ist die Summe einer großen Anzahl von unabhängigen, identisch verteilten Zufallsvariablen annähernd normalverteilt.

Als Realisierungen $x_i$ werden im Folgenden die Summe von 12 gleichverteilten Zufallszahlen normiert auf das Intervall $[0; 1]$ verwendet. Nach dem zentralen Grenzwertsatz wird erwartet, dass hierbei die folgende Nullhypothese

$$H_0: \quad F_0(x) \text{ ist normalverteilt}$$

angenommen wird und die Gleichverteilung als zweite hypothetische Verteilungsfunktion abgelehnt wird.

Für diese Realisierungen erhält man die in Tabelle 2 dargestellten durchschnittlichen Prüfgrößen.

| *n* | *Rückweisungspunkt* | *Durchschnittliche Prüfgröße* | |
|---|---|---|---|
| | | *Gleichverteilung* | *Normalverteilung* |
| 100 | 0,122 | 0,337 | 0,064 |
| 1000 | 0,039 | 0,322 | 0,022 |
| 5000 | 0,017 | 0,317 | 0,009 |

**Tabelle 2: Durchschnittliche Prüfgrößen für Gleich- und Normalverteilung des Anpassungstests bei Summe von gleichverteilten Zufallszahlen**

Wie erwartet, wird die Gleichverteilung in allen drei Fällen abgelehnt, da die Prüfgröße größer ist als der Rückweisungspunkt. Die Normalverteilung wird dagegen als solche erkannt.
In Abbildung 11 sind die hypothetischen Verteilungsfunktionen und die empirische Verteilungsfunktion für den Test mit der Summe von gleichverteilten Zufallsvariablen dargestellt. Man erkennt deutlich, dass die hypothetische Verteilung der Gleichverteilung sehr große Differenzen bezüglich der empirischen Verteilungsfunktion aufweist und demnach das Modell gegen die Gleichverteilung spricht.

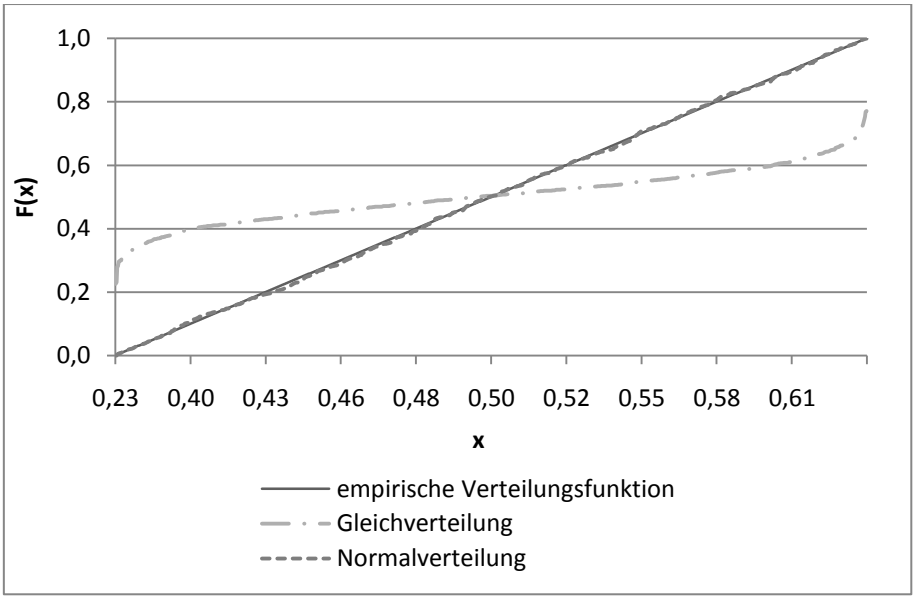


**Abbildung 11: Hypothetische und empirische Verteilungsfunktionen des Anpassungstest für die Summe von gleichverteilten Zufallszahlen**

## 5.3 Vergleich mit dem Fließgleichgewicht

Im Folgenden wird der Quotient zweier gleichverteilter Zufallszahlen untersucht. Die Ergebnisse des KSA-Tests zeigen, dass die sonst häufig in der Praxis Anwendung findenden Verteilungen, wie Gleichverteilung und Normalverteilung, nicht auf diese erzeugten Daten passen. Bei dieser Zufallsgröße wird das Fließgleichgewicht als bestes Modell erkannt. Tabelle 3 zeigt die Prüfgrößen des Anpassungstest für 1000 simulierte Daten.

| ***Rückweisungs-punkt*** | ***Prüfgröße*** | | |
|---|---|---|---|
| | ***Gleichverteilung*** | ***Normalverteilung*** | ***Fließgleichgewicht*** |
| 0,039 | 287,641 | 0,416 | 0,026 |

**Tabelle 3: Durchschnittliche Prüfgrößen für Gleich-, Normalverteilung und Fließgleichgewicht des Anpassungstests bei Quotient zweier gleichverteilter Zufallszahlen**

Folglich werden andere dem Fließgleichgewicht konkurrierende Verteilungen ausgewählt. Die in der Literatur am häufigsten diskutierte Verteilung ist die Lognormalverteilung. Weiterhin werden die Weibullverteilung sowie die Gammaverteilung zum Vergleich mit herangezogen. Alle drei Verteilungsfunktionen lassen sich ganz gut an die Werte anpassen. Führt man nun den KSA-Test mit 1000 simulierten Daten mit diesen Verteilungsfunktionen durch, erhält man die in Abbildung 12 dargestellten Ergebnisse. Der Rückweisungspunkt liegt bei 0,039. Die Zufallsgröße ist wie auch hier der Quotient zweier Zufallszahlen. Um zufällige Fehler etwas zu minimieren, wurde der Test 15 mal durchgeführt und anschließend der Mittelwert $\bar{P}$ der 15 erhaltenen Prüfgrößen $P_i$, $i = 1; \dots; 15$, berechnet

$$\bar{P} = \frac{1}{15}\sum_{i=1}^{15} P_i \, .$$

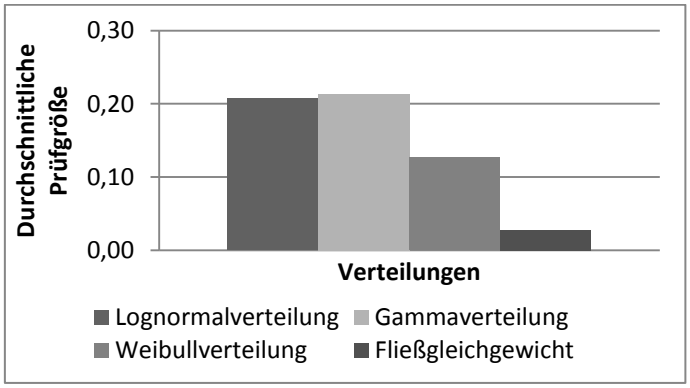


**Abbildung 12: Durchschnittliche Prüfgrößen für Lognormal-, Weibull-, Gammaverteilung und Fließgleichgewicht des Anpassungstests bei Quotient zweier gleichverteilter Zufallszahlen**

Es ist leicht ersichtlich, dass das Fließgleichgewicht das passende Modell ist, da dieses Modell als Einziges, mit einer durchschnittlichen Prüfgröße von 0,028, eine Prüfgröße hat, die unter dem Rückweisungspunkt von 0,039 liegt. Die Prüfgrößen der anderen Verteilungen sind zu hoch. Demzufolge können die drei konkurrierenden Verteilungen bei dieser Zufallsgröße ausgeschlossen werden.
In dem folgenden Abschnitt geht es darum, Störungen auf die Zufallsgröße aufzutragen, um das Verhalten der einzelnen Verteilungen bewerten zu können.

# 6 Untersuchung von Störeinflüssen

Auf den Quotient zweier gleichverteilter Zufallsgrößen werden nun verschiedene Störungen aufgetragen, um feststellen zu können, welches Modell am besten passt, welches Modell auf Ausreißer bzw. Störungen reagiert und welche Art von Störungen in der Praxis relevant sind.

Es wird nun zu dem Quotient zweier gleichverteilter Zufallszahlen eine Zufallsgröße multipliziert, die um 1 schwankt. Hierbei sei s eine Konstante, die die Werte $0{,}1;\ 0{,}2;\ \ldots;\ 1{,}0$ annehmen kann

$$ZG = \frac{ZZ}{ZZ} \cdot (1 + s \cdot ZZ)\ .$$

Der KSA-Test wird anschließend für jede Zufallsgröße 15 mal durchgeführt. Danach berechnet man den Mittelwert $\bar{P}$ um die durchschnittliche Prüfgröße zu erhalten. In Abbildung 13 sind die Ergebnisse der Tests dargestellt.

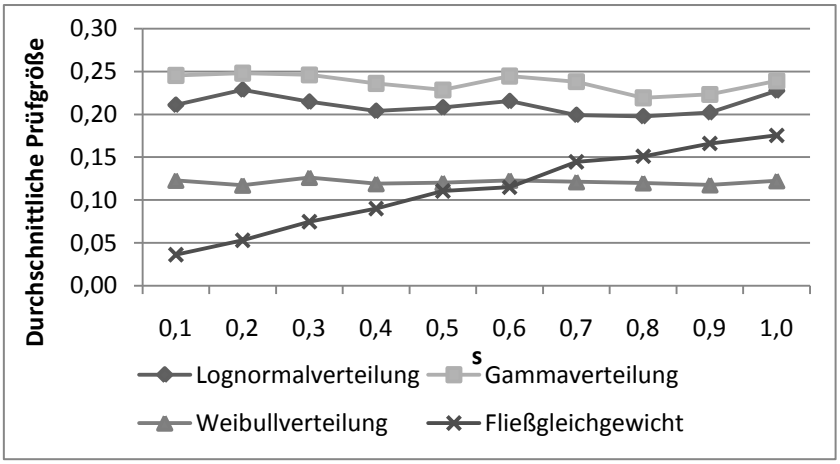


**Abbildung 13: Durchschnittliche Prüfgrößen für Lognormal-, Weibull-, Gammaverteilung und Fließgleichgewicht des Anpassungstests für $ZG = \frac{ZZ}{ZZ} \cdot (1 + s \cdot ZZ)$ mit zunehmender Störung $s$**

Die drei dem Fließgleichgewicht konkurrierenden Verteilungen stellen sich als sehr robust gegenüber Störungen heraus. Die Prüfgrößen der verschiedenen Modelle verhalten sich mit zunehmender Störung ähnlich. Am schlechtesten jedoch ist das Modell der Lognormalverteilung. Man erkennt, dass das Fließgleichgewicht lange die favorisierte Verteilung bleibt, jedoch wird das Modell der Weibullverteilung für $s = 0{,}7;\ \ldots;\ 1{,}0$ besser. Es tritt demnach der Fehler 1. Art ein, das heißt die Verteilung des Fließgleichgewichts wird für $s > 0{,}6$ nicht mehr erkannt, obwohl das Modell vorliegt.

Bei einem zweiten Störungstyp, bei dem die Zufallsgröße additiv gestört wird, erhält man ein ähnliches Ergebnis. Es wird hierbei folgende Zufallsgröße für den KSA-Test verwendet

$$ZG = \frac{ZZ}{ZZ} + (s \cdot ZZ)\,.$$

Auch hier ist $s$ eine Konstante, welche die Werte $0{,}1;\ 0{,}2;\ \ldots;1{,}0$ annimmt. Nach 15-maliger Durchführung erhält man die in Abbildung 14 dargestellten durchschnittlichen Prüfgrößen $\bar{P}$ der einzelnen Verteilungen mit zunehmender Störung.

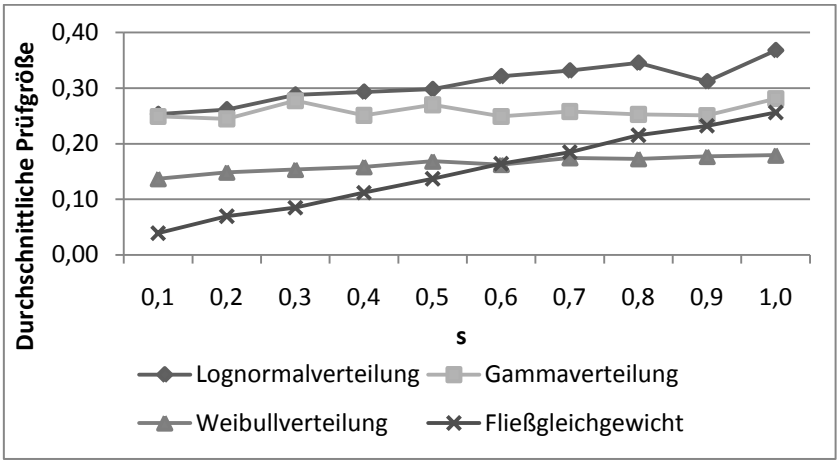


**Abbildung 14: Durchschnittliche Prüfgrößen für Lognormal-, Weibull-, Gammaverteilung und Fließgleichgewicht des Anpassungstests für $ZG = \frac{ZZ}{ZZ} + (s \cdot ZZ)$ mit zunehmender Störung $s$**

Die Modelle der Weibullverteilung sowie der Gammaverteilung sind bei diesem Störungstyp ebenfalls wieder sehr robust gegenüber den Störungen. Nur bei der Lognormalverteilung erkennt man einen leichten Anstieg der Prüfgröße $\bar{P}$. Für $s \geq 0{,}6$ wird auch hier das Modell der Weibullverteilung dem Modell des Fließgleichgewichts vorgezogen.

Bei den beiden ersten Störungstypen wurden die Parameter der jeweiligen Verteilungen während der Ausführung des KSA-Tests an die jeweilige Stichprobe angepasst. Um die einzelnen Verteilungsfunktionen jedoch noch besser vergleichen zu können, werden die Parameter nun fest gesetzt, so dass die Verteilungsfunktionen in etwa denselben Verlauf haben. Hierbei hat sich herausgestellt, dass folgende Parameter verwendet werden sollten. Für das Fließgleichgewicht wählt man für $a = 0{,}4$ und für $b = 0{,}8$. Der Mittelwert $\mu$ der Lognormalverteilung wird auf $-0{,}5$ festgesetzt und die Varianz $\sigma^2$ auf 1,0. Für die Gammaverteilung wählt man die Parameter $\alpha = 0{,}965$ sowie $\beta = 1{,}0$.
Nun werden auf den Quotient zweier gleichverteilter Zufallszahlen additiv zufällige Störungen aus dem Intervall $[-s; +s]$ mit $s = 0{,}1; 0{,}2; \ldots; 1{,}0$ aufgetragen und anschließend der KSA-Test ausgeführt (Abbildung 15).

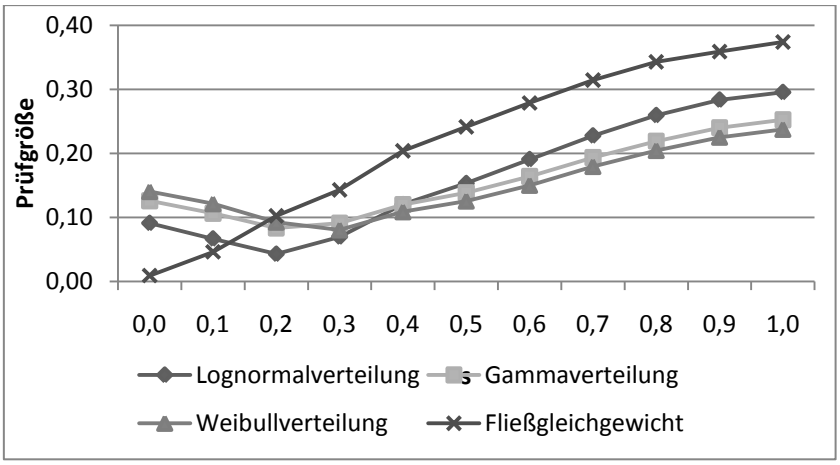


**Abbildung 15: Prüfgrößen für Lognormal-, Weibull-, Gammaverteilung und Fließgleichgewicht des Anpassungstests für $ZG = \frac{ZZ}{ZZ} + (s \cdot ZZ)$ mit zunehmender Störung $s$ mit angepassten Parametern**

Bei diesen angepassten Modellen wird ein ganz anderes Verhalten der Prüfgrößen mit zunehmender Störung ersichtlich. Der Schnittpunkt des Fließgleichgewichts mit einer anderen Verteilung, in diesem Fall mit der Lognormalverteilung, hat sich weiter nach vorn verlagert. Es wird also schon für $s > 0{,}2$ ein besseres Modell gefunden und somit das Modell des Fließgleichgewichts nicht mehr erkannt, obwohl dieses eigentlich vorliegt.
Da in der Praxis Fehler bzw. Störungen nicht ausgeschlossen werden können, wird somit bei realen Daten fast immer das ursprüngliche von der Modellierung genutzte Modell nicht mehr erkannt. In Abschnitt 7 wird deshalb ein mögliches Verfahren zur Beseitigung dieses Problems vorgeschlagen.

Es wird nun noch überprüft, ob Daten die nicht der Verteilung des Fließgleichgewichts unterliegen auch als falsches Modell erkannt werden. Als Zufallsgröße werden hierbei lognormalverteilte Daten verwendet. Auch diese Zufallsgröße wird mit zunehmender Störung untersucht. Dies wird, genau wie die vorausgehenden Daten, erzeugt, indem auf die lognormalverteilte Zufallsgröße eine zufällige Störung aus dem Intervall $[-s; +s]$ mit $s = 0{,}1;\ 0{,}2; \ldots, 1{,}0$. Nach Durchführung des KSA-Tests erhält man das in Abbildung 16 dargestellte Ergebnis.

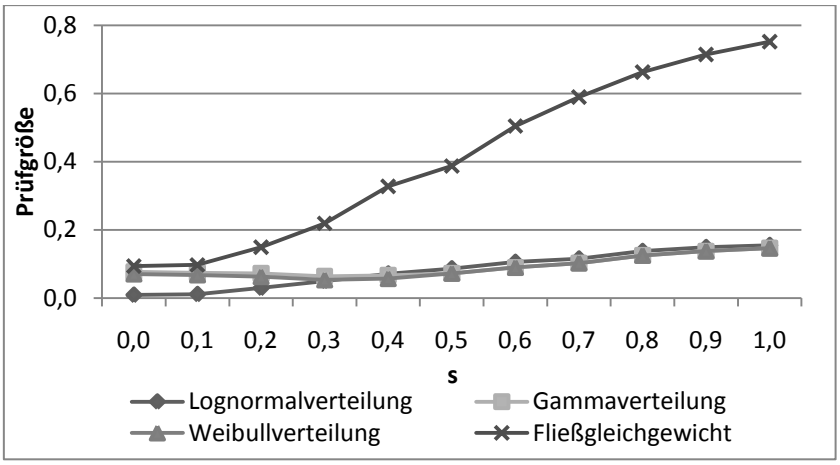


**Abbildung 16: Prüfgrößen für Lognormal-, Weibull-, Gammaverteilung und Fließgleichgewicht des Anpassungstests für $ZG = Log\_ZZ + (s \cdot ZZ)$ mit zunehmender Störung $s$**

Bei kleiner Störung wird die Lognormalverteilung als richtiges Modell erkannt. Ab einer Störung von $s = 0{,}4$ werden andere Verteilungen als ein besseres Modell erkannt. Nur die Prüfgröße des Fließgleichgewichts wird mit zunehmender Störung immer schlechter. Demnach liegt der Fehler 2. Art im Bezug auf das Fließgleichgewicht nicht vor. Das vorliegende Modell wird also nicht fälschlicherweise als Fließgleichgewicht erkannt.

# 7 Faltung

Um die einzelnen Verteilungsfunktionen besser vergleichen zu können, wird im Folgenden über das Verfahren der Faltung eine integrale Analyse von empirischen Verteilungsfunktionen mit Hilfe einer sogenannten Fensterfunktion geschaffen. Dadurch wird nicht nur anhand der maximalen Abweichung der hypothetischen Verteilungsfunktion von der empirischen Verteilungsfunktion entschieden, ob die jeweilige Nullhypothese verworfen oder angenommen wird, sondern der gesamte Kurvenverlauf kann stärker berücksichtigt werden. Es werden auch lokale Veränderungen in der hypothetische Verteilungsfunktion sichtbar und eine qualitativ bessere Testentscheidung ist möglich.

Zur Modellierung einer solchen Fensterfunktion bildet man das Produkt zweier Zufallsgrößen $X$ und $Y$.
Für die Dichtefunktion $f_Z$ des Produktes $Z = X * Y$ gilt allgemein:

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{(X,Y)}\left(t, \frac{z}{t}\right) * \frac{1}{|t|} dt$$

Unter der Voraussetzung der Unabhängigkeit von $X$ und $Y$ erhält man

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(t) * f_Y\left(\frac{z}{t}\right) * \frac{1}{|t|} dt$$

Um die Wirkung einer solchen Fensterfunktion zu demonstrieren, wird folgendes Beispiel betrachtet.

Im Folgenden sei $f_X$ eine lineare Funktion in dem Intervall $[0; 1]$. Die Dichtefunktion sei

$$f_X(t) = \begin{cases} 2 * t, & \text{für } 0 \leq t \leq 1 \\ 0, & \text{sonst} \end{cases}$$

Die Dichtefunktion $f_Y$ wird nun so gewählt, dass sich das Integral des Produktes beider Zufallsgrößen vereinfacht. Man definiert eine Fensterfunktion $f_Y$, für die gilt

$$f_Y\left(\frac{z}{t}\right) \begin{cases} > 0 & \text{für } 0 \leq i \leq \frac{z}{t} \leq j \leq 1 \\ = 0 & \text{sonst} \end{cases}$$

Aus $z = t * s$ und damit $s = z/t$ folgt

$$f_Y(s) = \begin{cases} 0 & \text{für } 0 \leq i \leq s \leq j \leq 1 \\ 0 & \text{sonst} \end{cases}.$$

Um nun die Produktdichte

$$f_{X*Y}(z) = \int_0^1 2 * t * f_Y\left(\frac{z}{t}\right) * \frac{1}{|t|}\, dt$$

berechnen zu können, ist es notwendig die Integrationsgrenzen zu bestimmen. Diese erhält man aus $f_Y(z/t)$ wie folgt

$$i \leq \frac{z}{t} \leq j$$

$$\frac{i}{z} \leq \frac{1}{t} \leq \frac{j}{z}$$

$$\frac{z}{j} \leq t \leq \frac{z}{i}.$$

Da die Dichtefunktion nur im Intervall $[0; 1]$ definiert ist, wird im Weiteren

$$0 \leq \frac{z}{j} \leq t \leq \frac{z}{i} \leq 1$$

gefordert. Die Produktdichte $f_{X*Y}$ wird im Folgenden nur für $z \leq i$ betrachtet.
Es wird nun von $z/j$ bis $z/i$ integriert

$$f_{X*Y}(z) = \int_{\frac{z}{j}}^{\frac{z}{i}} 2 * t * f_Y\left(\frac{z}{t}\right) * \frac{1}{|t|}\, dt\,.$$

Da $i, j \geq 0$ gilt, kann die Betragsfunktion vernachlässigt werden. Nach weiteren Vereinfachungen folgt

$$f_{X*Y}(z) = 2 * \int_{\frac{z}{j}}^{\frac{z}{i}} f_Y\left(\frac{z}{t}\right)\, dt\,.$$

Anschließend wird eine Substitution vorgenommen, damit $f_Y(s)$ durch $f_Y(z/t)$ ersetzt werden kann. Es gilt

$$s = \frac{z}{t}$$

$$t = \frac{z}{s}.$$

Da von $t = z/j$ bis $t = z/i$ integriert wird, erhält man durch Einsetzen in die obigen Formeln

$$\frac{z}{s} = \frac{z}{j} \qquad \rightarrow s = j$$

$$\frac{z}{s} = \frac{z}{i} \qquad \rightarrow s = i.$$

Es wird nun von $j$ bis $i$ integriert. Außerdem wird das Differenzial $dt$ wie folgt ersetzt

$$dt = -\frac{z}{s^2}\, ds.$$

Man erhält für die Produktdichte

$$f_{X*Y}(z) = 2 * \int_j^i f_Y(s) * \left(-\frac{z}{s^2}\right) ds$$

$$f_{X*Y}(z) = -2 * z * \int_j^i \frac{f_Y(s)}{s^2}\, ds.$$

Aus der Integraleigenschaft $\int_i^j f(t)\, dt = -\int_j^i f(t)\, dt$ folgt schließlich

$$f_{X*Y}(z) = 2 * z * \int_i^j \frac{f_Y(s)}{s^2}\, ds.$$

Für $z \leq i$ erhält man für $f_X(t) = 2 * t$ für jedes $f_Y(s)$ immer eine lineare Funktion.

**Beispiel 1:**

Im Folgenden wird die gewonnene allgemeine Formel auf ein konkretes Beispiel angewandt. Es sei $c \in \mathbb{R}$, $i, j \geq 0$ und $f_Y$ die Dichtefunktion einer gleichverteilten Zufallsgröße $Y$ in dem Intervall $[i; j]$ mit

$$f_Y(s) = \begin{cases} c & \text{für } i \leq s \leq j \\ 0 & \text{sonst} \end{cases}.$$

$c$ muss so berechnet werden, dass $f_Y$ eine Dichtefunktion wird. Da die Fläche unter einer Dichtefunktion immer gleich 1 sein muss, muss folgende Gleichung gelten:

$$\int_{-\infty}^{+\infty} f_Y(s)\ ds = 1.$$

Damit die Eigenschaft der Normierung einer Dichtefunktion erfüllt ist, berechnet man $c$ wie folgt:

$$\int_a^b c\ ds = 1$$

$$c * [s]_i^j = 1$$

$$c = \frac{1}{j-i}.$$

Demzufolge lautet die Dichtefunktion

$$f_Y(s) = \begin{cases} \frac{1}{j-i} & \text{für } i \leq s \leq j \\ 0 & \text{sonst} \end{cases}.$$

Für die Produktdichte $f_{X*Y}$ erhält man durch Einsetzen der obigen Formel für $z \leq i$

$$f_{X*Y}(z) = 2 * z * \int_i^j \frac{1}{s^2 * (j-i)}\ ds$$

$$f_{X*Y}(z) = \frac{2*z}{j-i} * \left[-\frac{1}{s}\right]_i^j$$

$$f_{X*Y}(z) = \frac{2*z}{i*j}.$$

**Beispiel 2:**

Es sei $c \in \mathbb{R}$, $i, j \geq 0$ und

$$f_Y(s) = \begin{cases} \frac{c}{s} & \text{für } i \leq s \leq j \\ 0 & \text{sonst} \end{cases}$$

Damit die Funktion $f_Y$ eine Dichtefunktion ist, wird zu Beginn die Konstante $c$ berechnet.

$$\int_i^j \frac{c}{s}\, ds = 1$$

$$c * [\ln s]_i^j = 1$$

$$c = \frac{1}{\ln j - \ln i}$$

$$c = \frac{1}{\ln \frac{j}{i}}.$$

Für $f_Y$ folgt

$$f_Y(s) = \begin{cases} \frac{1}{s * \ln \frac{j}{i}}, \text{ für } i \leq s \leq j \\ 0, \text{ sonst} \end{cases}.$$

Für die Produktdichte $f_{X*Y}(z)$ erhält man

$$f_{X*Y}(z) = 2 * z * \int_i^j \frac{1}{s^3 * \ln \frac{j}{i}}\, ds$$

$$f_{X*Y}(z) = \frac{2 * z}{\ln \frac{j}{i}} * \left[-\frac{1}{2 * s^2}\right]_i^j$$

$$f_{X*Y}(z) = \frac{2 * z}{\ln \frac{j}{i}} * \left(-\frac{1}{2}\right) * \left(\frac{1}{j^2} - \frac{1}{i^2}\right)$$

$$f_{X*Y}(z) = \frac{z}{\ln \frac{j}{i}} * \left(\frac{1}{j^2} - \frac{1}{i^2}\right).$$

Als Ergebnis erhält man in beiden Beispielen eine lineare Funktion für $z \leq i$.

Um die Faltung auf andere Dichtefunktionen anzuwenden, wird anstelle der linearen Funktion $f_X$ eine andere Dichtefunktion eingesetzt.

**Anwendung auf das Fließgleichgewicht**

Die Zufallsgröße $X$ sei im Folgenden nach der Verteilung des Fließgleichgewichts verteilt. Die Dichtefunktion $f_X$ sei

$$f_X(t) = \begin{cases} \frac{b}{2 * a}, \text{für } 0 \leq t \leq \frac{a}{b} \\ \frac{a}{2 * b * t^2}, \text{für } t > \frac{a}{b} \end{cases}$$

und $f_Y$ sei

$$f_Y(s) = \begin{cases} \frac{1}{s * \ln \frac{j}{i}}, \text{für } i \leq s \leq j \\ 0, \text{ sonst} \end{cases}.$$

Es wird nun die Fensterfunktion $f_Y$ über die Dichtefunktion des Fließgleichgewichtes „geschoben“. Im untenstehenden Beispiel sei $f_Y(s)$ im Intervall $[i_k; j_k]$ mit $i = 0{,}1 * (k-1)$ und $j = 0{,}1 * (k+1)$ für $k = 2; 3; \ldots; 8$ größer Null. In Abbildung 17 sind die Dichtefunktionen $f_X$ mit $a = 0{,}4$ und $b = 0{,}8$ sowie $f_Y(s)$ für $k = 3$ dargestellt. Demzufolge ist

$$f_X(t) = \begin{cases} 1, \text{für } 0 \leq t \leq 0{,}5 \\ \frac{4}{t^2}, \text{für } t > 0{,}5 \end{cases}$$

und

$$f_Y(s) = \begin{cases} \frac{1}{s * \ln \frac{0{,}4}{0{,}2}}, \text{für } 0{,}2 \leq s \leq 0{,}4 \\ 0, \text{ sonst} \end{cases}.$$

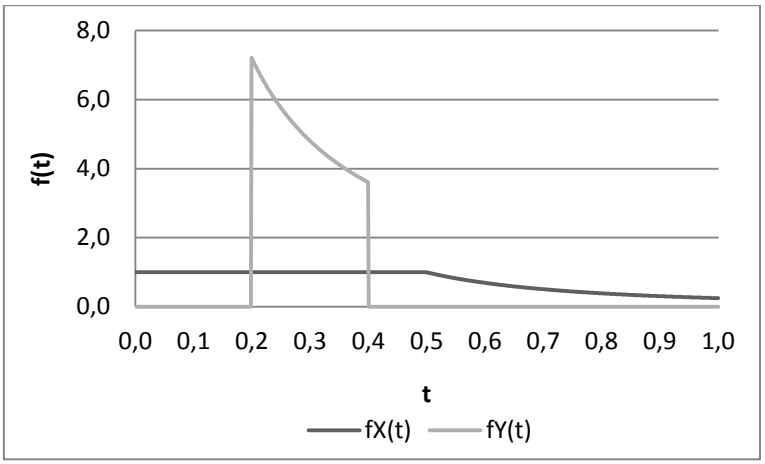


**Abbildung 17: Dichtefunktion $f_X$ mit $a = 0{,}4$ und $b = 0{,}8$ und Fensterfunktion $f_Y$ für $k = 3$**

Man erhält somit sieben Faltungsintegrale, welche anschließend mit Hilfe von numerischer Integration berechnet werden. Hierbei wird die Trapezregel verwendet. Diese besagt, dass die Fläche unter einer Kurve $f(x)$ durch ein Trapez beziehungsweise bei Stückelung des Intervalls durch mehrere Trapeze ersetzt wird. Man kann nun die Kurve $f(x)$ näherungsweise durch eine Sehne zwischen den Funktionswerten an den Stellen $a$ und $b$ ersetzen. Dies führt zur Sehnentrapezformel. Sie ergibt sich aus dem Flächeninhalt des beschriebenen Trapezes

$$Q(f) = \frac{b-a}{2} \cdot \big(f(a) + f(b)\big).$$

Nach Ausführung der numerischen Integration entstehen die in Abbildung 18 dargestellten Produktdichten $f_{X*Y}(z)$.

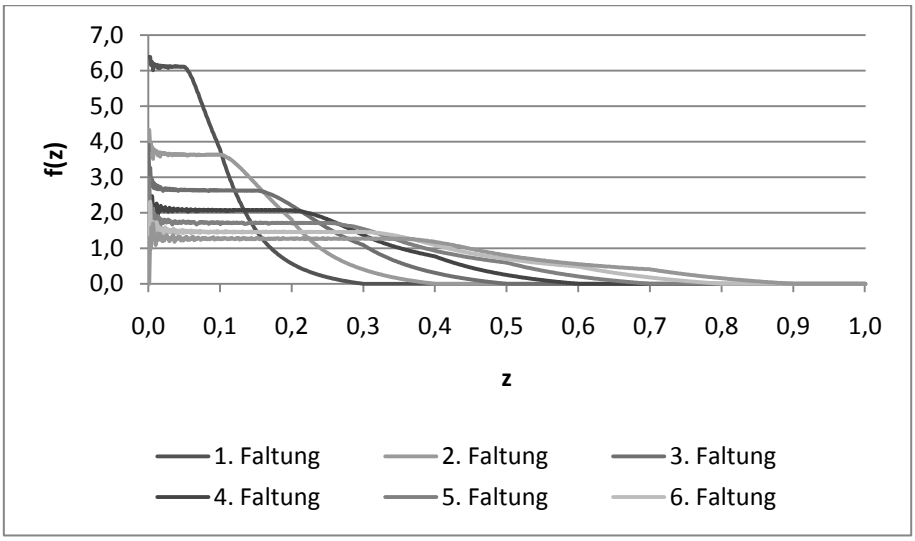


**Abbildung 18: Produktdichten $f_{X*Y}(z)$**

Jede einzelne Dichtefunktion $f_{X*Z}$ in Abbildung 18 weist bei den $z$-Werten nahe 0 einige Schwankungen auf, welche numerisch begründet werden. Denn die Integrale werden nicht analytisch berechnet, sondern durch numerische Integration. Durch die numerisch entstehenden Fehler ist es nicht möglich den Kolmogorov-Smirnov-Anpassungstest auf die zugehörigen Verteilungsfunktionen der erhaltenen Produktdichten anzuwenden, da die Fehler nahe 0 zu hoch sind und diese als maximale Abweichung im Anpassungstest erkannt werden würden, obwohl an dieser Stelle $z$ nach analytischen Berechnungen eventuell gar nicht die maximale Abweichung wäre. Somit wird ein anderes Verfahren benötigt, um verschiedene Produktdichten miteinander vergleichen zu können. Es wird im Folgenden von jeder einzelnen Kurve jeweils nur ein Punkt betrachtet. Bei der Faltung der Dichte des Fließgleichgewichts mit einer Fensterfunktion würde ich drei Möglichkeiten vorschlagen, diesen Punkt aus der Kurve auszuwählen. Zum einen wäre das eine parameterfreie Auswahl. Hierbei wird der Funktionswert an der Stelle $z_0 = i$ betrachtet. Dies hat den Vorteil, dass die beiden Parameter des Fließgleichgewichts $a$ und $b$ nicht bekannt sein müssen. Das ist auch der Fall, wenn man den Funktionswert an der Stelle $z_0 = (i+j)/2$ betrachtet. Die numerisch bedingten Schwankungen werden mit zunehmendem $z$ immer kleiner

und damit erhält man für $z_0 = (i + j)/2$ einen genaueren, mit weniger Fehlern behafteten Wert $f(z_0)$. Wenn der Punkt $f(z_0)$ jedoch in Abhängigkeit der Parameter $a$ und $b$ gewählt werden soll, kann man $f(z_0)$ an der Stelle $z_0 = a/b * i$ bestimmen. Für das obige Beispiel werden in Tabelle 4 alle drei Verfahren angewandt, um die einzelnen Punkte jeder Kurve zu bestimmen.

| | $f_{X*Y}(z_0)$ - ungestört | | |
|---|---|---|---|
| $k$ | $z_0 = i$ | $z_0 = (i+j)/2$ | $z_0 = a/b * i = 0,5 * i$ |
| 1 | 3,8124 | 0,5773 | 6,1004 |
| 2 | 1,8182 | 0,4063 | 3,6279 |
| 3 | 1,0966 | 0,3109 | 2,6267 |
| 4 | 0,7763 | 0,2520 | 2,0739 |
| 5 | 0,5979 | 0,2112 | 1,7132 |
| 6 | 0,4851 | 0,1818 | 1,4610 |
| 7 | 0,4100 | 0,1596 | 1,2741 |

**Tabelle 4: Funktionswerte der Funktion $f_{X*Y}$ an den Stellen $z_0$ der entstandenen Faltungskurven 1 bis 7**

Es wird nun eine abstrakte gestörte Dichtefunktion $f_X{}^*$ des Fließgleichgewichts betrachtet, um die Effekte einer Fensterfunktion zu erarbeiten. Auf die ungestörte Dichte $f_X$ werden additiv zufällige Fehler aufgetragen. Dies geschieht in diesem Beispiel mit einer gleichverteilten Zufallszahl im Intervall $[-0,5; 0,5]$. Der Verlauf dieser Funktion ist in Abbildung 19 dargestellt.

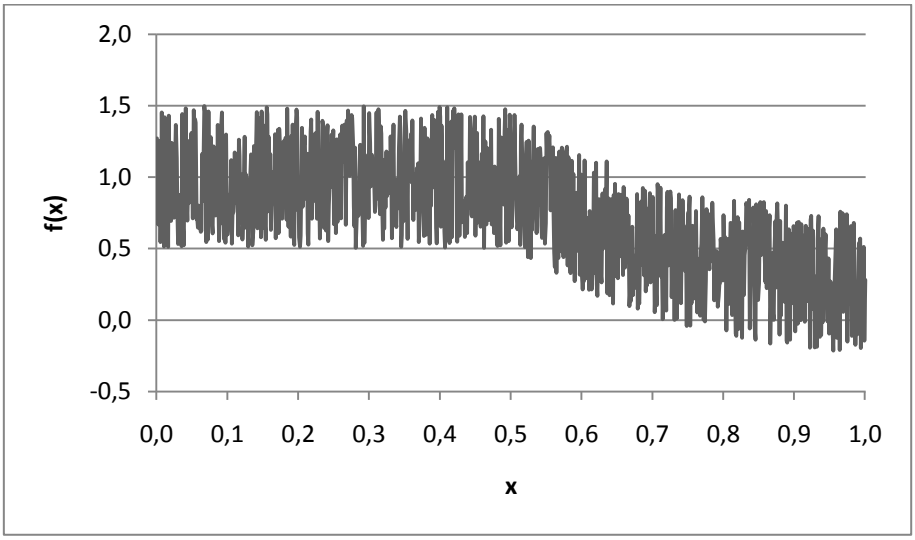


**Abbildung 19: Gestörte Dichtefunktion $f_X^*$**

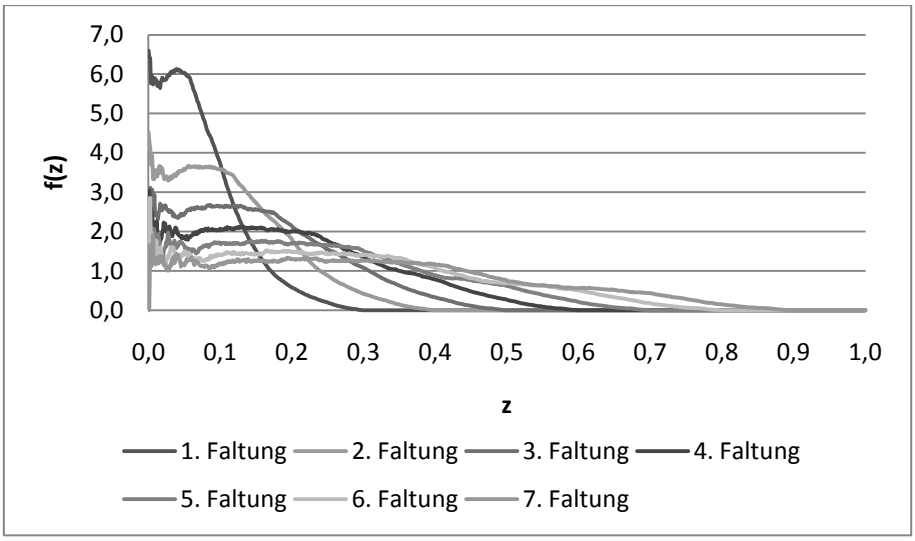


**Abbildung 20: Produktdichten $f_{X*Y}{}^{*}(z)$**

Im Vergleich zu den Produktdichten der ungestörten Dichtefunktion $f_X$ (Abbildung 18), stellt man fest, dass sich der Kurvenverlauf der Produktdichten nicht erheblich ändert, sondern die Werte der Funktionen $f_{X*Y}(z)$ und $f_{X*Y}{}^{*}(z)$ in etwa gleich bleiben (siehe Abbildung 20).
Auch für diese Faltung wird nun aus jeder Kurve der entstandenen Produktdichten ein Punkt entnommen (Tabelle 5). Anschließend wird das ungestörte mit dem gestörten Fließgleichgewicht verglichen.

| | $f_{X*Y}{}^{*}(z_0)$ - gestört | | |
|---|---|---|---|
| $k$ | $z_0 = i$ | $z_0 = (i+j)/2$ | $z_0 = a/b * i = 0,5 * i$ |
| 1 | 3,7785 | 0,5896 | 6,0343 |
| 2 | 1,8344 | 0,4282 | 3,5836 |
| 3 | 1,0902 | 0,3347 | 2,5711 |
| 4 | 0,7917 | 0,2770 | 2,0095 |
| 5 | 0,6252 | 0,2212 | 1,6651 |
| 6 | 0,5141 | 0,1831 | 1,4120 |
| 7 | 0,4441 | 0,1533 | 1,2357 |

**Tabelle 5: Funktionswerte der Funktion $f_{X*Y}{}^{*}$ an den Stellen $z_0$ der entstandenen Faltungskurven 1 bis 7**

Aus den beiden Tabellen wird nun jeweils die Summe der quadratischen Abweichungen

$$PG = \sum \left(f_{X*Y}{}^{*}(z_0) - f_{X*Y}(z_0)\right)^2$$

ermittelt.

Man erhält die in Tabelle 6 dargestellten Werte.

| | $z_0 = i$ | $z_0 = (i+j)/2$ | $z_0 = a/b * i = 0,5 * i$ |
|---|---|---|---|
| $PG$ | 0,0044 | 0,0020 | 0,0198 |

**Tabelle 6: Summe der quadratischen Abweichungen an verschiedenen Stellen $z_0$**

Um Aussagen über die in Tabelle 6 tabellierten Abweichungen zwischen der ungestörten und der gestörten Kurve treffen zu können, wird das Fließgleichgewicht im Folgenden mit zunehmender Störung untersucht. Die Abweichungen zur ungestörten Funktion sind in Abbildung 21 dargestellt, wobei wie oben auf die Dichtefunktion eine zufällige gleichverteilte Zahl aus dem Intervall $[-s;\ +s]$ aufgetragen wird.

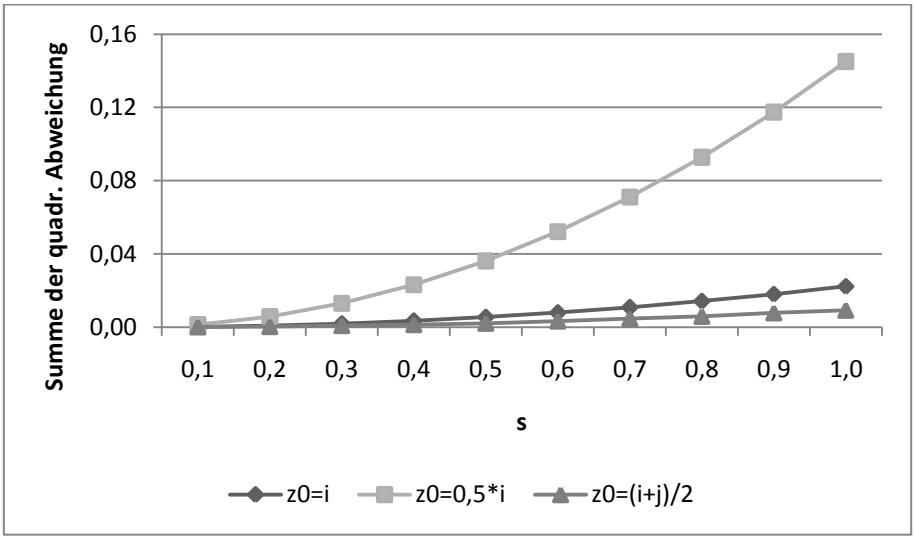


**Abbildung 21: Abweichungen von der gestörten Funktion $f^*_{X*Y}$ zur ungestörten Funktion $f_{X*Y}$ mit zunehmender Störung $s$**

Man stellt fest, dass die Werte für $z_0 = (i+j)/2$ der ungestörten und der gestörten Zufallsgröße sehr nah beieinander liegen. Im Gegensatz dazu entstehen für $z_0 = i$ sowie $z_0 = a/b * i$ größere quadratische Abweichungen. Daraus kann man schlussfolgern, dass die einzelnen Punkte der Produktdichten der gestörten Funktion für $z_0 = (i+j)/2$ nicht so sehr von der ungestörten Funktion abweichen, als für $z_0 = i$ und $z_0 = a/b * i$. Demnach sollte man im Folgenden den Punkt $z_0 = (i+j)/2$ betrachten. Somit wird für die Auswahl der Punkte bei $z_0 = (i+j)/2$ die Modellerkennung noch robuster.

Das Ziel ist es, zu zeigen, dass bei der Faltung die Zuordnung des Modells des Fließgleichgewichts auch bei größeren Störeinflüssen gelingt. Denn wie die Ergebnisse des KSA-Tests zeigen, wird schon bei kleinen Störungen das Modell des Fließgleichgewichts nicht mehr erkannt, sondern die Lognormalverteilung wird als bessere auf die Daten passende Verteilung angegeben. Die Prüfgrößen des KSA-Tests sind in Abbildung 22 dargestellt.

Es wurden 5000 Werte mithilfe der Verteilungsfunktion des Fließgleichgewichts

$$F(x) = \begin{cases} \frac{b * x}{2a} & , 0 < x \leq \frac{a}{b} \\ 1 - \frac{a}{2b * x} & , x > \frac{a}{b} \end{cases}$$

erzeugt. Dabei wurden als $x$-Werte zufällige gleichverteilte Zahlen aus dem Intervall $[0; 1]$ verwendet. Anschließend kommt zu $F(x)$ eine zufällige additive Störung aus dem Intervall $[-s; +s]$ hinzu.

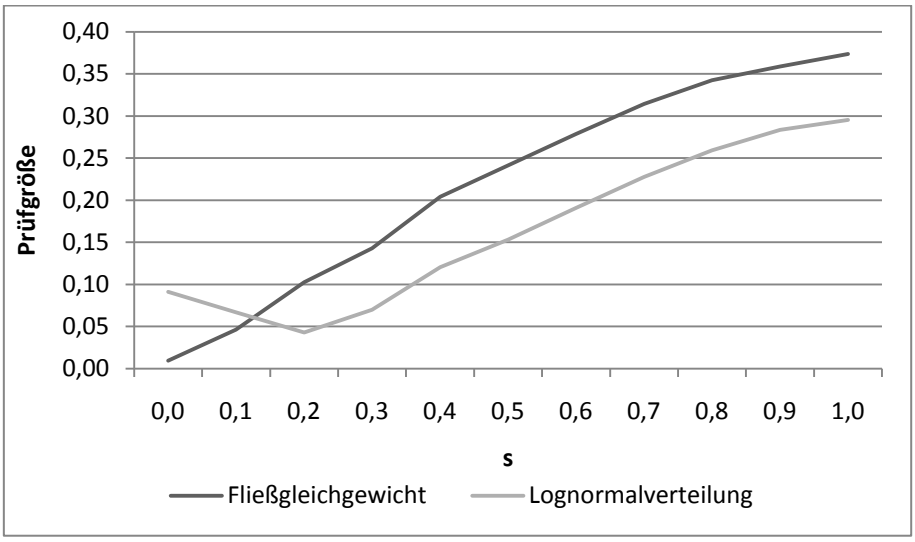


**Abbildung 22: Prüfgröße der Lognormalverteilung und des Fließgleichgewichts des Anpassungstests mit Daten die der Verteilung des Fließgleichgewichts unterliegen mit zunehmender Störung *s***

Um dieses Ergebnis mit der Faltung vergleichen zu können, sollten bei der Faltung möglichst die gleichen Daten verwendet werden, welche auch beim Anpassungstest verwendet wurden. Bei der Faltung wird jedoch die Dichtefunktion benötigt. Um die bereits verwendeten Werte so aufzubereiten, dass sie für die Faltung genutzt werden können, werden zu den vorliegenden 5000 Werten zusätzlich noch 45000 Daten auf die gleiche Art und Weise erzeugt. Es liegen nun insgesamt 50000 Daten vor, die der Verteilung des Fließgleichgewichts unterliegen. Die gegebenen Daten werden anschließend in 1000 äquidistante Klassen eingeteilt. Zunächst wird die absolute Häufigkeit berechnet, anschließend die relative Häufigkeit. Diese erhaltenen relativen Häufigkeiten stellen nun die Dichtefunktion $f_X$ des Fließgleichgewichts für die Faltung dar. Da diese Werte simuliert wurden, treten einige Schwankungen innerhalb der Daten auf, selbst bei der ungestörten Dichtefunktion des Fließgleichgewichts, wie man in Abbildung 23 sieht.

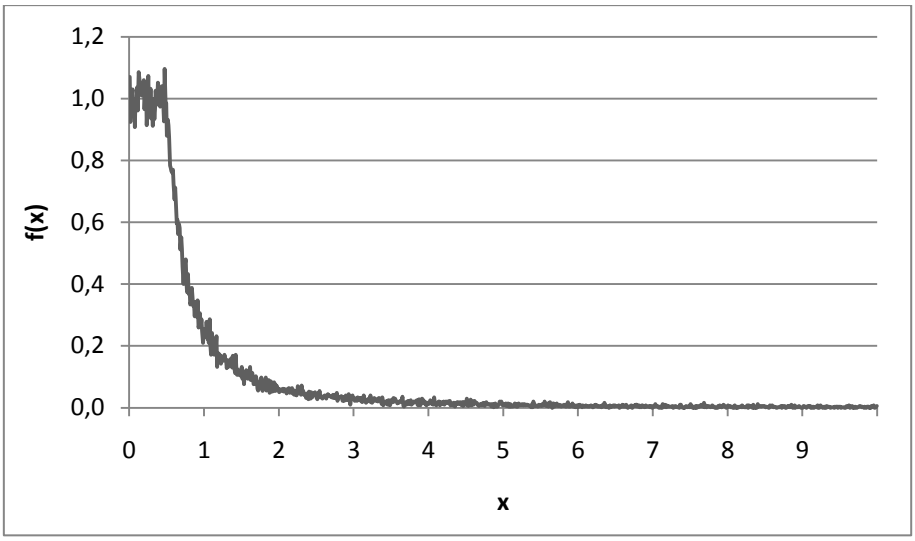


**Abbildung 23: Ungestörte Dichtefunktion $f_X$ des Fließgleichgewichts für die Faltung**

Auch bei dieser Methode der Erzeugung der Dichtefunktion wurden zufällige additive Störungen aufgetragen. Die Störungen wurden jedoch nicht auf die Dichtefunktion selbst aufgetragen, sondern auf die Originaldaten, um einen besseren Vergleich mit dem KSA-Test zu gewährleisten. Aus diesen gegebenen Originaldaten mit Störungen wurde dann wie oben beschrieben, die Dichtefunktion erzeugt. Diese Störungen sind wieder zufällige gleichverteilte Zahlen aus dem Intervall $[-s; +s]$, wobei $s = 0{,}1; 0{,}2; \ldots; 1{,}0$. Bei dieser Methode der Störmodifikation ändert sich schon bei kleinen Störungen der Verlauf der Kurve der Dichtefunktion des Fließgleichgewichts und nimmt damit einen ähnlichen Kurvenverlauf an, wie die konkurrierenden Dichtefunktionen der Lognormalverteilung und der Gammaverteilung. In Abbildung 24 ist die gestörte Dichte des Fließgleichgewichts dargestellt, mit $s = 0{,}2$.

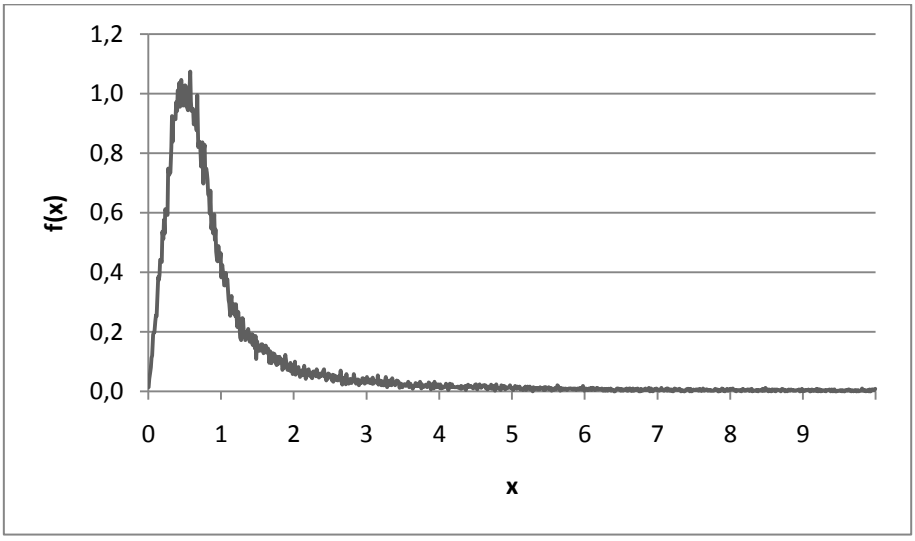


**Abbildung 24: Gestörte Dichtefunktion $f_X$ des Fließgleichgewichts für die Faltung mit $s = 0,2$**

Da man mithilfe der Faltung falsche Modelle mit zunehmender Störung länger als falsch erkennen möchte als beim KSA-Test, wird im Folgenden angenommen, dass die gegebenen Daten einer anderen, dem Fließgleichgewicht konkurrierenden Verteilung unterliegen, wie zum Beispiel der Lognormalverteilung. Die Dichtefunktion der Lognormalverteilung wird ebenfalls nach dem obenstehenden beschriebenen Prinzip erzeugt. Es wird nun die Faltung für das Fließgleichgewicht und für die Lognormalverteilung sowohl ohne Störung als auch mit Störungen ausgeführt. Anschließend werden Funktionswerte an der Stelle $z_0 = (i + j)/2$ ermittelt und folglich die Summe der quadratischen Abweichungen im Bezug auf das ungestörte Fließgleichgewicht berechnet, d.h.

$$PG = \sum \left( {f_{X*Y}}^*(z_0) - f_{X*Y}(z_0) \right)^2$$

und

$$PG = \sum \left( {f_{X*Y}^{LN}}^*(z_0) - f_{X*Y}(z_0) \right)^2,$$

wobei ${f_{X*Y}^{LN}}^*(z_0)$ die gestörte Lognormalverteilung ist. Da diese Abweichungen sehr klein sind, werden sie jeweils mit $10000$ multipliziert, um ein besser handhabbares Zahlenformat zu erhalten. Man erhält damit die in Tabelle 7 aufgelisteten Werte für die Abweichungen.

| **s** | **Summe der quadratischen Abweichungen** | |
|---|---|---|
| | ***Fließgleichgewicht*** | ***Lognormalverteilung*** |
| 0,0 | - | 0,0607 |
| 0,1 | 0,0002 | 0,2514 |
| 0,2 | 0,0016 | 0,2222 |
| 0,3 | 0,0031 | 0,2079 |
| 0,4 | 0,0079 | 0,1876 |
| 0,5 | 0,0088 | 0,1929 |
| 0,6 | 0,0136 | 0,1488 |
| 0,7 | 0,0189 | 0,1322 |
| 0,8 | 0,0388 | 0,1036 |
| 0,9 | 0,0408 | 0,0872 |
| 1,0 | 0,0620 | 0,0766 |

**Tabelle 7: Summe der quadratischen Abweichungen des Fließgleichgewichts und der Lognormalverteilung im Bezug auf das ungestörte Fließgleichgewicht mit zunehmender Störung**

Man erkennt, dass die Abweichungen der lognormalverteilten Daten im Bezug auf das ungestörte Fließgleichgewicht deutlich größer sind. Die tabellierten Werte werden in Abbildung 24 noch einmal veranschaulicht, jedoch bis zu einer Störung von $s = 2{,}0$.

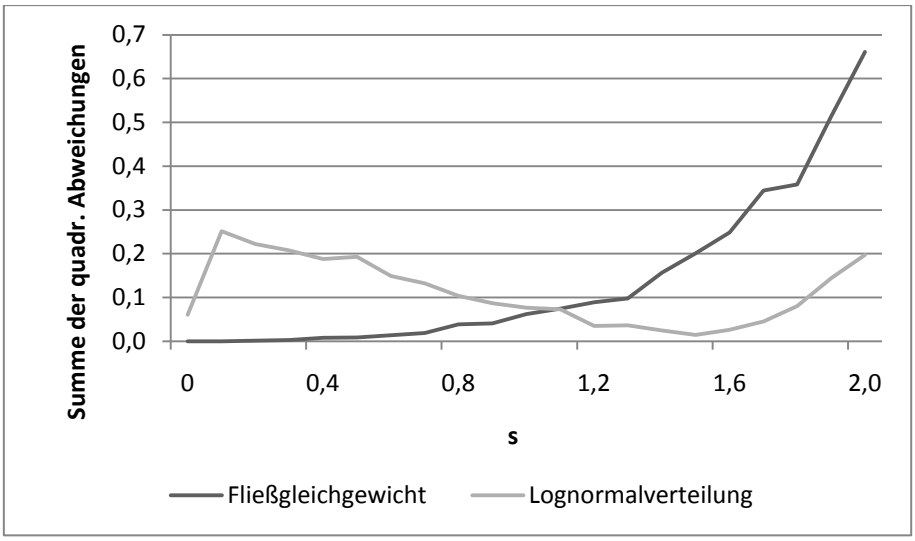


**Abbildung 25: Summe der quadratischen Abweichungen des Fließgleichgewichts und der Lognormalverteilung im Bezug auf das ungestörte Fließgleichgewicht mit zunehmender Störung *s***

Man sieht, dass das Fließgleichgewicht deutlich länger die favorisierte Verteilung bleibt. Im Gegensatz zur Faltung wurde die Lognormalverteilung beim KSA-Test schon ab einer zufälligen additiven Störung im Intervall von $[-0{,}2; +0{,}2]$ besser als das Fließgleichgewicht. Demnach ist das Verfahren der Faltung robuster in der Modellerkennung. Der Fehler 2. Art tritt erst wesentlich später ein, als beim Anpassungstest.

Daraus kann man schlussfolgern, dass die Faltung nicht so stark auf Störungen reagiert wie der Kolmogorov-Smirnov-Anpassungstest. Dies ist ein Vorteil, denn reale Daten können Messfehler beinhalten, welche solche Störungen, wie in diesem Beispiel simuliert wurden, hervorrufen können. Somit würden bei dem Anpassungstest die Daten nicht als Fließgleichgewicht erkannt werden, obwohl das Modell vorliegt. Im Gegensatz dazu wird das Modell bei der Faltung erkannt, obwohl die Daten Störungen enthalten.

Es stellt sich abschließend noch die Frage, ob eine Möglichkeit besteht, aus den ermittelten Abweichungen zwischen ungestörter und gestörter Funktion bei der Faltung eine Prüfgröße zu erzeugen und damit einen statistischen Test zu entwickeln.

## 7.1 Anwendung auf reale Daten

Abschließend wird ein realer Datensatz auf die Verteilung des Fließgleichgewichts getestet. Hierbei wird der Datensatz *Muskel GSM 120719* verwendet. Die $22645$ Realisierungen werden zunächst mithilfe des Anpassungstests geprüft. Dazu ist es notwendig durch systematisches Suchen dasjenige Verhältnis $a/b$ zu finden, welches die kleinste Prüfgröße liefert. Für den gegebenen Datensatz ergibt sich für $a = 54$ und $b = 1/8$ eine minimale Prüfgröße von $0{,}063$. Werden dieselben Daten auf Lognormalverteilung getestet, so ergibt sich für diese Verteilung eine Prüfgröße in Höhe von $0{,}123$. Da bei $22645$ Daten der Rückweisungspunkt sehr klein ist $(0{,}008)$, wird die Nullhypothese beim Anpassungstest immer abgelehnt, da sowohl die Prüfgröße der Lognormalverteilung als auch die des Fließgleichgewichts größer als $0{,}008$ ist. Um nun zu prüfen, ob eine Entscheidung für das Fließgleichgewicht über die Faltung möglich ist, werden die Parameter $a$ und $b$ so angepasst, dass im Anpassungstest die Prüfgröße des Fließgleichgewichts gleich der Prüfgröße der Lognormalverteilung ist. Für $a = 44{,}5$ und $b = 0{,}125$ ergibt sich für das FG eine Prüfgröße von $0{,}123$.

Zunächst wird eine ungestörte Dichtefunktion $f_X$ mit den angepassten Parametern $a = 44{,}5$ und $b = 0{,}125$ aus dem KSA-Test simuliert, um die Ergebnisse der Faltung der realen Dichtefunktion mit den simulierten ungestörten Fließgleichgewicht vergleichen zu können. Diese erzeugte Dichte $f_X$ ist in Abbildung 26 dargestellt.

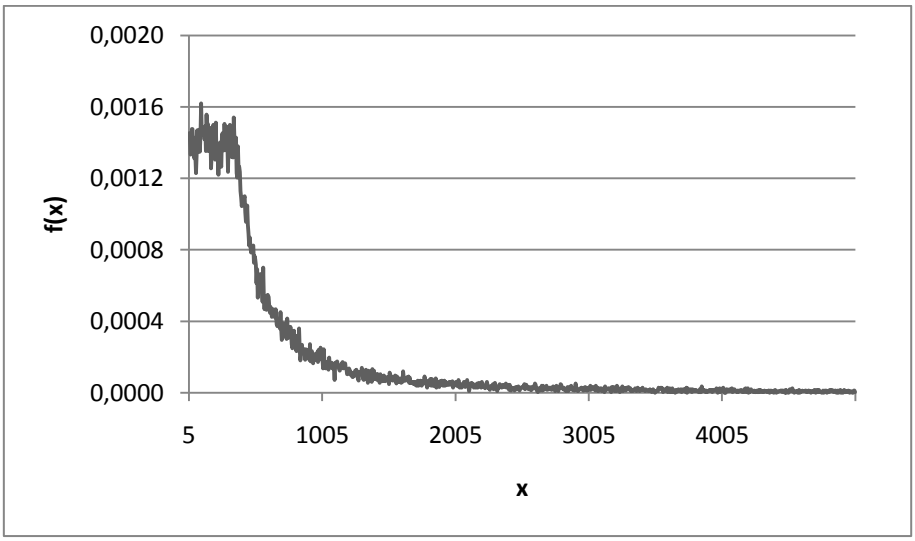


**Abbildung 26: Ungestörte Dichtefunktion $f_X$ des Fließgleichgewichts mit $a = 44{,}5$ und $b = 0{,}125$**

Anschließend wird für diese ungestörte Dichte $f_X$ die Faltung ausgeführt. Die Produktdichten $f_{X*Y}$ sind in Abbildung 27 dargestellt.

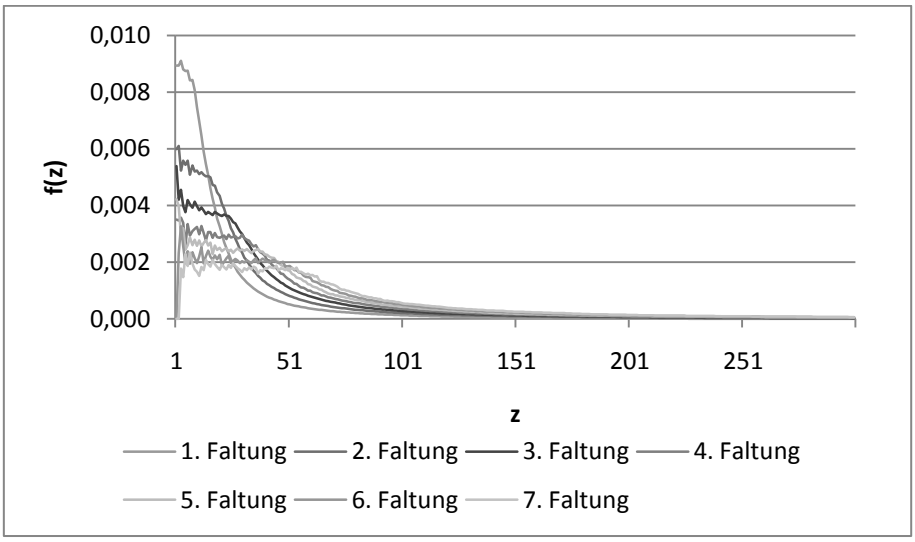


**Abbildung 27: Produktdichten $f_{X*Y}$ des ungestörten Fließgleichgewichts**

Um einen Vergleich zwischen den simulierten Daten des ungestörten Fließgleichgewichts und den realen Daten herzustellen, wird nun aus den gegebenen Daten eine Dichtefunktion $f_X^r$ erzeugt (Abbildung 28).

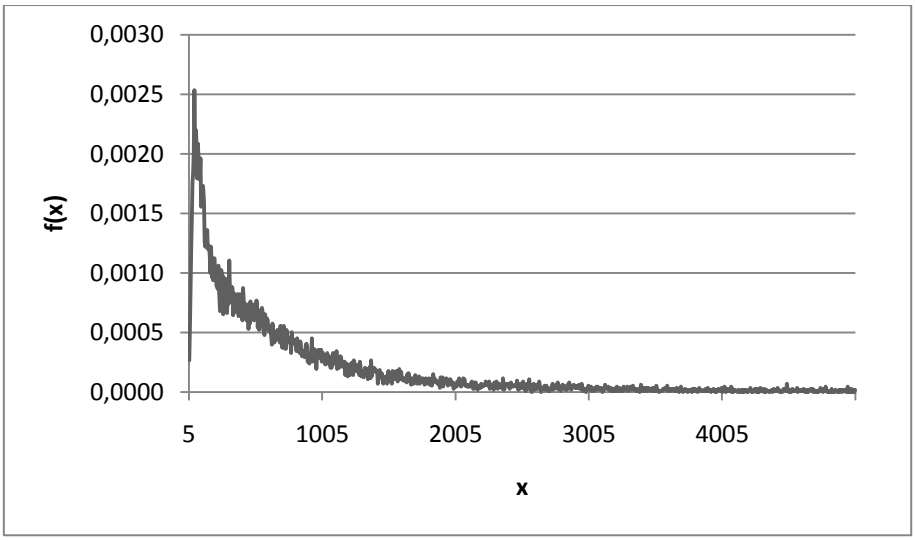


**Abbildung 28: Dichtefunktion $f_X^r$ aus realen Daten *Mensch Muskel GSM 120719***

Anschließend wurde die Faltung durchgeführt und man erhält das in Abbildung 29 dargestellte Ergebnis.

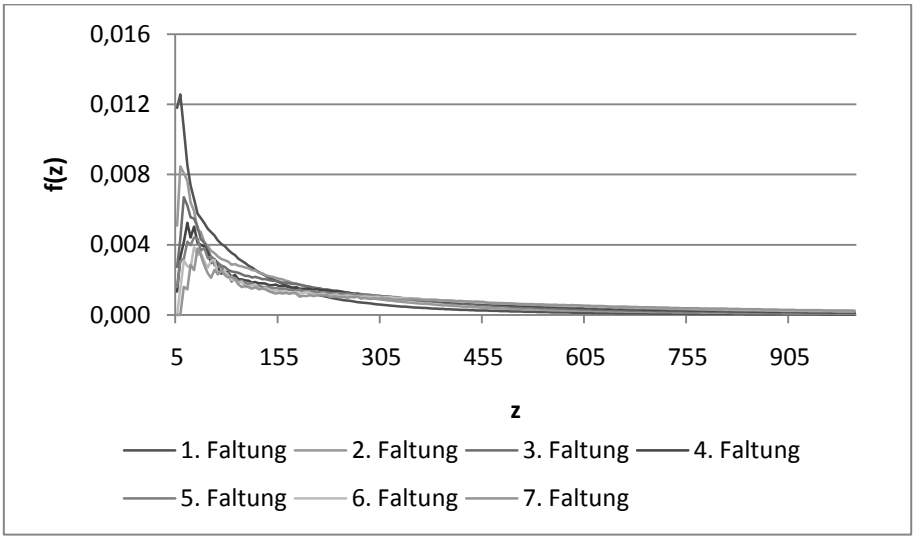


**Abbildung 29: Produktdichten $f^{r}_{X*Y}$ der realen Daten *Mensch Muskel GSM 120719***

Die Summe der quadratischen Abweichungen $PG_r$ zwischen den simulierten ungestörten Daten und den realen Daten beträgt $9{,}12 * 10^{-7}$.

Es wird nun eine ungestörte Lognormalverteilung simuliert und anschließend mithilfe der Faltung die Produktdichten berechnet. Bei der Anwendung der realen Daten *Mensch Muskel GSM 120719* ergaben sich im KSA-Test für den Mittelwert der Lognormalverteilung ein Wert von $5{,}705$ sowie für die Varianz ein Wert von $2{,}76$. Mit diesen Parametern wurde nun die Dichtefunktion $f_X^{LOG}$ erzeugt. In Abbildung 30 ist die Dichtefunktion $f_X^{LOG}$ grafisch dargestellt.

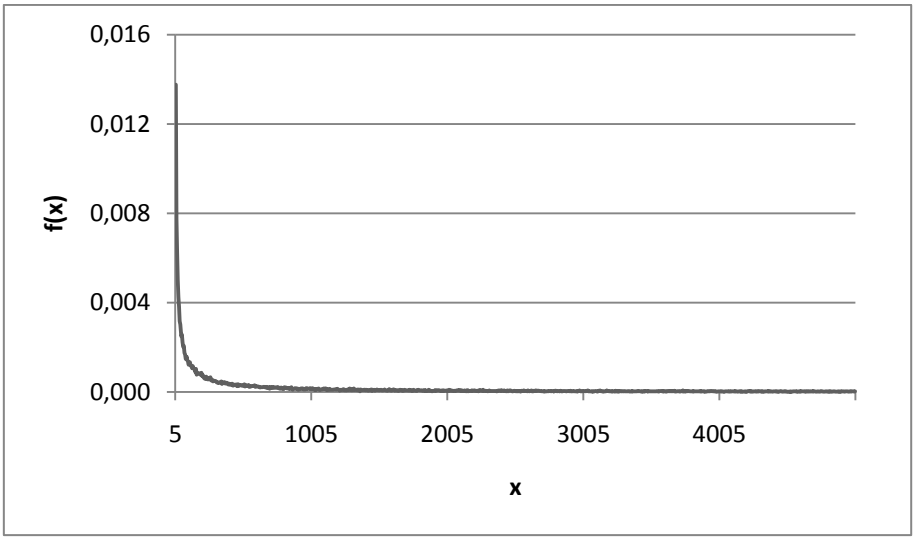


**Abbildung 30: Dichtefunktion $f_X^{LOG}$ der Lognormalverteilung**

Nach anschließender Faltung ergeben sich die in Abbildung 31 dargestellten Produktdichten $f_{X*Y}^{LOG}$.

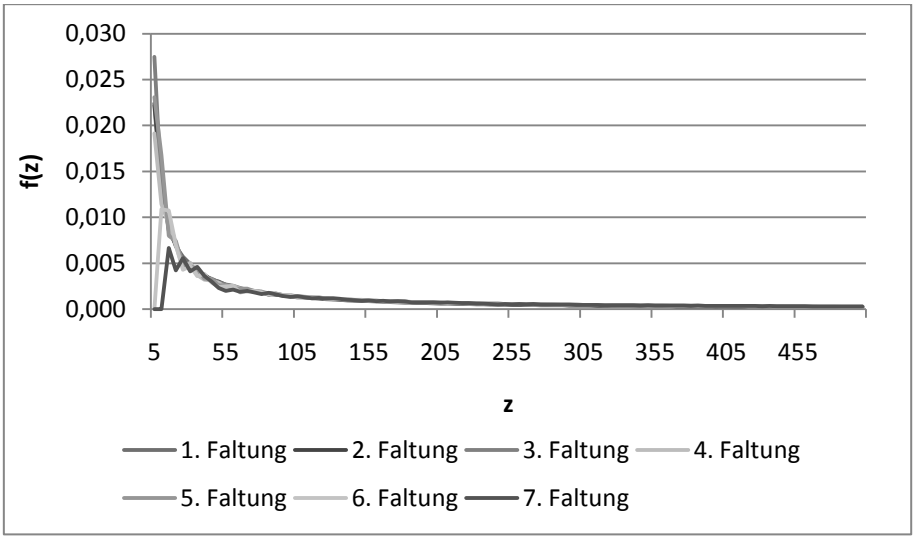


**Abbildung 31: Produktdichten $f_{X*Y}^{LOG}$**

Für die Summe der quadratischen Abweichungen $PG_{LOG}$ zwischen der ungestörten Lognormalverteilung und den realen Daten erhält man ein Wert von $5{,}04 * 10^{-6}$.

Demzufolge werden die realen Daten durch das Verfahren der Faltung eher als Fließgleichgewicht erkannt, da $PG_r$ kleiner ist als $PG_{LOG}$. Die Lognormalverteilung kann demnach über die Faltung ausgeschlossen werden, was bei dem KSA-Test nicht möglich war.

## 7.2 Zusammenfassung

Für die Hypothesenprüfung, ob ein gegebener Datensatz eine bestimmten Verteilungsfunktion unterliegt, kann ein Anpassungstests verwendet werden.

Nach vorliegenden Untersuchungen hat sich bei dem KSA-Test herausgestellt, dass schon bei kleinen Störungseinflüssen die statistisch signifikante Zuordnung des passenden Modells zu simulierten Daten nicht mehr gelingt. Um diesen Effekt zu verbessern wurde das Verfahren der Faltung entwickelt. Tests haben gezeigt, dass diese Methode wesentlich robuster gegenüber Störungen ist.

Zusammenfassend kann man dieses Verfahren der Faltung wie folgt beschreiben.

Zu Beginn sollten aus dem gegebenen Datensatz die charakteristischen Kenngrößen zu vermuteten und der alternativen Verteilungen geschätzt und ggf. bezüglich der Prüfgröße des KSA-Tests optimiert werden. Ist auf diese die Modellanpassung statistisch signifikant abzuweisen, die Prüfgröße also nicht kleiner als der Schwellwert, so werden die Kenngrößen modifiziert, sodass die Prüfgröße des Anpassungstests der favorisierten Verteilung in etwa genauso groß ist, wie die Prüfgröße der konkurrierenden Verteilung.
Ist die konkurrierende Verteilung des Fließgleichgewichts beispielsweise die Lognormalverteilung, so wird anschließend mit Hilfe der Parameter aus dem KSA-Test eine ungestörte Dichtefunktion des Fließgleichgewichts sowie eine ungestörte Dichtefunktion der Lognormalverteilung erzeugt. Desweiteren ermittelt man aus den gegebenen Daten die empirische Dichtefunktion. Die Faltung wird nun mit allen drei erzeugten Dichten durchgeführt. Abschließend berechnet man die Summe der quadratischen Abweichungen nach dem in Kapitel 7 beschriebenen Verfahren zwischen dem ungestörten Fließgleichgewicht und den gegebenen Daten sowie der ungestörten Lognormalverteilung und den Daten. Erhält man als Ergebnis, dass die Summe der quadratischen Abweichungen zwischen ungestörtem Fließgleichgewicht und den Daten kleiner ist, als die andere Prüfgröße, dann folgt daraus, dass man die Lognormalverteilung ausschließen kann. Demzufolge unterliegt der Datensatz dem Modell des Fließgleichgewichts. Umgekehrt könnte man die Verteilung des Fließgleichgewichts ausschließen und damit eher die Lognormalverteilung zuordnen.

# 8 Ausblick

Bei der Anwendung der Faltung auf reale Daten ist es notwendig, das Verhältnis $a/b$ des Fließgleichgewichts zu schätzen. Durch systematische Suche ist es effektiv möglich, die Parameter entsprechend anzupassen. Es wäre jedoch zu prüfen, inwieweit die beiden Parameter aus dem gegebenen Datensatz zu schätzen und damit $a$ und $b$ optimal zu wählen sind.

In vorliegender Arbeit wird sich nur auf die mRNA-Konzentration

$$c = \frac{S}{D}$$

bezogen. In der Literatur wird beschrieben, dass man die Synthese- und Abbauraten jeweils als Produkte von unabhängigen gleichverteilten Zufallsgrößen

$$c = \frac{S^{(m)}}{D^{(n)}}$$

auffassen und modellieren kann. Es zeigt sich, dass die zugehörigen Verteilungsfunktionen für beliebige $m$ und $n$ iterativ als Faltungsprodukte aus Verteilungsfunktionen mit kleineren Indizes ermittelt werden können. Es sollte deshalb untersucht werden, wie sich der Anpassungstest sowie das Verfahren der Faltung auf derart modellierte Zufallsgröße auswirken.

Die im Verfahren der Faltung ermittelte Prüfgröße stellt derzeit ein relatives Vergleichskriterium dar. Es erhebt sich deshalb die Frage, ob diese Prüfgröße im Sinne eines statistischen Signifikanztests auch als absolutes Bewertungskriterium eingesetzt werden kann.

# Literaturverzeichnis


Monod/Pappenheimer/Cohen-Bazaire, 1952
Monod J., Pappenheimer A., Cohen-Bazaire G.: The kinetics of the biosynthesis of beta-galactosidase in Escherichia coli as a function of growth. In: Biochim Biophys Acta 1952; 9: 648-660

Bohley, 1992
Bohley, Peter: Statistik: Einführendes Lehrbuch für Wirtschafts- und Sozialwissenschaftler. – 5., überarb. und erg. Aufl. – München: Oldenbourg, 1992

Rinne, 1997
Rinne, Horst: Taschenbuch der Statistik. – 2., überarb. und erw. Aufl. – Thun: Verlag Harri Deutsch, 1997

Dorner, 1999
Dorner, William <wdorner@qualitydigest.com>: Using Microsoft Excel for Weibull Analysis. URL: <http://www.qualitydigest.com/jan99/html/body_weibull.html>, 1999

Storm, 2001
Storm, Regina: Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle. – 11., verb. Aufl. – Leipzig: Fachbuchverlag Leipzig, 2001

Hoffmann, 2003
Hoffmann, Georg <hoffmann@trillium.de>: SimChip – Computer Simulation of Biochip Signals. URL: <http://www.simchip.de>, 2003

Konishi, 2004
Konishi T.: Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. In: BMC Bioinformatics 2004; 5; 5:82

Hoffmann/Ostermeir/Müller, 2007
Hoffmann G., Ostermeir R., Müller H. et al.: SimChip - Simulation von Genexpressions-Signalen auf der Basis von mRNA-Synthese- und Abbauraten. In: Klinische Chemie – Mitteilungen 2007; 38; 43-48.

Matthäus, 2007
Matthäus, Wolf-Gert: Statistische Tests mit Excel leicht erklärt. – 1.Aufl. – Wiesbaden: Teubner, 2007

Bitterlich/Hoffmann, 2008
Bitterlich N., Hoffmann G.: Über die Schätzung der Wahrscheinlichkeitsmodelle für Genexpressionsprofile. 53. JT GMDS, Stuttgart 2008, Tagungsband S. 41

Bortz/Lienert/Boehnke, 2008
Bortz J., Lienert G.A., Boehnke K.: Verteilungsfreie Methoden in der Biostatistik. – 3. Aufl. – Heidelberg: Springer Medizin Verlag, 2008

Hoffmann/Ostermeir/Müller, 2008
Hoffmann G., Ostermeir R., Müller H. et al.: SimChip - Computer Simulation of mRNA Steady States. In: Clin. Lab. 2008; 54; 19-24.

Microsoft Deutschland GmbH, 2008
Microsoft GmbH <kunden@microsoft.com>: Beschreibung der Funktion ZUFALLSZAHL in Excel 2007 und Excel 2003. URL: <http://support.microsoft.com/kb/828795/de>, 07.01.2008

Biosicherheit, 2009
Biosicherheit <info@biosicherheit.de>: DNA-Microarray. URL: <http://www.biosicherheit.de/de/lexikon/202.dna_microarray.html>, 29.09.2009

ITWissen, 2009
ITWissen <klaus.lipinski@itwissen.info>: Simulation. URL: <http://www.itwissen.info/definition/lexikon/Simulation-simulation.html>, 24.11.2009

# Erklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Sabrina Karthe
Chemnitz, 25.11.2009