Diplomarbeit aus Technischer Mathematik, TU-Graz

# Die Lognormalverteilung als Modell zur Schätzung von betrieblichen Energieverbrauchsfaktoren

Günter Blümel

Juli 1997

Vorgelegt der Technisch-Naturwissenschaftlichen Fakultät an der Technischen Universität Graz

Begutachter: Prof. Dr. Ernst Stadlober
Betreuer: Dr. Herwig Friedl

Institut für Statistik der Technischen Universität Graz

Ich versichere, diese Arbeit selbständig verfaßt, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfsmittel bedient zu haben.

Ich möchte mich bei Herrn Prof. Dr. Ulrich Dieter für die großzügige Bereitstellung der Institutseinrichtungen, die für die Durchführung dieser Arbeit notwendig waren, bedanken.

Danken möchte ich auch Herrn Dr. Herwig Friedl für die Unterstützung bezüglich Statistik-Software und sonstige zahlreiche nützliche Tips im Fachgebiet und im Umgang mit dem Textsystem.

Speziell möchte ich mich bei Herrn Prof. Dr. Ernst Stadlober für die intensive fachliche Betreuung meiner Diplomarbeit bedanken.

Besonderer Dank gilt jedoch meinen Eltern für ihre jahrelange Unterstützung meiner Studientätigkeit.

## Zusammenfassung

Für die Erstellung eines Emissionskatasters der Stadt Graz wurden 311 der 1811 emissionsrelevanten Betrieben statistisch erfaßt. Neben emissionsbezogenen Daten (Lackverbrauch, Brennstoffart, Brennstoffverbrauch etc.) standen für die Analyse auch Energieverbrauchsdaten zur Verfügung. Ausgehend von den Energieverbrauchsfaktoren und Emissionswerten (pro Mitarbeiter und Jahr) werden Schätzungen für den Energieverbrauch und Emissionsausstoß der nicht erfaßten Betriebe -nach Branchen gegliedert- ermittelt. Die Analyse wurde im Rahmen von zwei Diplomarbeiten durchgeführt. Während in dieser Arbeit das theoretische Modell zur Beschreibung von Energieverbrauchs- und Emissionsfaktoren behandelt und eine Energieverbrauchsanalyse durchgeführt wird, analysiert G. Glawar (1997) die Emissionsfaktoren und führt einen Vergleich der Schätzwerte mit Literaturwerten durch.

Die relevanten Kenngrößen können durch zwei- oder dreiparametrische Lognormalverteilungsmodelle beschrieben werden. In Kapitel 1 werden verschiedene Typen von Schätzern für die Verteilungsmomente und -parameter diskutiert und in Kapitel 2 über Monte-Carlo- bzw. Bootstrap-Methoden miteinander verglichen, um davon Entscheidungskriterien für deren Anwendung abhängig von Stichprobenumfang und Parameterkonstellation abzuleiten.

Neben komplexen Methoden zur Berechnung von exakten Konfidenzintervallen für die Verteilungsmomente werden einfache Verfahren für approximative Konfidenzintervalle angegeben.

Die Verteilung der Summe von lognormalverteilten Zufallsvariablen ist nicht in geschlossener Form angebbar. Es wird die Güte der Approximationen durch geeignete Normal - , Lognormal- und Gammaverteilungen bezüglich der Schiefe der Summenverteilung diskutiert.

In Kapitel 5 werden die theoretischen Erkenntnisse verwendet, um Aussagen über den Strom- bzw. Fernwärmeverbrauch pro Mitarbeiter und Jahr zu treffen. Es werden Schätzungen und Konfidenzintervalle für die Lokationscharakteristiken berechnet. Aufbauend auf den Schätzungen werden Entscheidungshilfen für die Wahl eines geeigneten Stichprobenumfanges gegeben. Neben der Globalanalyse werden noch jene Branchen getrennt analysiert, für die eine ausreichende Stichprobe zur Verfügung stand.

# Inhaltsverzeichnis

# Kapitel 1

# Die Lognormalverteilung

## 1.1 Einleitung

Die Lognormalverteilung kann als die Verteilung einer Zufallsvariablen, deren Logarithmus normalverteilt ist, definiert werden. Aufgrund dieser Definition kann schon die erste, wichtige Eigenschaft von lognormalverteilten Zufallsvariablen, die Positivität, abgeleitet werden. Da eine Vielzahl von Prozessen und Zuständen aus allen Lebensbereichen (Biologie, Meteorologie, Wirtschaftswissenschaften, etc.) durch positive Variablen beschrieben werden, ergibt sich ein breites Anwendungsgebiet für das Lognormalverteilungsmodell. Die Restriktion der Positivität kann durch eine Verallgemeinerung der herkömmlichen Lognormalverteilung aufgehoben werden (siehe nächster Abschnitt).

Die wohl charakterisierendste Eigenschaft der Lognormalverteilung ist die positive Schiefe. Vor allem in der Biologie und Epidemiologie wird man häufig mit Daten konfrontiert, die rechtsschief verteilt sind. In den meisten Fällen sind das Beobachtungen beziehungsweise Ereignisse, die in Abhängigkeit von der Zeit erhoben werden. Beispiele wären die Analyse von Überlebensdaten, Analyse von Wachstumsprozessen, Untersuchung des Alters bei erstmaliger Hochzeit, etc. Ein spezieller Bereich, aus dem in weiterer Folge immer wieder Beispiele gebracht werden, ist die Ermittlung branchenbezogener Emissions- und Energieverbrauchsfaktoren für Betriebe, wobei die Faktoren immer auf die Mitarbeiteranzahl bezogen sind.

Die aus der Definition gegebene enge Beziehung zur Normalverteilung wird zur Ermittlung der theoretischen Momente ebenso ausgenutzt, wie bei der Bestimmung von Schätzmethoden für die Verteilungsparameter, von kritischen Bereichen bei Hypothesentests und bei der Berechnung von Konfidenzintervallen.

## 1.2 Definition und Notation

Eine positive Zufallsvariable $X$ ist lognormalverteilt mit den Parametern $\mu$ und $\sigma^2$, wenn die Zufallsvariable $Z = \ln X$ normalverteilt mit dem Erwartungswert $\mu$ und der Varianz $\sigma^2$ ist. Die zweiparametrische Lognormalverteilung wird mit $\Lambda(\mu, \sigma^2)$ bezeichnet, die korrespondierende Normalverteilung mit $N(\mu, \sigma^2)$. Die Verteilungsfunktion von $X \sim \Lambda(\mu, \sigma^2)$

ist

$$
\begin{aligned}
F_X(x) &= P(X \leq x) = P(\ln X \leq \ln x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right) \\
&= \int\limits_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma\xi} \exp\left\{-\frac{(\ln\xi - \mu)^2}{2\sigma^2}\right\} d\xi \qquad (1.1)
\end{aligned}
$$

und die Dichte der Verteilung von $X$ ist

$$
f_X(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\} & x > 0 \\ 0 & x \leq 0 \end{cases} \qquad (1.2)
$$

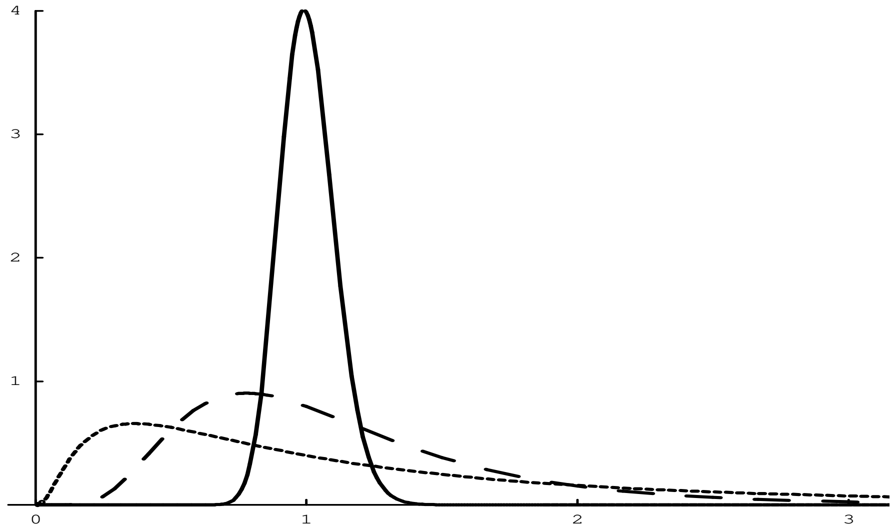



Abbildung 1.1: Dichtefunktion der Lognormalverteilung: durchgezogene Linie: $\Lambda(0, 0.01)$, strichlierte Linie: $\Lambda(0, 0.25)$, punktierte Linie: $\Lambda(0, 1)$

Die Abbildung 1.1 zeigt die Dichtefunktion von $\Lambda(0, \sigma^2)$ für $\sigma = 0.1, 0.5$ und 1. Die Dichte von $\Lambda(\mu, \sigma^2)$ kann als $f(x \mid \mu, \sigma^2) = e^{-\mu} f(xe^{-\mu} \mid 0, \sigma^2)$ geschrieben werden, daher haben die Dichtefunktionen von $\Lambda(\mu, \sigma^2)$ und $\Lambda(0, \sigma^2)$ dieselbe Gestalt. Diese Beziehung wird im nächsten Abschnitt auch aus der Gestalt der Verteilungsmomente ersichtlich. Wie aus Abbildung 1.1 ersichtlich, ist die Lognormalverteilung unimodal und rechtsschief (= linkssteil).

Vermutet man bei vorliegenden Daten auf Grund einer explorativen Analyse, daß die Lognormalverteilungsannahme zutreffen könnte, die Daten einen Wert $\tau$ jedoch nicht unterschreiten, also beschränkt sind, so kann man auf den allgemeineren Fall, die dreiparametrische Lognormalverteilung zurückgreifen. Es wird ein Lokationsparameter $\tau$, der sogenannte *threshold*-Parameter eingeführt, sodaß gilt:

$$
Z = \ln(X - \tau) \sim N\left(\mu, \sigma^2\right)
$$

Die dreiparametrische Lognormalverteilung wird mit $\Lambda(\tau, \mu, \sigma^2)$ bezeichnet. Ein epidemiologisches Beispiel für die Anwendung der dreiparametrischen Lognormalverteilung ist, wie bereits erwähnt, die Beschreibung des Alters bei (erstmaliger) Hochzeit. Die Notwendigkeit für die Einführung des threshold-Parameters läßt sich an diesem Beispiel anschaulich erklären: Die Variable 'Alter bei (erstmaliger) Hochzeit' wird immer größer als ein fixer Wert (z.B. 14) sein; führt man eine Transformation der Daten um diesen Wert durch $(X - 14)$, wird das zweiparametrische Lognormalverteilungsmodell angewandt auf die transformierten Daten, eine bessere Anpassung ergeben. Der Definitionsbereich der Dichtefunktion beginnt bei $X = 14$, Extremwerte in positiver x-Richtung werden zugelassen (Rechtsschiefe). Führt man die Transformation nicht durch, so ergibt sich eine linke Grenze des Definitionsbereiches bei $X = 0$, was bei vorliegendem Problem nicht sinnvoll ist. Durch den threshold-Parameter besteht die Möglichkeit auch negative Daten durch die Lognormalverteilung zu beschreiben ($\tau$ wird negativ, betragsmäßig jedoch groß genug gewählt, damit gilt: $X - \tau > 0$).

Um Verwirrungen in der Notation zu vermeiden, bezeichne künftig $X$ eine zweiparametrisch lognormalverteilte Zufallsvariable und $Y$ eine dreiparametrisch lognormalverteilte Zufallsvariable.

Formal läßt sich die (dreiparametrische) Lognormalverteilung folgend definieren: Existiert eine Größe $\tau$ derart, daß die Variable $Z = \ln(Y - \tau)$ normalverteilt ist, nennt man die Verteilung von $Y$ Lognormalverteilung. $Y$ kann jeden Wert größer $\tau$ annehmen, die Wahrscheinlichkeit einen Wert kleiner als $\tau$ anzunehmen ist 0. Der Ausdruck 'lognormal' kann auch auf die Verteilung von $Y$ angewandt werden, wenn $\ln(\tau - Y)$ normalverteilt ist; die Wahrscheinlichkeit, daß $Y$ einen Wert größer $\tau$ annimmt ist 0. Dieser Fall kann durch Substitution von $Y$ durch $-Y$ und $\tau$ durch $-\tau$ auf den ersten Fall reduziert werden. Die Verteilung von $Y$ wird durch die Gleichung

$$U = \frac{\ln(Y - \tau) - \mu}{\sigma} \tag{1.3}$$

definiert, wobei U eine standardnormalverteilte Variable ist. Aus (1.3) folgt für die Dichtefunktion von Y

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma(y-\tau)} \exp\left\{-\frac{(\ln(y-\tau)-\mu)^2}{2\sigma^2}\right\} & y > \tau \\ 0 & y \leq \tau \end{cases} \tag{1.4}$$

Die zweiparametrische Lognormalverteilung $\Lambda(\mu, \sigma^2)$ ist also ein Spezialfall der dreiparametrischen Lognormalverteilung $\Lambda(\tau, \mu, \sigma^2)$ mit $\tau = 0$.

## 1.3 Momente und andere Charakteristiken

Sei $X \sim \Lambda(\mu, \sigma^2)$ mit der Dichte (1.2). Die Momente der Lognormalverteilung können über die Eigenschaften der erzeugenden Funktion der Normalverteilung gefunden werden. Sei $Z \sim N(\mu, \sigma^2)$. Bezeichne

$$\psi_Z(t) = E[e^{tZ}] = e^{t\mu + \frac{t^2\sigma^2}{2}} \tag{1.5}$$

die momentenerzeugende Funktion von $Z$. Mit $X = e^Z$ ergibt sich für das *r-te Moment um Null*

$$\begin{aligned} \mu_r^{'} &= E[X^r] = E\left[\left(e^Z\right)^r\right] = E\left[e^{rZ}\right] \\ &= \psi_Z(r) = e^{r\mu + \frac{r^2\sigma^2}{2}} \end{aligned} \tag{1.6}$$

Für des *r-te zentrale Moment* gilt

$$\begin{aligned} \mu_r &= E\left[(X - E[X])^r\right] = E\left[(X - \psi_Z(1))^r\right] \\ &= E\left[\sum_{k=0}^{r} (-1)^k \binom{r}{k} (\psi_Z(1))^k X^{r-k}\right] \\ &= \sum_{k=0}^{r} (-1)^k \binom{r}{k} (\psi_Z(1))^k \psi_Z(r-k) \\ &= \sum_{k=0}^{r} (-1)^k \binom{r}{k} \exp\left(\mu(r-k) + \frac{1}{2}\sigma^2(r-k)^2 + \mu k + \frac{1}{2}\sigma^2 k\right) \\ &= \sum_{k=0}^{r} (-1)^k \binom{r}{k} \exp\left(r\mu + \frac{1}{2}\left[(r-k)^2 + k\right]\sigma^2\right) \\ &= e^{r\mu} \sum_{k=0}^{r} (-1)^k \binom{r}{k} e^{\frac{\sigma^2\left[(r-k)^2+k\right]}{2}} \end{aligned} \tag{1.7}$$

Aus (1.6) folgt für den Erwartungswert

$$\mu_1^{'} = e^{\mu + \frac{1}{2}\sigma^2}$$

und aus (1.7) für die Varianz

$$\mu_2 = e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$$

Für die Schiefe und Kurtosis gilt daher:

$$\begin{aligned} \alpha_3 &= \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \left(e^{\sigma^2} - 1\right)^{\frac{1}{2}} \left(e^{\sigma^2} + 2\right) \\ \alpha_4 &= \frac{\mu_4}{\mu_2^2} - 3 = \left(e^{\sigma^2}\right)^4 + 2\left(e^{\sigma^2}\right)^3 + 3\left(e^{\sigma^2}\right)^2 - 3 \end{aligned}$$

Für den Variationskoeffizienten gilt

$$VC = \frac{\sqrt{\mu_2}}{\mu_1'} = \sqrt{e^{\sigma^2} - 1}$$

Die Quantile einer Variablen $X \sim \Lambda(\mu, \sigma^2)$ sind wegen (1.3) definiert durch

$$\theta_q = e^{\mu + z_q \sigma}, \tag{1.8}$$

wobei $q$ ein gegebener Anteil und $z_q$ das $q$-Quantil der Standardnormalverteilung sei. Für den Median gilt daher

$$Med(X) = e^{\mu}$$

Der Modalwert ist

$$Mode(X) = e^{\mu - \sigma^2}$$

Die Momente einer dreiparametrischen lognormalverteilten Zufallsvariable bestimmt man, analog zum zweiparametrischen Fall, über die momentenerzeugende Funktion der korrespondierenden Normalverteilung (1.5).

Es sei $Z = \ln(Y - \tau) \sim N(\mu, \sigma^2)$. Dann gilt für das *r-te Moment um Null* unter Verwendung von

1. $E[(Y-\tau)^r] = E\left[e^{rZ}\right] = \psi_Z(r)$ und
2. $E[(Y-\tau)^r] = \sum_{k=0}^{r}(-1)^k \binom{r}{k} (\tau)^k E\left[Y^{r-k}\right]$

$$\begin{aligned} E[Y^r] &= E[(Y-\tau)^r] - \sum_{k=1}^{r}(-1)^k \binom{r}{k} \tau^k E\left[Y^{r-k}\right] \\ &= \psi_Z(r) - \sum_{k=1}^{r}(-1)^k \binom{r}{k} \tau^k E\left[Y^{r-k}\right] \end{aligned} \tag{1.9}$$

In Gleichung (1.9) sieht man die Auswirkung der Einführung des threshold-Parameters $\tau$ bei der Bestimmung der Momente um Null. Die *zentralen Momente* der dreiparametrischen Lognormalverteilung sind jedoch unabhängig von diesem Parameter:

$$\begin{aligned} E[(Y - E[Y])^r] &= E[((Y-\tau) + \tau - (E[Y-\tau] + \tau))^r] \\ &= E[((Y-\tau) - E[Y-\tau])^r] \\ &= E[((Y-\tau) - \psi_Z(1))^r] \\ &= E\left[\sum_{k=0}^{r}(-1)^k \binom{r}{k} (\psi_Z(1))^k (Y-\tau)^{r-k}\right] \\ &= \sum_{k=0}^{r}(-1)^k \binom{r}{k} (\psi_Z(1))^k \psi_Z(r-k) \\ &= E[(X - E[X])^r], \end{aligned} \tag{1.10}$$

wobei $Y \sim \Lambda(\tau, \mu, \sigma^2)$ und $X \sim \Lambda(\mu, \sigma^2)$.

Die Lokationscharakteristiken der dreiparametrischen Lognormalverteilung $\Lambda(\tau, \mu, \sigma^2)$ sind also um den threshold- Parameter größer als jene von $\Lambda(\mu, \sigma^2)$. Ist $Y \sim \Lambda(\tau, \mu, \sigma^2)$, so gilt:

$$\begin{aligned} E(Y) &= \tau + e^{\mu + \frac{\sigma^2}{2}} \\ Med(Y) &= \tau + e^{\mu} \\ Mode(Y) &= \tau + e^{\mu - \sigma^2} \end{aligned}$$

Die Quantile verschieben sich ebenfalls um den Wert $\tau$ von $\theta_q$ zu $\tau + \theta_q$. Die zentralen Momente (und deren Kombinationen) von $Y$ sind mit jenen von $X$ ident.

## 1.4 Schätzmethoden für die Parameter der dreiparametrischen Lognormalverteilung

Sei $Y = (Y_1, \ldots, Y_n)$ eine Zufallsstichprobe der Größe $n$, wobei $Y_i \overset{iid}{\sim} \Lambda(\tau, \mu, \sigma^2)$. Seien $y_i$ die beobachteten Werte von $Y$. $T = T(Y_1, \ldots, Y_n)$ heißt *Stichprobenfunktion*. Schätzt $T$ einen unbekannten Parameter, dann nennt man $T$ einen *Schätzer*. Eine Realisierung $t = t(y_1, \ldots, y_n)$ des Schätzers $T$ nennt man *Schätzung*.
Güteeigenschaften von Schätzern:

1. Eine Folge von Schätzer $(T_n)_{n \in \mathbb{N}}$ des Parameters $u$ heißt *konsistent*, wenn $\lim_{n \to \infty} P(\mid T_n - u \mid > \epsilon) = 0$ gilt.

2. Eine Folge von Schätzer $(T_n)_{n \in \mathbb{N}}$ des Parameters $u$ heißt *asymptotisch erwartungstreu*, wenn $\lim_{n \to \infty} E(T_n) = u$ gilt.

3. Ein Schätzer $T$ heißt *erwartungstreu* , wenn $E(T) = u$ gilt. Ein nicht erwartungstreuer Schätzer heißt *biased*.

4. Seien $T_1$ und $T_2$ zwei Schätzer für den Parameter $u$. $T_1$ heißt *wirksamer* als $T_2$, wenn $E\left((T_1 - u)^2\right) < E\left((T_2 - u)^2\right)$ gilt. Sind $T_1$ und $T_2$ erwartungstreu, so gilt $Var(T_1) < Var(T_2)$. $T_w$ heißt *wirksamster* Schätzer von $u$, falls für alle Schätzer $T$ von $u$ die Ungleichung $E\left((T_w - u)^2\right) \leq E\left((T - u)^2\right)$ gilt. In vielen Fällen existiert der wirksamste Schätzer jedoch nicht.

### 1.4.1 Die Maximum-Likelihood-Methode (MLE)

Bei der Maximum-Likelihood-Methode erhält man für einen unbekannten Parameter $u$ eine Schätzung, indem man die Dichte $f_u(y_1, \ldots, y_n)$ des Zufallvektors $Y$ maximiert, d.h. man bestimmt sich den Wert $\hat{u}(y_1, \ldots, y_n)$, der unter Annahme einer bestimmten Verteilung die größte Wahrscheinlichkeit ergibt. Man nennt die Funktion $L(u; y_1, \ldots, y_n) := f_u(y_1, \ldots, y_n)$ (betrachtet als Funktion von $u$ unter festen Beobachtungen $y_i$) *Likelihoodfunktion*. Sind die $Y_i$ unabhängig identisch verteilt mit der Dichte $f_u(y_i)$, dann gilt: $L(u; y_1, \ldots, y_n) = \prod_{i=1}^{n} f_u(y_i)$. In diesem Fall ist es sinnvoll, die Likelihoodfunktion durch die logarithmierte Funktion, der sogenannten *Loglikelihoodfunktion*, zu ersetzen (wegen der Monotonie des Logarithmus ist eine derartige Transformation bei Extremwertbestimmung zulässig). Der Maximum-Likelihood-Schätzer muß nicht erwartungstreu sein, es gilt aber:

1. Falls die *Likelihoodgleichung* $\frac{\partial \ln L(u;y_1,\ldots,y_n)}{\partial u} = 0$ eindeutig lösbar ist, dann ist der Schätzer *konsistent*.

2. Existiert eine wirksamste Schätzung, dann wird sie durch die Maximum- Likelihood-Methode bestimmt.

Im Fall der dreiparametrischen Lognormalverteilung erhält man durch partielle Ableitung der Loglikelihoodfunktion nach den drei Parametern $\tau, \mu$ und $\sigma^2$ folgendes Gleichungssystem (Dichte siehe (1.4)):

$$\begin{aligned}
\frac{\partial \ln L\left(\hat{\tau}, \hat{\mu}, \hat{\sigma^2}; y_1, \ldots, y_n\right)}{\partial \hat{\mu}} &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} \left(\ln\left(y_i - \hat{\tau}\right) - \hat{\mu}\right) = 0 \\
\frac{\partial \ln L\left(\hat{\tau}, \hat{\mu}, \hat{\sigma^2}; y_1, \ldots, y_n\right)}{\partial \hat{\sigma}} &= -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3} \sum_{i=1}^{n} \left(\ln\left(y_i - \hat{\tau}\right) - \hat{\mu}\right)^2 = 0 \\
\frac{\partial \ln L\left(\hat{\tau}, \hat{\mu}, \hat{\sigma^2}; y_1, \ldots, y_n\right)}{\partial \hat{\tau}} &= \frac{1}{\hat{\sigma}^3} \sum_{i=1}^{n} \frac{\left(\ln\left(y_i - \hat{\tau}\right) - \hat{\mu}\right)}{y_i - \hat{\tau}} + \\
&+ \sum_{i=1}^{n} \left(y_i - \hat{\tau}\right)^{-1} = 0 \qquad (1.11)
\end{aligned}$$

Eliminiert man in dem Gleichungssystem $\hat{\mu}$ und $\hat{\sigma}^2$, so erhält man:

$$\begin{aligned}
\lambda(\hat{\tau}) &= \left(\sum_{i=1}^{n} \left(y_i - \hat{\tau}\right)^{-1}\right) \left(\sum_{i=1}^{n} \ln\left(y_i - \hat{\tau}\right) - \sum_{i=1}^{n} \ln^2\left(y_i - \hat{\tau}\right) + \frac{1}{n}\left[\sum_{i=1}^{n} \ln\left(y_i - \hat{\tau}\right)\right]^2\right) \\
&- n \sum_{i=1}^{n} \frac{\ln\left(y_i - \hat{\tau}\right)}{y_i - \hat{\tau}} = 0 \qquad (1.12)
\end{aligned}$$

Diese Gleichung kann iterativ, beispielsweise durch das Newton-Raphson-Verfahren gelöst werden. Bei der Lösung werden nur Nullstellen zugelassen, für die $\hat{\tau} < y_{(1)}$ mit $y_{(1)} = \min(y_1, \ldots, y_n)$ gilt. Üblicherweise existiert nur eine einfache Nullstelle der Gleichung (1.12); sollten mehrere Lösungen existieren, so wählt man jene, für die $\overline{y}$ und $\widehat{E(Y)}$ am nächsten beieinander liegen. Sei nun $\hat{\tau}$ so bestimmbar, daß $\lambda(\hat{\tau}) = 0$. Dann gilt:

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} \ln\left(y_i - \hat{\tau}\right) \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} \ln^2\left(y_i - \hat{\tau}\right) - \left[\frac{1}{n} \sum_{i=1}^{n} \ln\left(y_i - \hat{\tau}\right)\right]^2 \qquad (1.13)
\end{aligned}$$

Die Varianz von Maximum-Likelihood-Schätzern kann asymptotisch über die Inverse der *Fisher-Informationsmatrix* (negative erwartete Werte für die 2. partiellen Ableitungen der Likelihoodfunktion) abgeschätzt werden (siehe COHEN (1951)): Sei $\omega = e^{\sigma^2}$, $\beta = e^{\mu}$, $H = \left[\omega\left(1+\sigma^2\right) - \left(1+2\sigma^2\right)\right]^{-1}$ und $K = \frac{\beta}{\sqrt{\omega}}$. Dann gilt:

$$\begin{array}{ll} Var(\hat{\tau}) \approx \frac{\sigma^2}{n}\frac{\beta^2}{\omega}H & Cov(\hat{\tau},\hat{\mu}) \approx -\frac{\sigma^2}{n}HK \\ Var(\hat{\mu}) \approx \frac{\sigma^2}{n}(1+H) & Cov(\hat{\tau},\hat{\sigma}) \approx \frac{\sigma^3}{n}HK \\ Var(\hat{\sigma}) \approx \frac{\sigma^2}{2n}(1+2\sigma^2 H) & Cov(\hat{\mu},\hat{\sigma}) \approx -\frac{\sigma^3}{n}H \end{array}$$

### 1.4.2 Die modifizierte Maximum-Likelihood-Methode (MMLE)

Grundsätzlich gibt es zwei modifizierte Maximum-Likelihood-Methoden. Beide basieren auf dem selben Prinzip: Die iterativ zu lösende Gleichung aus dem System (1.11) wird durch eine adäquate Gleichung ersetzt. Beide Methoden verwenden Ordnungsstatistiken zur Substitution der dritten Gleichung aus (1.11) (siehe COHEN (1951) und COHEN and WHITTEN (1980)).

#### MMLE-I

Die dritte Gleichung aus dem System (1.11) wird durch die Beziehung

$$E(F(Y_r)) = F(y_r) \tag{1.14}$$

ersetzt, wobei $Y_r$ die r-te Ordnungsstatistik der Zufallsstichprobe $Y = (Y_1, \ldots, Y_n)$ und $F(Y_r)$ der Wert der Verteilungsfunktion an der Stelle der r-ten Ordnungsstatistik ist. Bezeichnet man mit $u_r = \frac{\ln(y_r-\tau)-\mu}{\sigma}$, dann gilt: $\Phi(u_r) = F(y_r)$, wobei $\Phi$ die Verteilungsfunktion der Standardnormalverteilung und $F$ die Verteilungsfunktion von $\Lambda\left(\tau,\mu,\sigma^2\right)$ ist. Wegen $E(F(Y_r)) = \frac{r}{n+1}$ gilt: Die Bedingung (1.14) ist äquivalent zu

$$\ln\left(y_r - \tau\right) = \mu + \sigma k_r \tag{1.15}$$

wobei $k_r = \Phi^{-1}\left(\frac{r}{n+1}\right)$. Es folgt

$$\lambda_1(\hat{\tau}) = \ln\left(y_r - \hat{\tau}\right) - \hat{\mu}(\hat{\tau}) - \hat{\sigma}(\hat{\tau})k_r \quad = \quad 0, \tag{1.16}$$

wobei $\hat{\mu}(\hat{\tau})$ und $\hat{\sigma}(\hat{\tau})$ aus (1.13) zu bestimmen sind.

Diese Gleichung ist wiederum iterativ zu lösen. Liegen mehrere Lösungen vor, so wählt man jene, für die der Abstand zwischen $\overline{y}$ und dem Schätzer für den Erwartungswert minimal wird. Hat man den Schätzer für $\tau$ eruiert, so erhält man Schätzer für $\mu$ und $\sigma^2$ durch Einsetzen in Bedingung (1.13).

#### MMLE-II

Bei der MMLE-II-Methode geht man von einer zu MMLE-I ähnlichen Überlegung aus. Man substituiert die iterativ zu lösende Gleichung des MLE-Schemas durch

$$\ln\left(y_r - \tau\right) = \mu + \sigma E(Z_r), \tag{1.17}$$

wobei $E(Z_r)$ der Erwartungswert der r-ten Ordnungsstatistik einer standardnormalverteilten Zufallsstichprobe vom Umfang $n$ ist. Diese Bedingung ist der Substitutionsbedingung der MMLE-I-Methode sehr ähnlich; der Ausdruck $k_r$ wird durch $E(Z_r)$ ersetzt. Es folgt

$$\ln(y_r - \hat{\tau}) - \hat{\mu}(\hat{\tau}) - E(Z_r)\hat{\sigma}(\hat{\tau}) \quad = \quad 0 \tag{1.18}$$

Bei der Lösung dieser Bedingung bzw. bei der Bestimmung der Schätzer für $\mu$ und $\sigma$ geht man analog zur MLE-Methode bzw. zur MMLE-I-Methode vor. Eine Tabelle mit Werten von $E(Z_r)$ findet man bei HARTER (1961).

### 1.4.3 Die Momentenmethode (ME)

Bei der Momentenmethode erhält man für Parameter, die sich in bekannter Weise aus den Momenten der Verteilung zusammensetzen, Schätzungen, indem man die theoretischen Momente durch die empirischen Momente ersetzt. Aus dem *Schwachen Gesetz der großen Zahlen (Satz von Kolmogorov)* folgt, daß die Momentenmethode *erwartungstreue* und *konsistente* Schätzer für die Momente der Ordnung $k$ liefert.

Der Zusammenhang der theoretischen Momente mit den Parametern ist durch folgendes Gleichungssystem gegeben:

$$\begin{aligned}
\tau + e^{\mu + \frac{\sigma^2}{2}} &= \mu_1^{'} \\
e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right) &= \mu_2 \\
\left(e^{\sigma^2} - 1\right)^{\frac{1}{2}}\left(e^{\sigma^2} + 2\right) &= \alpha_3
\end{aligned} \tag{1.19}$$

Substituiert man in der dritten Gleichung von (1.19) $e^{\sigma^2}$ durch $\omega$ und schreibt diese dann in der Form $\omega^3 + 3\omega^2 - (a_3^2 + 4) = 0$ an, so gilt nach der *Vorzeichenregel von Descarte*: Die Gleichung hat eine positive Lösung und ein Paar konjungiert komplexer Lösungen. Damit ist $\omega$ und folglich auch $\sigma$ eindeutig bestimmbar. Folglich existieren auch für die ersten beiden Gleichungen aus (1.19) eindeutige Lösungen.

Schätzer für $\tau, \mu$ und $\sigma$ erhält man, indem man die ersten drei theoretischen Momente der dreiparametrischen Lognormalverteilung $\mu_1^{'}$, $\mu_2$ und $\alpha_3$ durch die ersten drei Stichprobenmomente $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ und $a_3 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^3/s^{\frac{3}{2}}$ ersetzt.

$$\hat{\mu} \quad = \quad \ln\left(\frac{s}{\sqrt{\hat{\omega}(\hat{\omega} - 1)}}\right) \tag{1.20}$$

$$\hat{\sigma} \quad = \quad \sqrt{\ln\hat{\omega}} \tag{1.21}$$

$$\hat{\tau} \quad = \quad \overline{y} - \frac{s}{\sqrt{\hat{\omega} - 1}} \tag{1.22}$$

Die Schätzung $\hat{\tau}$ könnte als Startwert für die Iteration zur Lösung der dritten Gleichung aus (1.11) bzw. der korrespondierenden Gleichungen der modifizierten Maximum-Likelihood-Methoden verwendet werden.

### 1.4.4 Die modifizierte Momentenmethode (MME)

Die Momentenmethode kann analog zur MLE-Methode modifiziert werden, indem man eine Gleichung aus dem Gleichungssystem (1.19) substituiert. Hierbei kann man dieselben Ersatzgleichungen wie bei MMLE-I (MMLE-II) verwenden.

**MME-I**

Man ersetzt die dritte Gleichung aus (1.19) durch die Beziehung (1.14). Diese Beziehung kann auch in der Form

$$y_r = \tau + e^{\mu + \sigma k_r} \tag{1.23}$$

geschrieben werden (Notation analog zu MMLE-I). Eliminiert man in dem resultierenden Gleichungssystem $\mu$ und $\tau$, so erhält man als Bestimmungsgleichung für $\hat{\omega}$:

$$\frac{s^2}{(\overline{y} - y_r)^2} = \frac{\hat{\omega}(\hat{\omega} - 1)}{\left(\sqrt{\hat{\omega}} - e^{\hat{\sigma} k_r}\right)^2} \tag{1.24}$$

Diese Gleichung muß wiederum iterativ nach $\hat{\omega}$ gelöst werden. Um Schätzungen der übrigen Parameter zu erhalten, setzt man in (1.20) ein.

**MME-II**

Ersetzt man die dritte Gleichung aus (1.19) durch die Beziehung (1.17), so erhält man folgende Beziehung für $\hat{\omega}$:

$$\frac{s^2}{(\overline{y} - y_r)^2} = \frac{\hat{\omega}(\hat{\omega} - 1)}{\left(\sqrt{\hat{\omega}} - e^{\hat{\sigma} E(Z_r)}\right)^2} \tag{1.25}$$

Ansonsten geht man analog zu MME-I vor.

## 1.5 Schätzmethoden für die Parameter der zweiparametrischen Lognormalverteilung

Neben dem Momentenschätzer und dem Maximum-Likelihood-Schätzer wird noch der sogenannte *UMVU-Schätzer* (uniform minimum variance unbiased) vorgestellt. Sei $X = (X_1, \ldots, X_n)$ eine Zufallsstichprobe der Größe $n$, wobei $X_i \overset{iid}{\sim} \Lambda(\mu, \sigma^2)$. Seien $x_i$ die beobachteten Werte von $X$.

### 1.5.1 Der Maximum-Likelihood-Schätzer (MLE)

Der ML-Schätzer der Parameter einer zweiparametrischen Lognormalverteilung ist ein Spezialfall des ML-Schätzers der dreiparametrischen Lognormalverteilung mit $\tau = 0$. Maximiert man die Likelihoodfunktion $L(\mu, \sigma^2; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_x(\mu, \sigma; x_i)$ (bzw. die Loglikelihoodfunktion) bezüglich $\mu$ und $\sigma^2$, so erhält man:

$$\begin{aligned}
\frac{\partial \ln L(\hat{\mu}, \hat{\sigma}; x_1, \ldots, x_n)}{\partial \hat{\mu}} &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (\ln(x_i) - \hat{\mu}) = 0 \\
\frac{\partial \ln L(\hat{\mu}, \hat{\sigma}; x_1, \ldots, x_n)}{\partial \hat{\sigma}} &= -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3} \sum_{i=1}^{n} (\ln(x_i) - \hat{\mu})^2 = 0
\end{aligned}$$

Löst man diese Gleichungen nach $\hat{\mu}$ bzw. nach $\hat{\sigma}^2$, so erhält man die Schätzer:

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} \ln(x_i) \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} \ln^2(x_i) - \left[ \frac{1}{n} \sum_{i=1}^{n} \ln(x_i) \right]^2
\end{aligned}$$

### 1.5.2 Momentenschätzer (ME)

Es gilt:

$$\begin{aligned}
e^{\mu + \frac{\sigma^2}{2}} &= \mu_1' \\
e^{2\mu + \sigma^2} \left( e^{\sigma^2} - 1 \right) &= \mu_2
\end{aligned}$$

Daher ergibt sich für die Schätzer von $\mu$ und $\sigma$:

$$\begin{aligned}
\hat{\mu} &= \ln(\overline{x}) - \ln\left[ \left( \frac{s}{\overline{x}} \right)^2 + 1 \right] \\
\hat{\sigma} &= \sqrt{\ln\left[ \left( \frac{s}{\overline{x}} \right)^2 + 1 \right]}
\end{aligned}$$

### 1.5.3 UMVU-Schätzer

Betrachtet man die Funktion

$$\theta_{a,b,c} = \sigma^{2c} e^{a\mu + b\sigma^2} \qquad a, b, c \in \mathbb{R}, \tag{1.26}$$

so lassen sich einige Populationscharakteristiken als Linearkombination dieser Funktion darstellen. Das r-te Moment um Null ist durch $\theta_{r,r^2/2,0}$, der Median durch $\theta_{1,0,0}$, der Modalwert durch $\theta_{1,1,0}$ und die Varianz durch $\theta_{2,2,0} - \theta_{2,1,0}$ gegeben. Um den UMVU-Schätzer (uniformly minimum variance unbiased) dieser Funktion bestimmen zu können, müssen einige Hilfsfunktionen definiert und einige Zwischenschritte durchgeführt werden (siehe CROW and SHIMIZU (1988)).

Für $Z_i = \ln X_i$ gilt $Z_i \overset{iid}{\sim} N(\mu, \sigma^2)$. Daher gilt für das empirische Mittel $\overline{Z} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ und demnach gemäß (1.5) auch:

$$E[e^{a\overline{Z}}] = e^{a\mu + \frac{a^2}{2n}\sigma^2} \qquad a \in \mathbb{R} \tag{1.27}$$

Weiters hat man mit $S_Z = \sum_{i=1}^n \left(Z_i - \overline{Z}\right)^2$ die Eigenschaft $\frac{S_Z}{\sigma^2} \sim \chi^2_{n-1}$ und $\overline{Z}$ und $S_Z$ sind unabhängig.

Um einen Schätzer für (1.26) zu erhalten, muß daher (wegen (1.27) und der Unabhängigkeit von $S_Z$ und $\overline{Z}$) nur eine Funktion $h\left(S_Z\right)$ gefunden werden mit $E\left(\hat{\theta}_{a,b,c}\right) = E\left(h(S_Z)e^{a\overline{Z}}\right) = E\left(h(S_Z)\right) E\left(e^{a\overline{Z}}\right) = E\left(h(S_Z)\right) e^{a\mu + \frac{a^2}{2n}\sigma^2} = \sigma^{2c} e^{a\mu + b\sigma^2}$ und somit

$$E\left[h\left(S_Z\right)\right] = \sigma^{2c} e^{\left(b - \frac{a^2}{2n}\right)\sigma^2} = \sum_{j=0}^{\infty} \frac{1}{j!}\left(b - \frac{a^2}{2n}\right)^j \sigma^{2(c+j)}$$

Da $\frac{S_Z}{\sigma^2} \sim \chi^2_{n-1}$ gilt für das $j$-te Moment von $S_Z$:

$$E\left[S_Z^j\right] = \frac{2^j \Gamma\left(\frac{n-1}{2} + j\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma^{2j} \quad ,$$

wobei $\Gamma(a)$ die vollständige Gammafunktion bezeichnet. Es folgt für $h\left(S_Z\right)$:

$$\begin{aligned} h\left(S_Z\right) &= \sum_{j=0}^{\infty} \frac{1}{j!}\left(b - \frac{a^2}{2n}\right)^j \frac{\Gamma\left(\frac{n-1}{2}\right)}{2^{j+c}\Gamma\left(\frac{n-1}{2} + j + c\right)} S_Z^{c+j} = \\ &= \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2} + c\right)} \left(\frac{S_Z}{2}\right)^c \sum_{j=0}^{\infty} \frac{1}{j!}\left(b - \frac{a^2}{2n}\right)^j \left(\frac{S_Z}{2}\right)^j \frac{1}{\left(\frac{n-1}{2} + c\right) \cdots \left(\frac{n-1}{2} + c + j - 1\right)} \end{aligned}$$

Definiert man

$$\begin{aligned} {}_0F_1(\alpha, w) &:= \sum_{j=0}^{\infty} \frac{w^j}{(\alpha)_j j!}, \\ \text{mit } (\alpha)_j &= \begin{cases} \alpha(\alpha+1)\cdots(\alpha+j-1) & j \geq 1 \\ 1 & j = 0 \end{cases}, \end{aligned}$$

dann kann man $h\left(S_Z\right)$ schreiben als:

$$h\left(S_Z\right) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}+c\right)}\left(\frac{S_Z}{2}\right)^c {}_0F_1\left(\frac{n-1}{2}+c;\frac{2bn-a^2}{4n}S_Z\right) \tag{1.28}$$

Aus (1.27) und (1.28) folgt für den UMVU-Schätzer der Funktion (1.26):

$$\hat{\theta}_{a,b,c} = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}+c\right)}e^{a\overline{Z}}\left(\frac{S_Z}{2}\right)^c {}_0F_1\left(\frac{n-1}{2}+c;\frac{2bn-a^2}{4n}S_Z\right)$$

Der Schätzer für eine Linearkombination der Form $\theta = \sum_{(a,b,c)} C_{a,b,c}\theta_{a,b,c}$ ist

$$\hat{\theta} = \sum_{(a,b,c)} C_{a,b,c}\hat{\theta}_{a,b,c}$$

Für die Schätzer von Erwartungswert und Varianz einer zweiparametrisch lognormalverteilten Zufallsvariable $X$ ergibt sich:

$$\begin{aligned}
\widehat{E[X]} &= \hat{\theta}_{1,1/2,0} = e^{\overline{Z}}{}_0F_1\left(\frac{n-1}{2};\frac{n-1}{4n}S_Z\right)\\
\widehat{Var[X]} &= \hat{\theta}_{2,2,0}-\hat{\theta}_{2,1,0} = e^{2\overline{Z}}\left[{}_0F_1\left(\frac{n-1}{2};\frac{n-1}{n}S_Z\right)-{}_0F_1\left(\frac{n-1}{2};\frac{n-2}{2n}S_Z\right)\right]
\end{aligned}$$

Eine Wertetabelle der Funktion ${}_0F_1(\alpha;w)$, die ein Spezialfall der verallgemeinerten hypergeometrischen Verteilung ist, kann bei SHIMIZU and IWASE (1981) gefunden werden. Mathematik-Softwarepakete wie `Mathematica` oder `Maple` enthalten die hypergeometrischen Funktionen ebenfalls.
Die Herleitung der Varianz der Schätzer für Erwartungswert und Varianz kann bei CROW and SHIMIZU (1988) nachgeschlagen werden. Es gilt:

$$\begin{aligned}
Var\left(\widehat{E[X]}\right) &= e^{2\mu+\sigma^2}\left[e^{\frac{\sigma^2}{n}}{}_0F_1\left(\frac{n-1}{2};\frac{(n-1)^2}{4n^2}\sigma^4\right)-1\right]\\
Var\left(\widehat{Var[X]}\right) &= e^{4\mu+2\sigma^2}\left[e^{2\sigma^2}\left[e^{\frac{4\sigma^2}{n}}{}_0F_1\left(\frac{n-1}{2};\frac{4(n-1)^2}{n^2}\sigma^4\right)-1\right]+\right.\\
&\quad + e^{\frac{4\sigma^2}{n}}{}_0F_1\left(\frac{n-1}{2};\frac{(n-1)^2}{n^2}\sigma^4\right)-1-\\
&\quad \left.- 2e^{\sigma^2}\left[e^{\frac{4\sigma^2}{n}}{}_0F_1\left(\frac{n-1}{2};\frac{2(n-1)(n-2)}{n^2}\sigma^4\right)-1\right]\right]
\end{aligned}$$

### 1.5.4 Quantil-Schätzer

Die Schätzer für die Quantile einer lognormalverteilten Zufallsvariable werden in dieser Arbeit nur angeführt. Die Herleitung der Schätzer und der Varianz der Schätzer kann bei CROW and SHIMIZU (1988) nachgeschlagen werden. Sei $q_p$ das $p$-te Quantil von $X \sim \Lambda\left(\mu,\sigma^2\right)$ und $z_p$ jenes der Standardnormalverteilung.

**UMVU-Schätzer für $q_p$**

$$_{UMVU}\hat{q}_p = e^{\overline{Z}} \sum_{k=0}^{\infty} H_k\left(\sqrt{n} z_q\right) \frac{\Gamma\left(\frac{n-1}{2}\right)}{k!(2n)^{\frac{k}{2}} \Gamma\left(\frac{k+n-1}{2}\right)} S_Z^{\frac{k}{2}},$$

wobei $H_k$ das Hermite'sche Polynom

$$H_k(x) = (-1)^k e^{\frac{x^2}{2}} \left(\frac{d^k}{dx^k} e^{-\frac{x^2}{2}}\right)$$

bezeichnet.

$$Var\left(_{UMVU}\hat{q}_p\right) = e^{2\mu}\left[e^{\frac{2\sigma^2}{n}} \sum_{k=0}^{\infty} W_k \frac{\Gamma\left(\frac{k+n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{\sigma}{\sqrt{n}}\right)^k - e^{2z_q\sigma}\right]$$

mit

$$W_k = \sum_{r+s=k} \frac{\Gamma^2\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{r+n-1}{2}\right) \Gamma\left(\frac{s+n-1}{2}\right) r!s!} H_r\left(\sqrt{n} z_q\right) H_s\left(\sqrt{n} z_q\right) \qquad \text{mit} \quad r \geq 0, s \leq k$$

**Maximum-Likelihood-Schätzer für $q_p$**

$$_{ML}\hat{q}_p = e^{\overline{Z} + z_p \sqrt{\frac{S_Z}{n}}}$$

**Schätzer für $q_p$ aufgrund der Stichprobenquantile**

$$_{Q}\hat{q}_p = X_{([pn+1])}$$

Der Schätzer ist asymptotisch normalverteilt mit dem Erwartungswert $q_p$ und der Varianz

$$Var\left(_{Q}\hat{q}_p\right) = \frac{1}{n}\left(2p(1-p)\pi\sigma^2 e^{z_p^2 + 2z_p\sigma + 2\mu}\right)$$

**Schätzer für $q_p$ aufgrund der Stichprobenmomente**

$$_{M}\hat{q}_p = \overline{Z}\left(1 + \frac{S_X}{n\overline{X}^2}\right)^{-\frac{1}{2}} \exp\left(z_p \sqrt{\ln\left(1 + \frac{S_X}{n\overline{X}^2}\right)}\right)$$

Auch dieser Schätzer ist asymptotisch normalverteilt mit dem Erwartungswert $q_p$ und der Varianz

$$\begin{aligned} Var\left(_{M}\hat{q}_p\right) &= \frac{1}{n} e^{2\mu - 2\sigma^2 + 2z_p\sigma}\left(e^{\sigma^2} - 1\right)\left[\left(1 + \left(e^{\sigma^2} - 1\right)\left(2 - \frac{z_p}{\sigma}\right)\right)^2 + \right. \\ &\quad + \left(e^{\sigma^2} - 1\right)\left(e^{\sigma^2} + 2\right)\left(1 + \left(e^{\sigma^2} - 1\right)\left(2 - \frac{z_p}{\sigma}\right)\right)\left(\frac{z_p}{\sigma} - 1\right) + \\ &\quad \left. + \frac{1}{4}\left(e^{\sigma^2} - 1\right)\left(e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 4\right)\left(\frac{z_p}{\sigma} - 1\right)^2\right] \end{aligned}$$

# Kapitel 2

# Vergleich der Schätzmethoden

## 2.1 Einleitung

In Kapitel 1 wurden sowohl für die zweiparametrische als auch für die dreiparametrische Lognormalverteilung einige Schätzmethoden für die Momente und Verteilungsparameter vorgestellt. In diesem Abschnitt werden diese Methoden miteinander verglichen. Dabei wurde folgende Vorgehensweise gewählt:

1. Getrennte Betrachtung der Methoden für den zwei- bzw. dreiparametrischen Fall.

2. Für beide Verteilungen werden die Schätzmethoden für die Verteilungsmomente und für die Verteilungsparameter verglichen. (Hier ergeben sich vor allem bei der dreiparametrischen Verteilung interessante Ergebnisse.)

3. Im zweiparametrischen Fall wird der Vergleich auf zwei Stufen durchgeführt:

   (a) Monte-Carlo-Simulation: In Abhängigkeit vom Stichprobenumfang (10, 20, 50, 100) und der Schiefe der Verteilung (0.2, 0.5, 1, 2, 4) wurden jeweils 100 lognormalverteilte Zufallsstichproben (mit `Mathematica`) generiert und auf jede dieser 100 Pseudo-Stichproben die verschiedenen Schätzmethoden angewandt. Aus den jeweils 100 Schätzwerten wurde das empirische Mittel und die empirische Standardabweichung der Schätzer ermittelt.

   (b) Bootstrapmethode: Mit Hilfe der Bootstrapmethode ist es möglich, die Verteilungen der Schätzer zu approximieren und anschließend miteinander zu vergleichen. Hierbei wurden nicht mehr alle Konstellationen bezüglich des Stichprobenumfangs und der Schiefe berücksichtigt. Bei der Schiefe wurden die Werte 0.2, 1 und 4 gewählt, um die Auswirkung der Gestalt der Ausgangsverteilung auf die Verteilung der Schätzer zu demonstrieren. Das Grundprinzip der Bootstrap-Methode wird im nächsten Abschnitt vorgestellt.

4. Im dreiparametrischen Fall wird der Vergleich der Schätzmethoden ähnlich wie bei der Simulationsstudie von COHEN and WHITTEN (1980) durchgeführt.

## 2.2 Die Bootstrap-Methode

Eine effiziente Resampling-Methode ist die von Bradley Efron im Jahre 1977 eingeführte Bootstrap-Methode. Das Bootstrapverfahren gewinnt seine Werte generell durch wiederholtes und unabhängiges Ziehen mit Zurücklegen aus der Ausgangsstichprobe, wobei die entsprechenden Verteilungsparameter durch Monte-Carlo-Simulation ermittelt werden, außer in den wenigen Fällen, in denen eine theoretische Berechnung möglich ist. Es werden nun die verschiedenen Bootstrapmethoden, die in dieser Arbeit eingesetzt werden, kurz beschrieben. Nähere Ausführungen findet man bei EFRON (1979) und RAPPITSCH (1991).

### 2.2.1 Der nichtparametrische Bootstrap

Das Prinzip des nichtparametrischen Bootstrap liegt darin, die Verteilung einer Zufallsvariablen $R(X, F)$, die von einem Stichprobenvektor $X = (X_1, \ldots, X_n)$ und einer unbekannten Verteilungsfunktion $F$ mit $X_i \overset{iid}{\sim} F$ abhängt, durch die Verteilung einer Zufallsvariablen $R^*$ zu approximieren. Diese Zufallsvariable $R^*$ erhält man folgendermaßen: Man ersetzt die unbekannte Verteilungsfunktion $F$ durch die empirische Verteilungsfunktion $\hat{F}$ der beobachteten Daten $x = (x_1, \ldots, x_n)$. Das führt zu einem neuen Stichprobenvektor $X^* = (X_1^*, \ldots, X_n^*)$ mit $X_i^* \overset{iid}{\sim} \hat{F}$. $R^*$ ist nun einfach die ursprüngliche Funktion $R$ angewandt auf $X^*$. Das Verfahren ist nichtparametrisch, da kein Wissen über die unbekannte Verteilungsfunktion $F$ oder etwaige Verteilungsparameter in die Konstruktion eingehen.

Die Motivation für diese Verfahren liegt darin, daß jeder nichtparametrische Schätzer für die Verteilung (oder einen Verteilungsparameter) von $R(X, F)$ nahe beim richtigen Wert liegen sollte, falls $F = \hat{F}$ gilt. Das Bootstrapverfahren liefert in diesem Fall genau die richtige Lösung. Das Prinzip, daß man die ursprüngliche Verteilung der Zufallsvariablen $R(X, F)$ durch die Bootstrapverteilung von $R\left(X^*, \hat{F}\right)$ approximiert, läßt sich in analoger Weise auf die Schätzung von Verteilungsparametern von $R$ übertragen.

### 2.2.2 Der parametrische Bootstrap

Während beim nichtparametrischen Bootstrap die Verteilung $F$ der zugrundeliegenden Grundgesamtheit als unbekannt vorausgesetzt wird, nimmt man beim parametrischen Bootstrap an, daß die Form der zugrundeliegenden Verteilung bis zu einem bestimmten Grad bekannt, aber nicht vollständig spezifiziert ist. In Abhängigkeit vom Grad der Bekanntheit resultieren verschiedene Formen des parametrischen Bootstrapverfahrens. Die stärkste Form läßt sich folgendermaßen darstellen:

Es sei $X = (X_1, \ldots, X_n)$ ein Stichprobenvektor mit $X_i \overset{iid}{\sim} F_\eta$ und $\eta = (\eta_1, \ldots, \eta_p)$ ein Vektor von unbekannten Verteilungsparametern. Dabei sei die Verteilungsfamilie $F_\eta$ bekannt und der Vektor $\eta$ unbekannt. Will man die Verteilung einer Zufallsvariablen $R(X, F)$ schätzen, geht man wie folgt vor:

1. Berechnung des ML-Schätzers $\hat{\eta}$ für $\eta$. Daraus erhält man die Verteilungsfunktion $F_{\hat{\eta}}$.

2. Man zieht eine Bootstrapstichprobe $X^* = (X_1^*, \ldots, X_n^*)$ mit $X_i^* \overset{iid}{\sim} F_{\hat{\eta}}$.

3. Approximation der Verteilung von $R(X, F)$ durch die Verteilung von $R(X^*, F_{\hat{\eta}})$.

### 2.2.3 Die Berechnung der Bootstrapverteilung

Die direkte, exakte Berechnung der Bootstrapverteilung der Zufallsvariablen $R\left(X^*, \hat{F}\right)$ ist meist nur in einfachen Fällen möglich. Die übliche Methode der Approximation der Bootstrapverteilung ist die Monte-Carlo-Methode, die die entsprechende Verteilung und die Verteilungsparameter durch Simulation ermittelt: Es sei $R(X, F)$ eine Zufallsvariable, deren Verteilung durch die Bootstrapverteilung von $R\left(X^*, \hat{F}\right)$ approximiert werden soll. Die Bootstrapverteilung selbst wird nun wieder durch wiederholtes Ziehen aus $\hat{F}$ approximiert. Dazu gibt man eine Anzahl $B$ von Ziehungen oder Replikationen vor. Die konkreten Schritte sind:

1. Konstruktion der empirischen Verteilungsfunktion $\hat{F}$, die jeden Datenpunkt $x_k$ mit Wahrscheinlichkeit $1/n$ gewichtet.

2. Ziehen einer Bootstrapstichprobe durch $n$-maliges unabhängiges Ziehen mit zurücklegen aus $\hat{F}$. Dadurch erhält man eine Bootstraprealisation $x^{*i} = (x_1^{*i}, \ldots, x_n^{*i})$.

3. Berechnung von $R\left(x^{*i}, \hat{F}\right)$.

4. $B$-malige Wiederholung der Schritte 2 und 3.

5. Approximation der Verteilung von $R\left(X^*, \hat{F}\right)$ durch das Histogramm der entsprechenden Werte $R\left(x^{*1}, \hat{F}\right), \ldots, R\left(x^{*B}, \hat{F}\right)$ und Approximation der Verteilungsparameter durch die entsprechenden empirischen Werte.

### 2.2.4 Bootstrap-Konfidenzintervalle

Eine ursprüngliche Motivation für die Konstruktion von Schätzern mittels der Bootstrapmethode war, ausgehend vom Schätzer für die Standardabweichung eines Schätzers $\hat{\theta}(X_1, \ldots, X_n)$, ein Konfidenzintervall für den zu schätzenden Parameter $\theta$ zu konstruieren. Die Bootstrapmethode bietet dafür spezielle Verfahren an, die auch unter allgemeinen Voraussetzungen gute Resultate liefern.

Im Abschnitt 3.2 wird eine Methode vorgestellt, die auf einem parametrischen Bootstrapverfahren basiert. Die zugrundeliegende Methodik wird dort erklärt werden.

Neben der in Kapitel 3 verwendeten Methode gibt es noch eine Vielzahl anderer Verfahren, über die Bootstrap-Methode ein Konfidenzintervall für einen Schätzer zu konstruieren (siehe EFRON (1987, 1990)).

## 2.3 Vergleich der Schätzer bei der zweiparametrischen Lognormalverteilung

Wie schon im Abschnitt 2.1 erwähnt, wird der Vergleich der Schätzmethoden auf verschiedenen Ebenen vorgenommen.

1. Es wird das Verhalten der Schätzer untersucht, wenn Momente der Verteilung geschätzt werden sollen. Wir werden den Erwartungswert und die Varianz herausgreifen und analysieren.

2. Das Verhalten der Schätzmethoden, wenn nicht Momente der Verteilung, sondern Verteilungsparameter geschätzt werden sollen, ist vor allem bei der Bestimmung von Funktionen lognormalverteilter Zufallsvariabler (Kapitel 4) und bei der Konstruktion von Konfidenzintervallen für Schätzer (Kapitel 3) von besonderem Interesse und wird deshalb hier untersucht. Besonders für den dreiparametrischen Fall ergeben sich interessante Ergebnisse. Die Analyse des zweiparametrischen Falls wird nur der Vollständigkeit wegen angegeben.

### 2.3.1 Schätzer für den Erwartungswert und die Standardabweichung

Die zu vergleichenden Schätzer sind (siehe Kapitel 1):

1. Momentenschätzer
2. Maximum-Likelihood-Schätzer
3. UMVU-Schätzer

#### Monte-Carlo-Simulation

Durch die Monte-Carlo-Simulation soll untersucht werden, ob und wie das Verhalten der Schätzmethoden vom Stichprobenumfang und von der Schiefe der Ausgangsverteilung abhängt. Es wurde deshalb folgende Vorgehensweise gewählt.

1. Es werden vier Stichprobenumfänge betrachtet:
   - $n = 10$
   - $n = 20$
   - $n = 50$
   - $n = 100$

2. Zusätzlich zu den Stichprobenumfängen wird die Schiefe der Verteilung variiert. Dafür wurde gewählt:
   - $\alpha_3 = 0.2$
   - $\alpha_3 = 0.5$

- $\alpha_3 = 1.0$
- $\alpha_3 = 2.0$
- $\alpha_3 = 4.0$

3. Durch die Wahl von $\alpha_3$ ist auch der Verteilungsparameter $\sigma$ der Grundgesamtheit festgelegt. Setzt man zusätzlich $E(X) = 10$, dann ist auch der zweite Verteilungsparameter $\mu$ der Grundgesamtheit festgelegt.

4. Erzeugung von 100 Zufallsstichproben vom Umfang $n$ aus einer lognormalverteilten Grundgesamtheit mit den Parametern $\mu$ und $\sigma^2$ (mit `Mathematica`). Dieser Schritt wird für alle 20 möglichen Kombinationen (bzgl. des Stichprobenumfangs und der Schiefe der Grundgesamtheit) durchgeführt.

5. Berechnung der Schätzungen für jede dieser Zufallsstichproben mit den zu untersuchenden Schätzmethoden (jeweils für den Erwartungswert und die Standardabweichung).

6. Berechnung des empirischen Mittels und der empirischen Standardabweichung der Schätzungen.

Durch die Festlegung des Erwartungswertes und der Schiefe ergeben sich folgende Verteilungsparameter und Momente für die Populationen, aus denen die Stichproben zu ziehen sind:

| $\alpha_3$ | $E(X)$ | $\sqrt{Var(X)}$ | $\mu$ | $\sigma$ |
|---|---|---|---|---|
| 0.2 | 10 | 0.66568 | 2.30037 | 0.06650 |
| 0.5 | 10 | 1.65165 | 2.28913 | 0.16406 |
| 1.0 | 10 | 3.22184 | 2.25320 | 0.31426 |
| 2.0 | 10 | 5.96070 | 2.15057 | 0.55138 |
| 4.0 | 10 | 9.99987 | 1.95601 | 0.83255 |

Tabelle 2.1: Verteilungsmomente und -parameter

**Ergebnisse**

Bei der Schätzung des Erwartungswertes liefern alle drei Schätzmethoden annähernd gleiche Ergebnisse, unabhängig vom Stichprobenumfang und der Schiefe der zugrundeliegenden Verteilung. Es ergaben sich beispielsweise für die Stichproben aus einer lognormalverteilten Grundgesamtheit mit $\alpha_3 = 0.2$ und dem Erwartungswert 10 erst in der vierten Dezimalstelle Abweichungen in den Schätzungen, bei $\alpha_3 = 2.0$ immerhin erst in der zweiten Dezimalstelle. Auch bei den empirischen Standardabweichungen der Schätzungen für den Erwartungswert waren keine Unterschiede festzustellen.

Bei der Schätzung der Standardabweichung konnten folgende Tendenzen beobachtet werden:

1. Für kleine Stichprobenumfänge ($n < 20$) und geringer Schiefe ($\alpha_3 < 1$) weichen die ML-Schätzungen von den beiden anderen ab: Die wahre Standardabweichung wird unterschätzt, der Grad dieser Abweichung ist jedoch gering.

2. Auffälliger hingegen reagiert die Momentenmethode bei Stichproben aus Populationen, die extrem rechtsschief verteilt sind ($\alpha_3 > 2$). Die Schätzungen weichen deutlich von den Populationswerten ab.

Diese Ergebnisse werden nun mit Hilfe der Bootstrap-Methode näher untersucht.

#### Bootstrap-Monte-Carlo-Simulation

Aus den Populationen, die zuvor untersucht worden sind, wurden jene mit $\alpha_3 = 0.2$, $\alpha_3 = 1.0$ und $\alpha_3 = 4.0$ analysiert. Aus diesen drei Grundgesamtheiten wurde jeweils eine Stichprobe vom Umfang 10 und jeweils eine vom Umfang 50 gezogen. Die Anzahl der Bootstrapreplikationen wurde mit 500 festgesetzt.

<u>$\alpha_3 = 0.2$, $n = 50$</u>

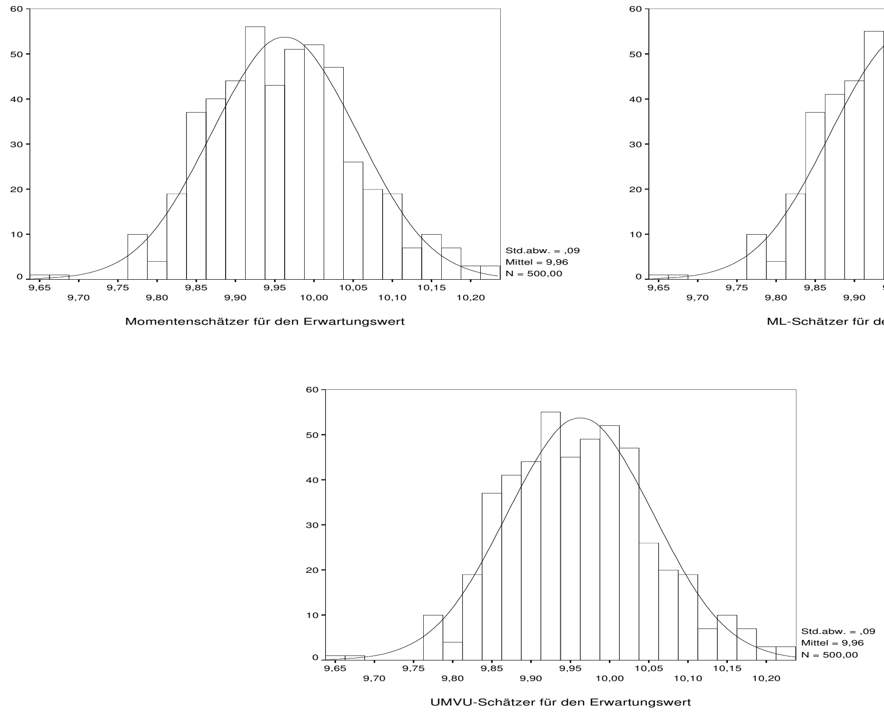



Abbildung 2.1: Histogramm der Schätzungen für den Erwartungswert; $\alpha_3 = 0.2, n = 50$, $E(X) = 10$

In Abbildung 2.1 sind die Histogramme der Schätzungen für den Erwartungswert angeführt. Es sind keine Unterschiede in den Dichteschätzern erkennbar. Dieser Eindruck wird auch von dem Boxplot und dem Matrixscatterplot (Abbildung 2.2) bestätigt. Der

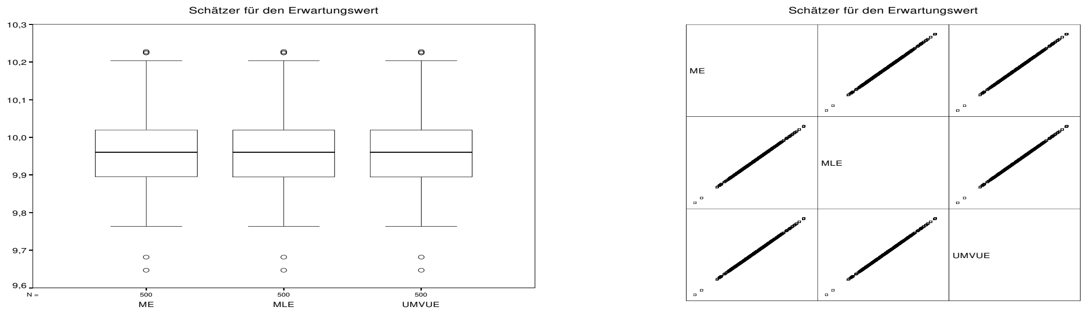



Abbildung 2.2: links: Boxplot der Schätzungen für den Erwartungswert, rechts: Matrixscatterplot der Schätzungen; $\alpha_3 = 0.2$, $n = 50$, $E(X) = 10$

Matrixscatterplot verdeutlicht die erwartet starke Korrelation zwischen den Ergebnissen der einzelnen Schätzmethoden bei einer Population mit geringer Schiefe.

Auch bei den Schätzungen für die Standardabweichung sind keine Unterschiede feststellbar (siehe Abbildung 2.3 und 2.4). Der Matrixscatterplot zeigt wiederum eine starke Korrelation zwischen den jeweilige Schätzungen, wobei die Schätzungen aus der UMVU-Methode und der ML-Methode vollständig korreliert sind. Diese starke Abhängigkeit ist dadurch erklärt, daß beide Methoden auf denselben Statistiken basieren.

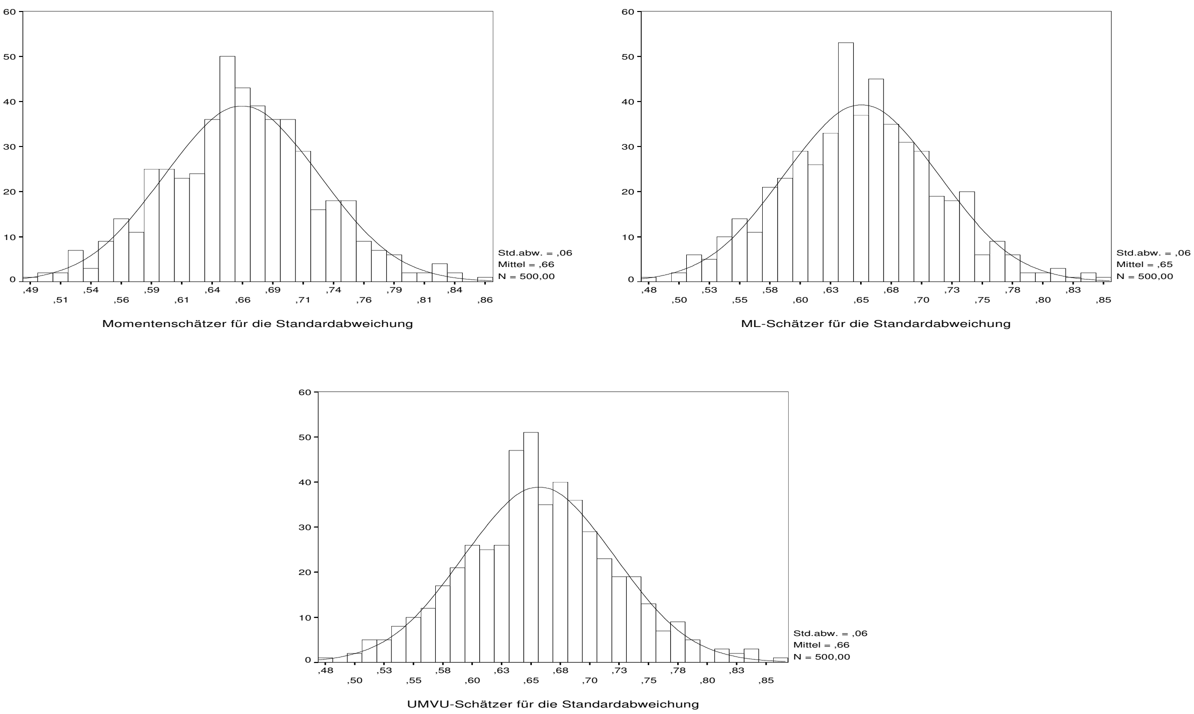



Abbildung 2.3: Histogramm der Schätzungen für die Standardabweichung; $\alpha_3 = 0.2$, $n = 50$, $SD(X) = 0.67$

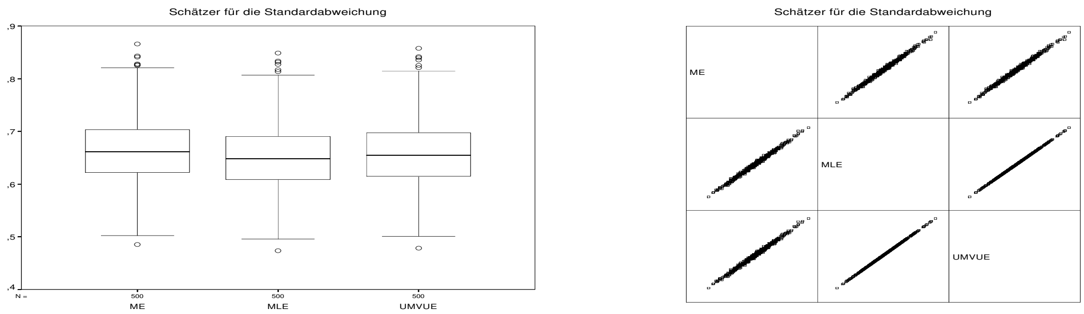



Abbildung 2.4: links: Boxplot der Schätzungen für die Standardabweichung, rechts: Matrixscatterplot der Schätzungen; $\alpha_3 = 0.2$, $n = 50$, $SD(X) = 0.67$

$\underline{\alpha_3 = 1.0,\ n = 50}$

Wie schon zuvor sind bei den Schätzungen für den Erwartungswert keine Unterschiede zwischen den Methoden feststellbar (Abbildung 2.5 und 2.6). Während die ML-Methode

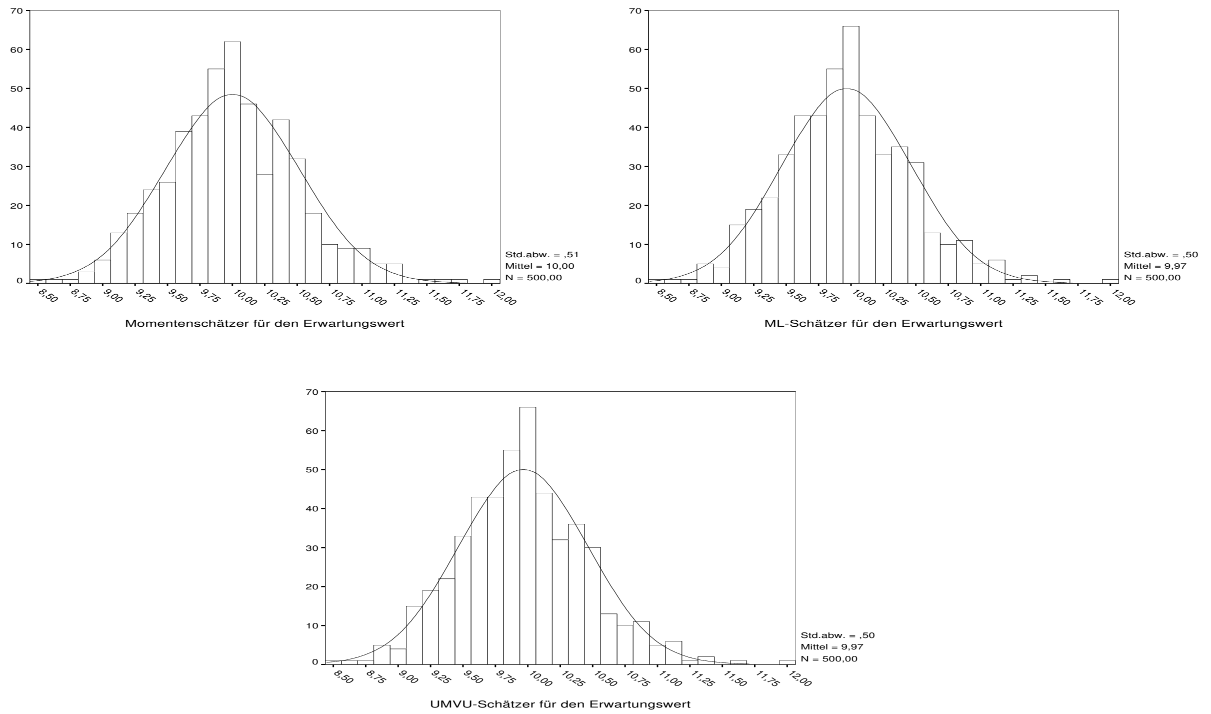



Abbildung 2.5: Histogramm der Schätzungen für den Erwartungswert; $\alpha_3 = 1.0$, $n = 50$, $E(X) = 10$

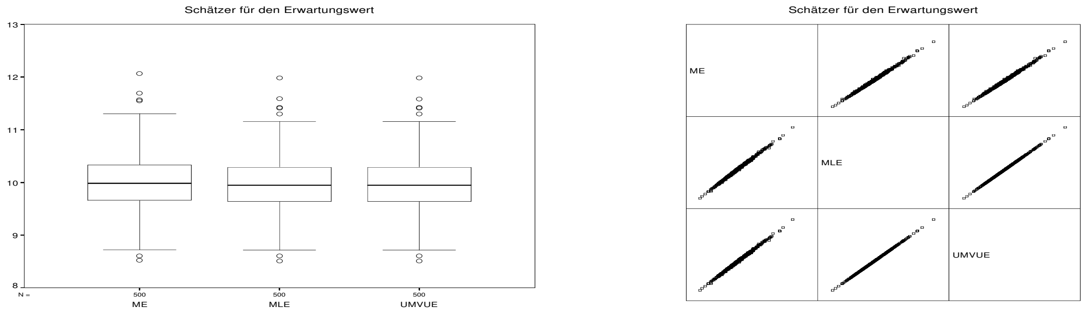



Abbildung 2.6: links: Boxplot der Schätzungen für den Erwartungswert, rechts: Matrixscatterplot der Schätzungen; $\alpha_3 = 1.0$, $n = 50$, $E(X) = 10$

und die UMVU-Methode bei der Schätzung der Standardabweichung wiederum annähernd dieselben Ergebnisse liefern, weichen die Schätzungen der Momentenmethode deutlich von den beiden anderen ab. Der wahre Wert $SD(X) = 3.22$ wird überschätzt. Das Histogramm (Abbildung 2.7) zeigt die Abhängigkeit der Momentenmethode von der 'Zusammenstellung' der Stichprobe, die von den beiden anderen Methoden ausgeglichen zu

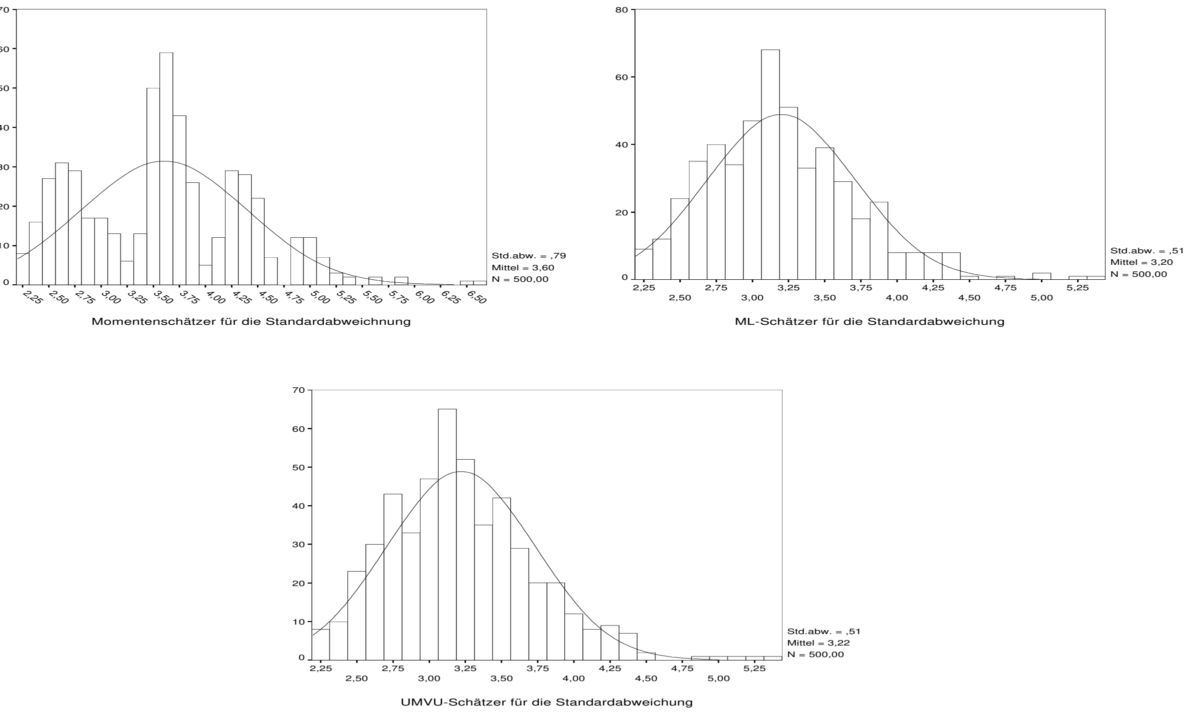



Abbildung 2.7: Histogramm der Schätzungen für die Standardabweichung; $\alpha_3 = 1.0$, $n = 50$, $SD(X) = 3.22$

werden scheint. Durch diese Abhängigkeit von der Stichprobe ist auch die größere Varianz der Schätzungen der Momentenmethode zu erklären. Die Dichteschätzer der Schätzungen für die Standardabweichung aus der ML-Methode und der UMVU-Methode zeigen erstmals rechtsschiefe Tendenzen.

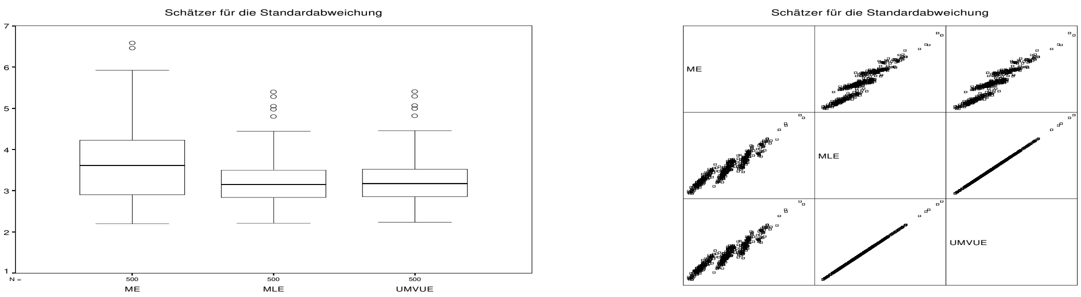



Abbildung 2.8: links: Boxplot der Schätzungen für die Standardabweichung, rechts: Matrixscatterplot der Schätzungen; $\alpha_3 = 1.0$, $n = 50$, $SD(X) = 3.22$

$\alpha_3 = 4.0$**,** $n = 50$

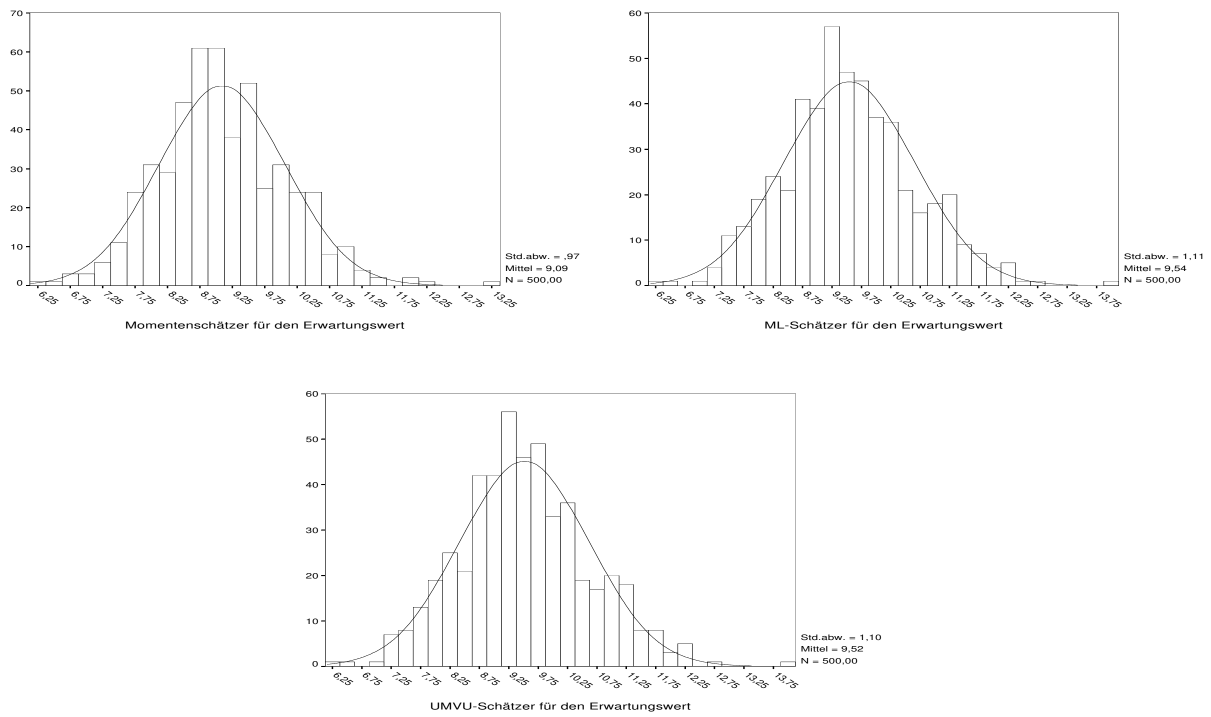



Abbildung 2.9: Histogramm der Schätzungen für den Erwartungswert; $\alpha_3 = 4.0$, $n = 50$, $E(X) = 10$

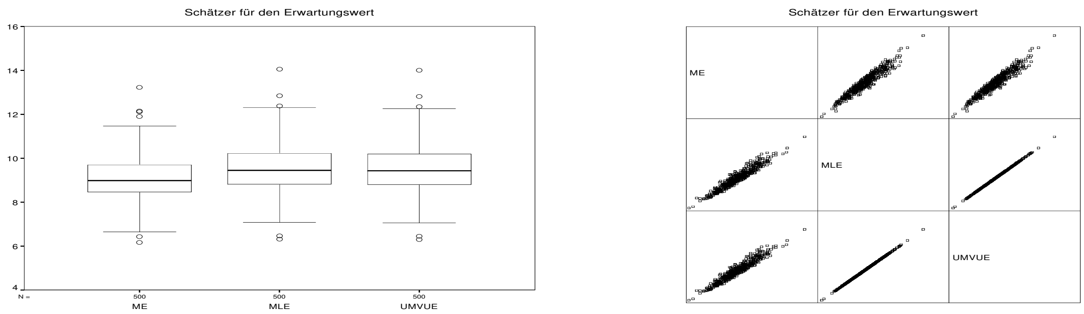



Abbildung 2.10: links: Boxplot der Schätzungen für den Erwartungswert, rechts: Matrixscatterplot der Schätzungen; $\alpha_3 = 4.0$, $n = 50$, $E(X) = 10$

Bei den Schätzungen des Erwartungswertes ergibt sich ein ähnliches Bild wie zuvor: Die UMVU- und ML-Methode liefern sehr ähnliche Resultate, die Momentenmethode weicht etwas von den beiden anderen ab. Zu erwähnen wäre, daß erstmals der wahre Wert nicht mehr in der 50%-Box um den Median der Schätzungen aus der Momentenmethode liegt. Diese Aussage ist jedoch zu relativieren, da auch bei den anderen beiden Methoden der wahre Wert an der oberen Grenze der Box liegt (siehe Abbildung 2.10 links). Bei den Schätzungen für die Standardabweichung sind erstmals gravierende Unterschiede zwischen den Methoden feststellbar: Die Schätzungen aus der Momentenmethode weichen

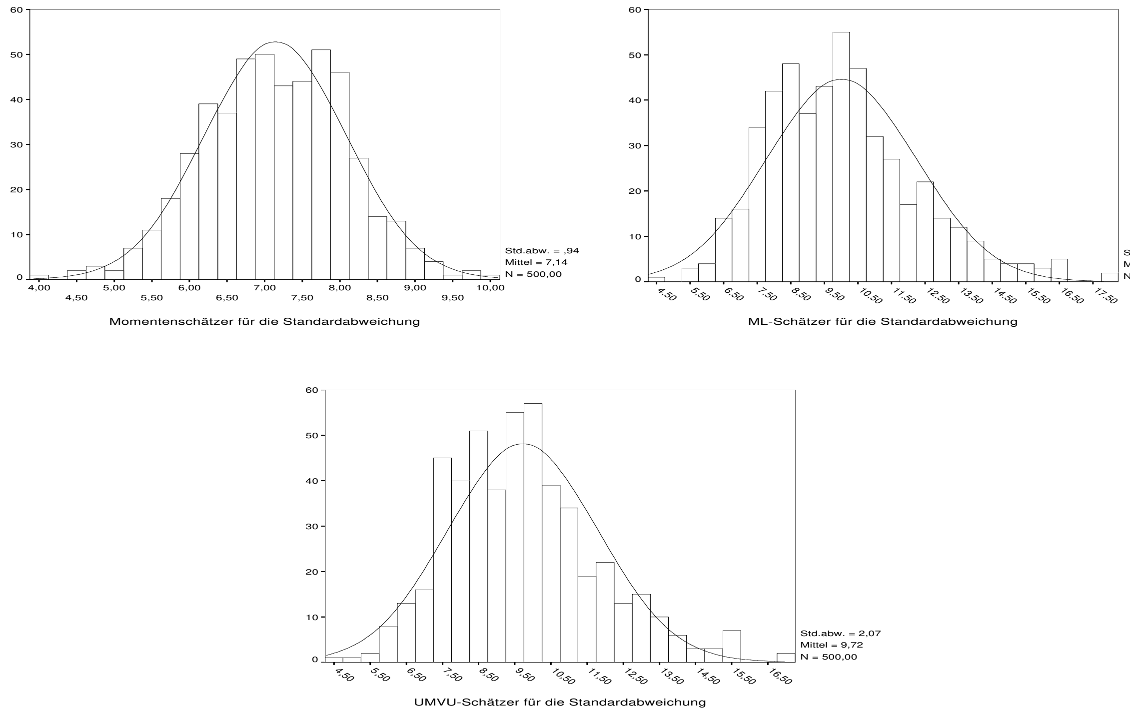



Abbildung 2.11: Histogramm der Schätzungen für die Standardabweichung; $\alpha_3 = 4.0$, $n = 50$, $SD(X) = 10$

stark von den Schätzungen aus der UMVU- und ML-Methode, die wiederum sehr ähnlich sind, ab. Während bei der UMVU- und ML-Methode aufgrund der stark rechtsschief verteilten Ausgangspopulation ($\alpha_3 = 4$) auch die Verteilung der Schätzer rechtsschief ist, tritt bei der Momentenmethode beinahe der konträre Fall, eine linksschiefe Verteilung des Schätzers, ein. Der wahre Wert der Standardabweichung wird deutlich unterschätzt; wäre der wahre Populationswert eine der Schätzungen, so würde sie bei der Box-Plot-Analyse als Extremwert ausgewiesen werden. Die beiden anderen Methoden liefern sehr gute Schätzungen.

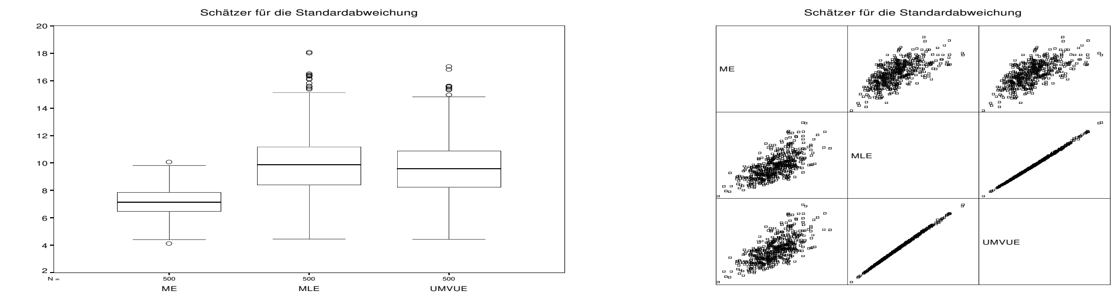



Abbildung 2.12: links: Boxplot der Schätzungen für die Standardabweichung, rechts: Matrixscatterplot der Schätzungen; $\alpha_3 = 4.0$, $n = 50$, $SD(X) = 10$

Das Verhalten der Momentenmethode bei einer stark rechtsschief verteilten Ausgangspopulation wurde nun hinreichend untersucht. Der zweite Punkt, der bei der Monte-Carlo-Simulation beobachtet wurde, das Verhalten des ML-Schätzers für die Standardabweichung bei geringer Schiefe und geringem Stichprobenumfang, soll jetzt untersucht werden. Bei dem Vergleich der Schätzungen für den Erwartungswert ergaben sich die selben Resultate wie zuvor bei den Stichproben vom Umfang 50. In Abbildung 2.13 (links)

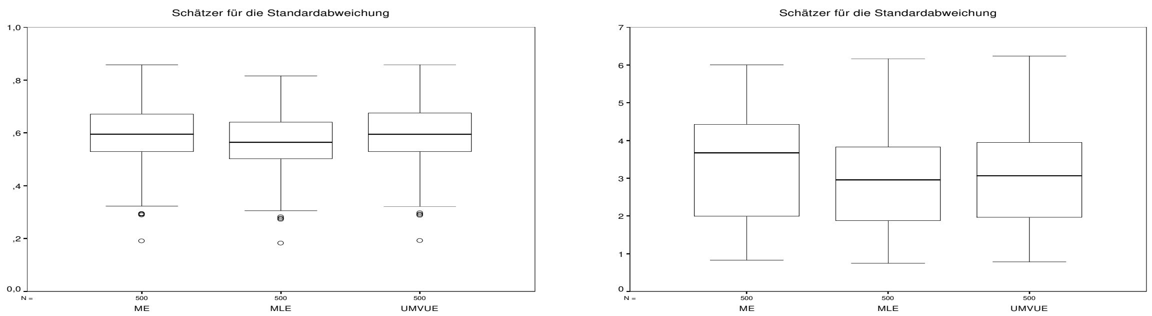



Abbildung 2.13: Boxplot der Schätzungen für die Standardabweichung; links: $\alpha_3 = 0.2$, $n = 10$, $SD(X) = 0.67$; rechts: $\alpha_3 = 1.0$, $n = 10$, $SD(X) = 3.22$

sieht man die geringe Abweichung der ML-Methode von den beiden anderen, die mit wachsendem $\alpha_3$ (siehe Abbildung 2.13 (rechts)) und/oder mit wachsendem $n$ (siehe Abbildung 2.4) abnimmt. Die Abweichung von den anderen Schätzern beziehungsweise vom Populationswert sind allerdings sehr gering.

### 2.3.2 Schätzer für die Verteilungsparameter

Eine Schätzung der Verteilungsparameter $\mu$ und $\sigma^2$ wird bei der Bestimmung von Funktionen lognormalverteilter Zufallsvariabler (Kapitel 4) und der Konstruktion von Konfidenzintervallen für einen Schätzer (Kapitel 3) benötigt. Es wurden analog zu vorigem Abschnitt Zufallsstichproben von unterschiedlichem Umfang aus fünf verschiedenen Grundgesamtheiten (siehe Tabelle 2.1) gezogen.

Im folgenden werden die Momentenschätzer und ML-Schätzer für $\mu$ und $\sigma^2$ miteinander verglichen. Der UMVU-Schätzer bezieht sich ausschließlich auf Funktionen der Form $\sigma^{2c}e^{a\mu+b\sigma^2}$ und kann daher nicht für die Schätzung der Verteilungsparameter verwendet werden.

Die Methodik bleibt dieselbe wie zuvor: Zuerst wird eine Monte-Carlo-Simulation (Anzahl der Runs = 100) durchgeführt, anschließend werden 'Auffälligkeiten' mit der Bootstrapmethode (Anzahl der Replikationen = 500) analysiert.

#### Monte-Carlo-Simulation

Es bestätigen sich im wesentlichen die Aussagen vom Vergleich der Schätzmethoden für die Verteilungsmomente. Auffällig ist wiederum das Verhalten des Momentenschätzers bei Stichproben aus einer extrem rechtsschief verteilten Population. Bei den Stichproben aus einer Lognormalverteilung mit $\alpha_3 = 2$ weichen die Schätzungen von $\mu$ und $\sigma$ deutlich von dem Populationswert ab, dieser Effekt verstärkt sich mit steigender Schiefe.

### Bootstrap-Monte-Carlo-Simulation

Bei Zufallsstichproben aus Populationen mit kleiner Schiefe sind keine Unterschiede in den approximativen Verteilungen der Schätzer erkennbar. Je größer die Schiefe der zugrundeliegenden Verteilung wird, umso instabiler verhält sich die Momentenmethode. Im folgenden werden die approximativen Bootstrapverteilungen der Momentenschätzer und ML-Schätzer für die Verteilungsparameter bzgl. der Populationen mit $\alpha_3 = 0.2$ und $\alpha_3 = 4.0$ analysiert, um einerseits die Äquivalenz der beiden Methoden bei Stichproben aus lognormalverteilten Populationen mit geringer Schiefe und andererseits die Diskrepanz zwischen den Methoden bei extrem rechtsschief verteilten Daten zu unterstreichen.

$\alpha_3 = 0.2$, $n = 50$, $\mu = 2.3$, $\sigma = 0.0665$

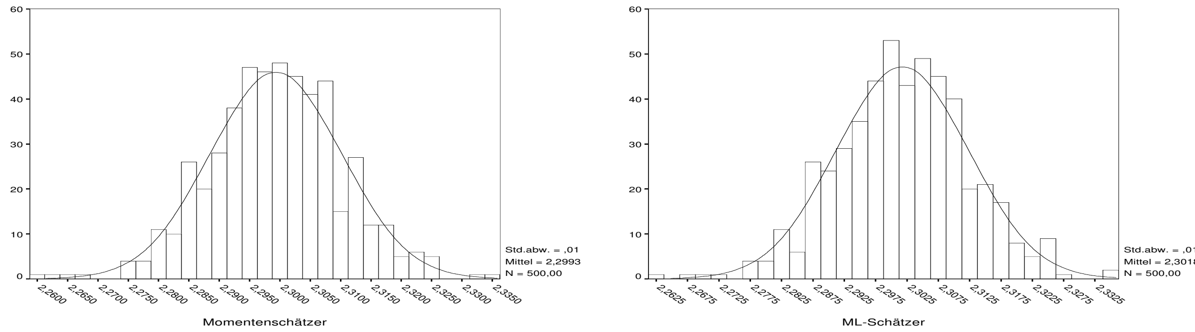



Abbildung 2.14: Histogramm der Schätzungen für $\mu = 2.3$; $\alpha_3 = 0.2$, $n = 50$

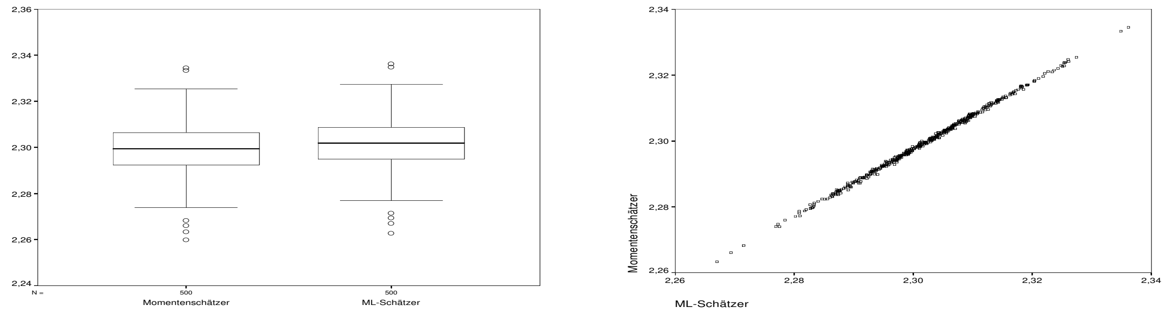



Abbildung 2.15: links: Boxplot der Schätzungen für $\mu = 2.3$, rechts: Scatterplot der Schätzungen für $\mu$; $\alpha_3 = 0.2$, $n = 50$

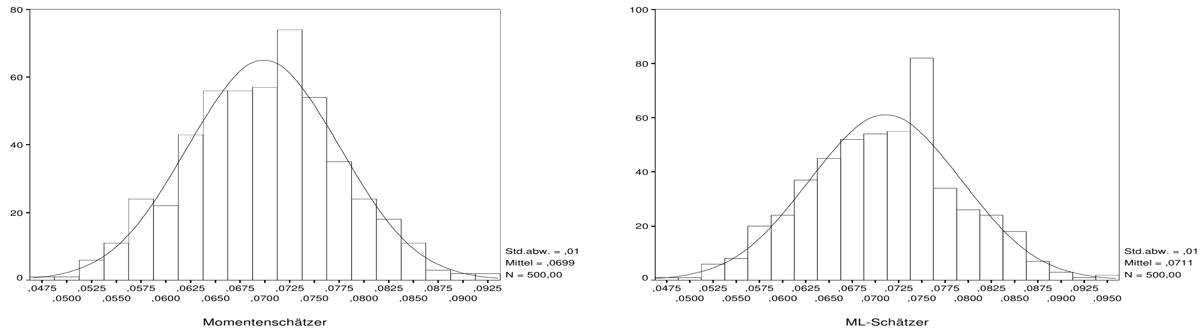



Abbildung 2.16: Histogramm der Schätzungen für $\sigma = 0.067$; $\alpha_3 = 0.2$, $n = 50$

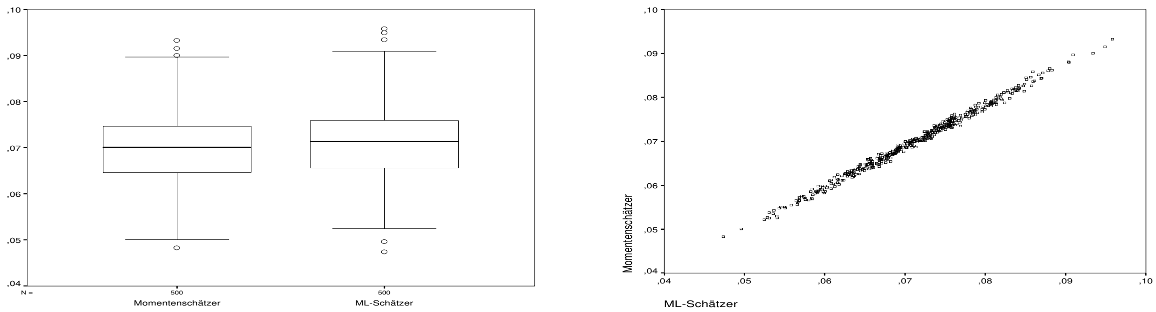



Abbildung 2.17: links: Boxplot der Schätzungen für $\sigma = 0.067$, rechts: Scatterplot der Schätzungen für $\sigma$; $\alpha_3 = 0.2$, $n = 50$

Beide Schätzmethoden liefern für $\mu$ und $\sigma$ annähernd gleiche Resultate, die nahe bei den Populationswerten liegen.

$\underline{\alpha_3 = 4.0,\ n = 50,\ \mu = 1.95,\ \sigma = 0.833}$

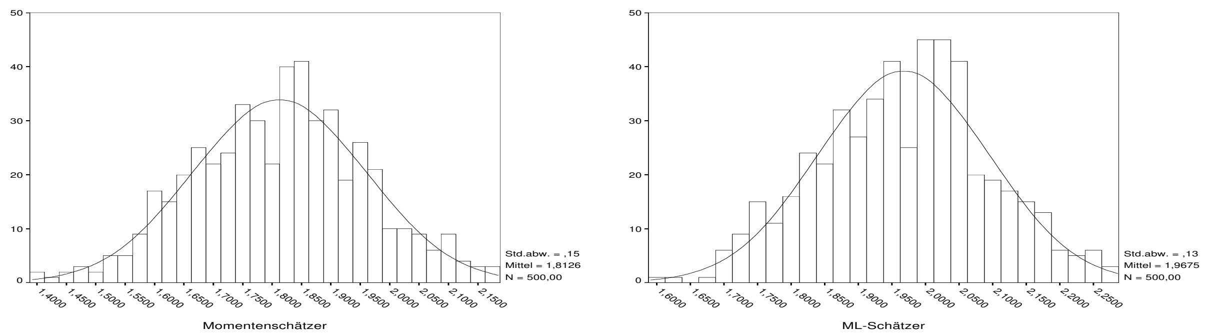



Abbildung 2.18: Histogramm der Schätzungen für $\mu = 1.95$; $\alpha_3 = 4.0$, $n = 50$

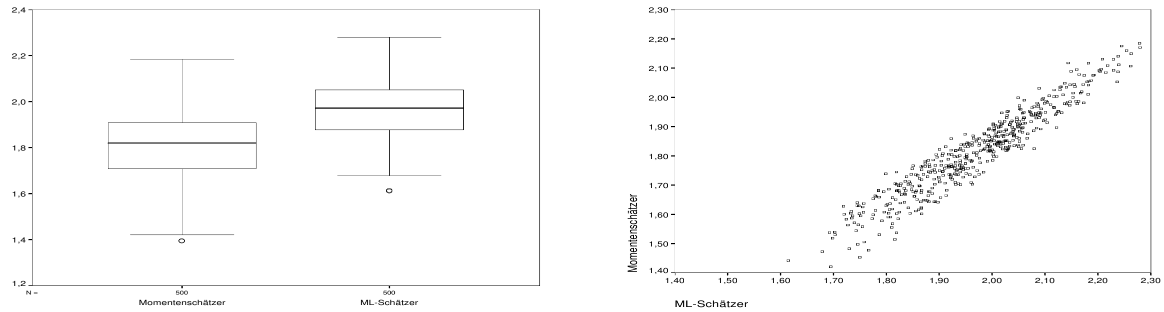



Abbildung 2.19: links: Boxplot der Schätzungen für $\mu = 1.95$, rechts: Scatterplot der Schätzungen für $\mu$; $\alpha_3 = 4.0$, $n = 50$

Die Stichprobe wurde aus einer lognormalverteilten Grundgesamtheit mit $\mu = 1.95061$ und $\sigma = 0.83255$ gezogen. Sowohl $\mu$ als auch $\sigma$ werden mit der Momentenmethode deutlich unterschätzt. Die Abbildungen 2.18 - 2.21 zeigen die zugehörenden Histogramme, Scatter- und Boxplots der einzelnen Schätzungen. Die Scatterplots zeigen, daß die Schätzungen der Momentenmethode größtenteils unter den Schätzungen der ML-Methode liegen. Interessant ist auch der Boxplot der Schätzungen für $\sigma$ (Abbildung 2.21):

1. Weit mehr als 75% der Schätzungen der Momentenmethode liegen unter dem wahren Wert $\sigma$.

2. Die Schätzungen aus der Momentenmethode liegen kompakter beieinander als die Schätzungen aus der ML-Methode, d.h. $\sigma$ wird systematisch unterschätzt.

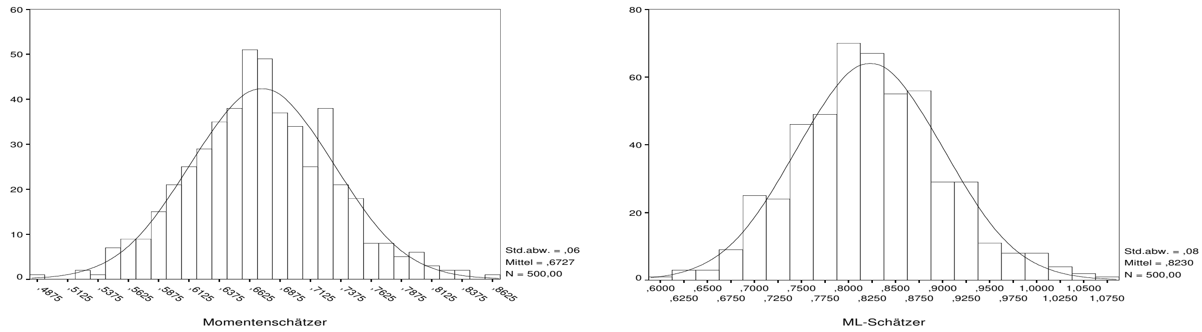



Abbildung 2.20: Histogramm der Schätzungen für $\sigma = 0.83$; $\alpha_3 = 4.0$, $n = 50$

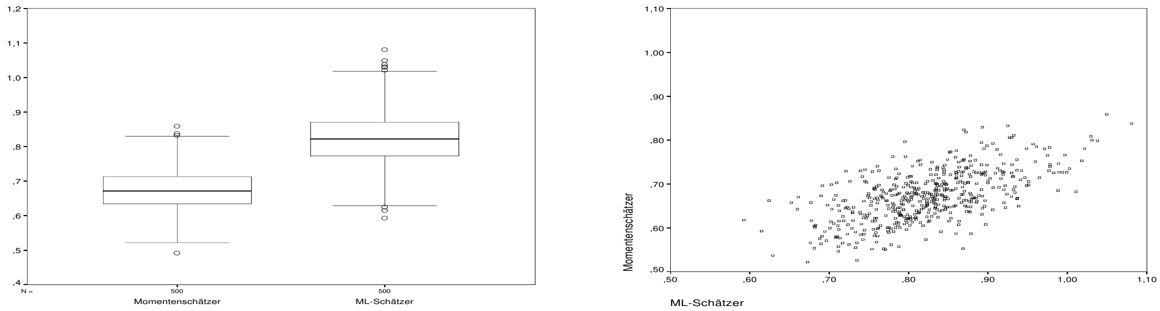



Abbildung 2.21: links: Boxplot der Schätzungen für $\sigma = 0.83$, rechts: Scatterplot der Schätzungen für $\sigma$; $\alpha_3 = 4.0$, $n = 50$

Abschließend können folgende Empfehlungen zur Verwendung der verschiedenen Schätzmethoden gegeben werden:

1. $\alpha_3 < 1$: UMVU-, ML-Methode oder Momentenmethode
2. $\alpha_3 \geq 1$: UMVU- oder ML-Methode

## 2.4 Vergleich der Schätzer bei der dreiparametrischen Lognormalverteilung

COHEN and WHITTEN (1980) veröffentlichten eine Simulationsstudie zum Vergleich der in Kapitel 1 vorgestellten Schätzmethoden für die Parameter einer Stichprobe aus einer dreiparametrisch lognormalverteilten Gundgesamtheit.

Sie verglichen die Ergebnisse der verschiedenen Methoden anhand einer Stichprobe vom Umfang 20 und führten anschließend eine Monte-Carlo-Simulation von 100 Stichproben vom Umfang 100 aus vier verschiedenen Grundgesamtheiten ($\alpha_3 = 0.301, = 0.980, = 1.625, = 2.475$) mit dem Erwartungswert 0 und der Varianz 1 durch. Die Ergebnisse werden im folgenden kurz wiedergegeben.

Ausgehend von der Stichprobe ermittelte sich Cohen Schätzungen für die Verteilungs-

| | | | |
|---|---|---|---|
| 148.290 | 144.328 | 174.800 | 168.554 |
| 184.101 | 166.475 | 131.375 | 145.788 |
| 135.880 | 137.338 | 164.304 | 155.369 |
| 127.211 | 132.971 | 128.709 | 201.415 |
| 133.143 | 155.680 | 153.070 | 157.238 |

Tabelle 2.2: Stichprobe aus einer dreiparametrisch lognormalverteilten Grundgesamtheit

parameter, für den Erwartungswert, die Varianz und die Schiefe mit folgenden Methoden:

1. ML-Methode(MLE)
2. modifizierte ML-Methoden:
   (a) MMLE-I Methode für $r$-te Ordnungsstatistik; $r$=1,..,3.
   (b) MMLE-II Methode für $r$-te Ordnungsstatistik; $r$=1,..,3.
3. Momentenmethode (ME)
4. modifizierte Momentenmethoden
   (a) MME-I Methode für $r$-te Ordnungsstatistik; $r$=1,..,3.
   (b) MME-II Methode für $r$-te Ordnungsstatistik; $r$=1,..,3.

Die Ergebnisse sind in Tabelle 2.3 angeführt (Erwartungswert=E, Standardabweichung=SD).

| Schätzmethode | $\tau$ | $\mu$ | $\sigma$ | E | SD | Schiefe |
|---|---|---|---|---|---|---|
| $MLE$ | 117.7206 | 3.3732 | 0.6043 | 152.7353 | 23.2471 | 2.2844 |
| $MMLE - I_1$ | 107.0445 | 3.7211 | 0.4297 | 152.3519 | 20.4039 | 1.4424 |
| $MMLE - I_2$ | 19.1641 | 4.8812 | 0.1412 | 152.2738 | 18.8876 | 0.4285 |
| $MMLE - I_3$ | 126.4777 | 2.8077 | 1.1418 | 158.2828 | 52.0992 | 9.3097 |
| $MMLE - II_1$ | 118.0314 | 3.3602 | 0.6122 | 152.7622 | 23.4184 | 2.3294 |
| $MMLE - II_2$ | 86.8243 | 4.1397 | 0.2876 | 152.2594 | 19.2133 | 0.9062 |
| $MMLE - II_3$ | 125.9937 | 2.8694 | 1.0491 | 156.5543 | 43.2821 | 7.0896 |
| $ME$ | 70.7067 | 4.3727 | 0.2410 | 152.3020 | 19.9537 | 0.7483 |
| $MME - I_1$ | 105.1667 | 3.7706 | 0.4060 | 152.3020 | 19.9537 | 1.3458 |
| $MME - I_2$ | 75.9154 | 4.3028 | 0.2569 | 152.3020 | 19.9537 | 0.8015 |
| $MME - I_3$ | -0.7037 | 5.0220 | 0.1299 | 152.3020 | 19.9537 | 0.3935 |
| $MME - II_1$ | 112.5581 | 3.5701 | 0.4741 | 152.3020 | 19.9537 | 1.6327 |
| $MME - II_2$ | 96.0955 | 3.9697 | 0.3445 | 152.3020 | 19.9537 | 1.1098 |
| $MME - II_3$ | 82.2976 | 4.2095 | 0.2795 | 152.3020 | 19.9537 | 0.8783 |
| *Population* | 100.0000 | 3.9120 | 0.4000 | 154.1644 | 22.5620 | 1.3219 |

Tabelle 2.3: Zusammenstellung der Schätzungen

Durch die Konstruktion des Momentenschätzers und der damit korrespondierenden modifizierten Schätzer ist klar: Unterschiede in den Schätzungen von Momenten werden erst im dritten (zentral oder um Null) auftreten; d.h. auch wenn die Verteilungsparameter unterschiedlich geschätzt werden, so sind die Schätzer für den Erwartungswert und die Varianz ident. Bei den ML-Methoden ist dies nicht der Fall, was durch die Definition der Schätzmethoden klar ist.

In Tabelle 2.3 sieht man, wie unterschiedlich die Schätzungen für die Verteilungsparameter, insbesondere für $\tau$, sind. Das wirft zwei Fragen auf:

1. Bei welcher Parameterkonstellation soll welche Schätzmethode verwendet werden?

2. Ist es überhaupt sinnvoll, eine gegebene Stichprobe durch ein dreiparametrisches Modell zu beschreiben, wenn die einzelnen Schätzmethoden derart inhomogene Ergebnisse liefern?

Wie bereits erwähnt, führten Cohen and Whitten eine Monte-Carlo-Simulation zum Vergleich der Schätzmethoden für die Verteilungsparameter durch. Es wurden dabei je 100 Stichproben vom Umfang 100 aus vier verschiedenen Populationen (Erwartungswert=0, Standardabweichung=1, Schiefe=0.301, =0.980, =1.625, =2.475) gezogen. Die Aussagen, die aufgrund der Simulation getroffen werden können, decken sich (bezüglich der Entscheidung ML- oder Momentenmethode) mit jenen, die im zweiparametrischen Fall gefunden wurden:

1. Ist $\alpha_3 < 1$, so sind die Momentenmethoden zu bevorzugen. Von den modifizierten Verfahren liefern jene, die auf der ersten Ordnungsstatistik basieren 'bessere' Ergebnisse. Cohen schlägt die $MME - II_1-$Methode vor. Es können aber auch $ME-$ oder $MME - I_1-$Schätzer verwendet werden.

2. Für $\alpha_3 \geq 1$ liefert die ML-Methode bessere Schätzungen. Auch hierbei sind bei den modifizierten Verfahren die auf der ersten Ordnungsstatistik basierenden den anderen vorzuziehen. Analog zu den Momentenschätzern schlägt Cohen die $MMLE - II_1$, $MMLE - I_1$ oder die $MLE-$Methode vor.

3. Die modifizierten Verfahren, die auf der zweiten oder dritten Ordnungsstatistik basieren, lieferten nicht mehr für alle der 100 generierten Stichproben Lösungen. Für die Momentenverfahren war dies vor allem bei geringer Schiefe und für die ML-Verfahren bei großem Schiefekoeffizienten, also genau im jeweiligen 'Anwendungsgebiet' der Fall. Das ist ein weiterer Grund dafür, auf die Verwendung dieser Methoden zu verzichten.

Aufbauend auf den Ergebnissen von Cohen wurden eigene Berechnungen angestellt. Es zeigte sich, daß sich bei den ML-Methoden bei geringem Stichprobenumfang und geringem Schiefekoeffizienten Probleme mit der iterativ zu lösenden Gleichung ergeben, während die korrespondierenden iterativ zu lösenden Gleichungen der ME-Methoden keine Probleme bereiteten.

Die Beantwortung der Frage nach der Sinnhaftigkeit des dreiparametrischen Modells zur Datenbeschreibung ist einerseits von der Problemstellung und andererseits von den Daten selbst (wie schief ist die emp. Verteilungsfunktion?) abhängig.

Ist man nur an Schätzungen für den Erwartungswert und für die Standardabweichung interessiert und ist die empirische Schiefe gering, so erhält man mit der Momentenmethode (2-parametrischer Ansatz) vernünftige Schätzungen (aufgrund der Definition der Momentenmethode im zwei- und dreiparametrischen Fall trivial).

Ist die empirische Schiefe jedoch größer, so wird (sowohl im zweiparametrischen als auch im dreiparametrischen Fall) die Verwendung des ML-Schätzers empfohlen. Betrachtet man sich die Schätzungen aus Tabelle 2.3, die die ML-Methoden liefern, so ist offensichtlich: Wird der treshold-Parameter $\tau$ unterschätzt, so wird $\mu$ überschätzt und somit $\sigma$ unterschätzt. Daraus resultiert eine Unterschätzung der Standardabweichung. Würde man auf einen Datensatz, der aus einer dreiparametrisch lognormalverteilten Grundgesamtheit stammt, die zweiparametrische ML-Methode anwenden, so bedeutet das: der treshold-Parameter wird als Null vorausgesetzt und somit über- oder unterschätzt. In Abhängigkeit von dem Wert von $\tau$ werden somit auch die anderen Verteilungsparameter und die Standardabweichung falsch geschätzt.

Der Bias der Schätzungen wurde in Abhängigkeit von der Größenordnung von $\tau$ untersucht. Dabei wurde je eine Stichprobe vom Umfang 100 aus einer normalverteilten Grundgesamtheit mit $(\mu, \sigma) = (0, 1)$ und $(\mu, \sigma) = (2, 2)$ generiert. Zu den exponentierten Werten wurde ein Skalar $\tau$ , $\tau = 0, \ldots, 99$, addiert, um Stichproben aus dreiparametrischen Lognormalverteilungen mit den Verteilungsparametern $(0 \ldots 99, \mu, \sigma^2)$ zu erhalten.

In den Abbildungen 2.22 und 2.23 sind die Schätzungen der dreiparametrischen ML-Methode (punktierte Linie), die der zweiparametrischen ML-Methode (strichliert) und die exakten Populationswerte (durchgezogene Linie) für die Verteilungsparameter $\mu$ und $\sigma$, für den Erwartungswert und für die Standardabweichung abgebildet. Dabei ist auf der $x$-Achse der Wert von $\tau$ und auf der $y$-Achse der entsprechende Schätzwert aufgetragen.

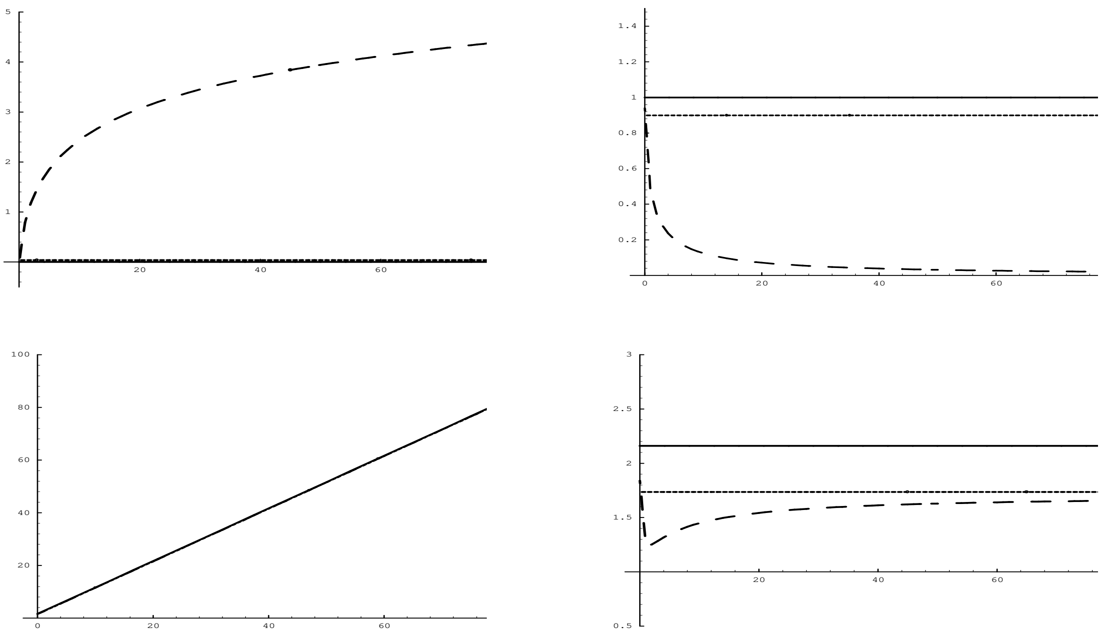



Abbildung 2.22: Vergleich zweiparametrische ML-Schätzer (strichliert), dreiparametrische ML-Schätzer (punktiert) und exakte Populationswerte (durchgezogen) für Stichprobe ($n = 100$) aus normalverteilter Grundgesamtheit mit $\mu = 0$ und $\sigma = 1$; links oben: Schätzer für $\mu$; rechts oben: Schätzer für $\sigma$; links unten: Schätzer für den Erwartungswert; rechts unten: Schätzer für die Standardabweichung.

Anmerkung: Jeder Wert $\tau = 0, .., 99$ wurde durch den dreiparametrischen ML-Schätzer im Fall der Stichprobe aus $\Lambda(\tau, 0, 1)$ um 0.28 unterschätzt und im Fall $\Lambda(\tau, 2, 2)$ um 0.034 überschätzt (unabhängig von der Größe des exakten Wertes $\tau$). Aus den Abbildungen 2.22 und 2.23 erkennt man:

1. Da der threshold-Parameter durch den dreiparametrischen ML-Schätzer unabhängig von der Größe immer gleich gut geschätzt wird, ergeben sich für alle gewählten Werte von $\tau$ dieselben Schätzungen für die Verteilungparameter und Verteilungsmomente (d.h. aufgrund der zugrundeliegenden Stichprobe aus der Normalverteilung werden die Charakteristiken entweder unter- oder überschätzt).

2. Bei der Schätzung der Verteilungsparameter versagen die (auf falschen) Verteilungsannahmen basierenden) zweiparametrischen ML-Schätzer natürlich. Auffallend ist, daß schon ein kleiner treshold-Parameter einen großen Fehler in der Schätzung verursacht. Der Bias wird mit wachsendem $\tau$ größer.

3. Die Standardabweichung wird durch den zweiparametrischen ML-Schätzer (bei positivem $\tau$) systematisch unterschätzt, während das bei der Schätzung des Erwartungswertes nicht zwingenderweise der Fall ist (es wurden Vergleiche mit anderen Stichproben durchgeführt, bei denen der Erwartungswert überschätzt wurde). Das wird auch durch die Definition des Schätzers klar: Der Einfluß des 'Unterschätzens'

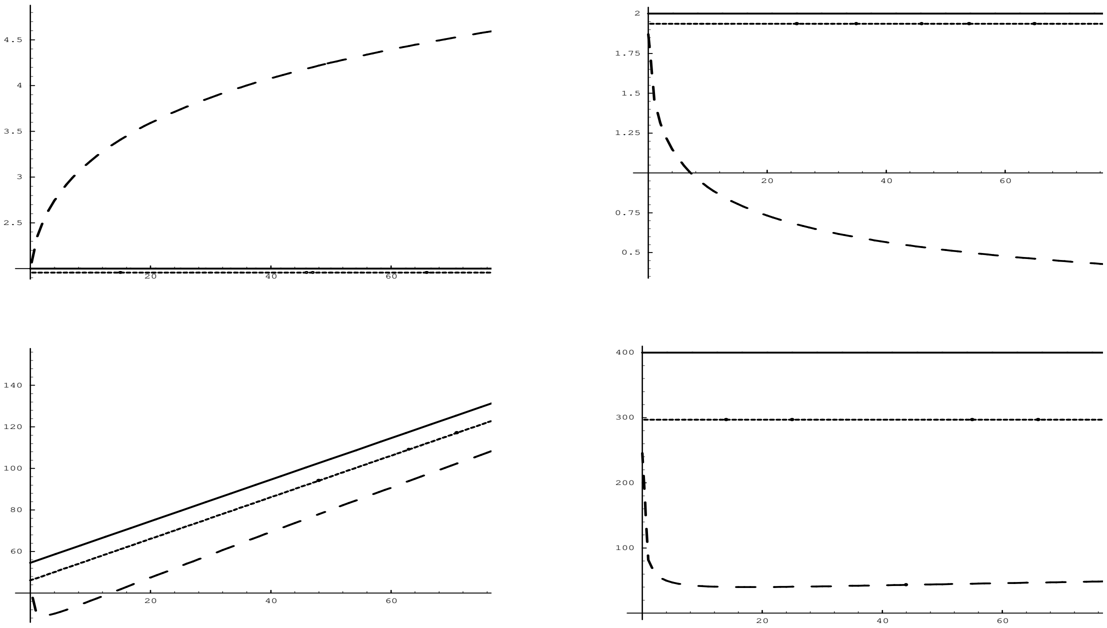



Abbildung 2.23: Vergleich zweiparametrische ML-Schätzer (strichliert), dreiparametrische ML-Schätzer (punktiert) und exakte Populationswerte (durchgezogen) für Stichprobe ($n = 100$) aus normalverteilter Grundgesamtheit mit $\mu = 2$ und $\sigma = 2$; links oben: Schätzer für $\mu$; rechts oben: Schätzer für $\sigma$; links unten: Schätzer für den Erwartungswert; rechts unten: Schätzer für die Standardabweichung.

des Verschiebungsparameters wird durch eine Überschätzung von $\mu$ ausgeglichen, wodurch aber wiederum $\sigma$ systematisch unterschätzt wird.

4. Der Einfluß des Parameters $\sigma$ der Grundgesamtheit wird deutlich. Ist $\sigma$ groß, so bewirkt eine kleine Abweichung in den Schätzungen von $\sigma$ eine große Abweichung in den Schätzungen der Standardabweichung.

Um den Vergleich der Schätzmethoden nicht nur in Abhängigkeit einer Stichprobe anzustellen, wurde eine Monte-Carlo-Simulation durchgeführt. Dabei wurde der treshold-Parameter, dessen Größe keinen Einfluß auf die Güte der dreiparametrischen Schätzungen hat, fixiert. Aus obigen Populationen wurden je 100 Stichproben vom Umfang 50 generiert (mit $\tau = 50$). Um den Einfluß der Schiefe (d.h. der Größe von $\sigma$) auf die Güte der Schätzungen der Verteilungsmomente zu demonstrieren, wurde noch 100 Stichproben (wiederum mit $n = 50$) aus der Population aus Tabelle 2.3 ($\tau = 100$, $\mu = 3.912$, $\sigma = 0.4$) generiert. Die Ergebnisse sind in den Tabellen 2.4 bis 2.6 angeführt.

| Schätzer | $\tau$ | | $\mu$ | | $\sigma$ | | E | | SD | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 2-param. | – | – | 3.94 | 0.01 | 0.03 | 0.01 | 51.63 | 0.32 | 1.77 | 0.63 |
| 3-param. | 50.03 | 0.09 | -0.07 | 0.21 | 1.04 | 0.19 | 51.68 | 0.35 | 2.46 | 1.13 |
| Population | 50.00 | – | 0.00 | – | 1.00 | – | 51.65 | – | 2.16 | – |

Tabelle 2.4: Zusammenstellung der Schätzungen; $\Lambda\,(50,0,1)$; $n = 50$, Runs=100

| Schätzer | $\tau$ | | $\mu$ | | $\sigma$ | | E | | SD | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 2-param. | – | – | 4.29 | 0.09 | 0.56 | 0.14 | 87.0 | 15.1 | 55.9 | 26.9 |
| 3-param. | 50.08 | 0.09 | 1.92 | 0.31 | 2.08 | 0.25 | 123.1 | 45.2 | 961.5 | 1265.9 |
| Population | 50.00 | – | 2.00 | – | 2.00 | – | 104.6 | – | 399.7 | – |

Tabelle 2.5: Zusammenstellung der Schätzungen; $\Lambda\,(50,2,4)$; $n = 50$, Runs=100

| Schätzer | $\tau$ | | $\mu$ | | $\sigma$ | | E | | SD | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 2-param. | – | – | 5.03 | 0.02 | 0.13 | 0.02 | 153.9 | 3.29 | 20.7 | 3.1 |
| 3-param. | 87.1 | 107.2 | 3.90 | 0.50 | 0.42 | 0.16 | 154.1 | 3.33 | 22.6 | 4.4 |
| Population | 100.0 | – | 3.91 | – | 0.40 | – | 154.2 | – | 22.6 | – |

Tabelle 2.6: Zusammenstellung der Schätzungen; $\Lambda\,(100,3.912,0.4)$; $n = 50$, Runs=100

Anmerkung zu Tabelle 2.6: Die große Streuung der Schätzungen von $\tau$ ist dadurch erklärbar, daß für zwei der 100 generierten Stichproben die iterativ zu lösende Bestimmungsgleichung zu den Werten 3.4 und -46.4 konvergierte.

Die Ergebnisse der Monte-Carlo-Simulation decken sich mit dem graphischen Vergleich zuvor: Bei (relativ) kleinen Werten von $\sigma$ liefern die dreiparametrischen Schätzer sehr gute Ergebnisse, der Bias der zweiparametrischen Methode ist auch gering (bezügl. der Schätzung von Erwartungswert und Varianz). Ist der Einfluß von $\sigma$ bei der Bestimmung der Standardabweichung groß (d.h. wenn $\sigma$ groß ist), so bewirken bereits kleine Abweichungen in der Schätzung von $\sigma$ große Abweichungen in der Schätzung der Standardabweichung.

Diese Untersuchungen wurden angestellt, um zu demonstrieren, wie sich die Verwendung des speziellen zweiparametrischen Lognormalverteilungsmodells auf Daten auswirkt, die aus einer dreiparametrisch lognormalverteilten Grundgesamtheit stammen. Alle Vergleiche gehen von der Voraussetzung aus, daß die beobachteten Daten aus einer dreiparametrisch lognormalverteilten Grundgesamtheit stammen und sind somit rein hypothetischer Natur.

In der Praxis wird das dreiparametrische Modell selten zur Datenbeschreibung eingesetzt, da die Bestimmung der Schätzungen und weitere Analysen (Konfidenzintervalle, Hypothesentests) aufwendig und kompliziert sind. Aus diesen Gründen wird in dem Fall, in dem offensichtlich nur verschobene Daten durch ein zweiparametrisches Modell beschrieben werden können (zum Beispiel bei negativen rechtsschiefen Daten), meist das Minimum der beobachteten Daten als Verschiebungs-Parameter verwendet. Im nächsten Abschnitt wird die Zulässigkeit dieser Vorgehensweise untersucht.

### 2.4.1 Der Praxisansatz: Shift um das Datenminimum

Bei Einsatz des dreiparametrischen Lognormalverteilungsmodells ergeben sich vor allem in weiterführenden Analysen Probleme durch den Einfluß des threshold-Parameters $\tau$. Um diese Probleme zu umgehen, wird in der Praxis meist nur ein Shift um das Stichprobenminimum (genauer $X_{(1)} - \epsilon$, $\epsilon > 0$) durchgeführt. In diesem Abschnitt wird die Güte dieser 'Approximation' für $\tau$ untersucht. Der Vergleich wird in zwei Schritten durchführt:

1. Theoretische Betrachtung des Erwartungswertes des Minimums einer Stichprobe aus einer lognormalverteilten Grundgesamtheit

2. Monte-Carlo-Simulation

Der Erwartungswert für die erste Ordnungsstatistik einer Stichprobe vom Umfang $n$ aus einer lognormalverteilten Grundgesamtheit kann über den Erwartungswert der ersten Ordnungsstatistik einer Stichprobe vom Umfang $n$ aus einer standardnormalverteilten Population berechnet werden (siehe HARTER (1961)). Der Erwartungswert des Minimums einer Stichprobe aus einer normalverteilten Grundgesamtheit ist vor allem vom Variationskoeffizenten der Population abhängig. Daher wurde der Bias der Schätzung für $\tau$ in Abhängigkeit vom Variationskoeffizienten $VC = \sqrt{\mu_2}/\mu_1^{'}$ der zugrundeliegenden Normalverteilung mit $VC \in (0, 1.5]$ untersucht. Unter Vorgabe von $VC$ wird für $\mu = 1, .., 4$ $\sigma = \mu VC$ bestimmt und der Erwartungswert der ersten Ordnungsstatistik berechnet. Abbildung 2.24 zeigt die Biasfunktion in Abhängigkeit vom Variationskoeffizienten der korrespondierenden Normalverteilung. Auf der $x$-Achse ist dabei $VC^2$ aufgetragen, auf der $y$-Achse der entsprechende Bias $(E\left(X_{(1)}\right) - \tau)$.

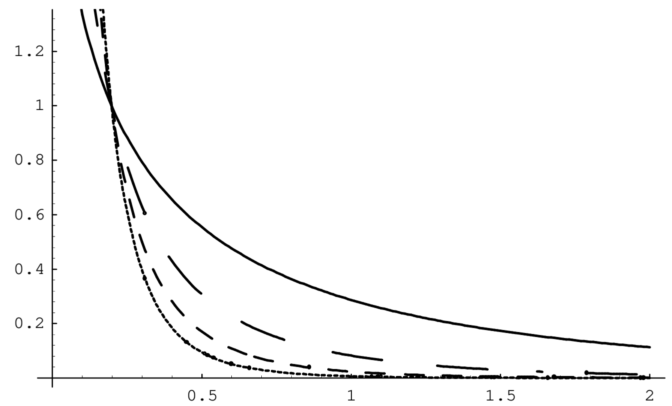



Abbildung 2.24: Biasfunktion der Schätzung für $\tau$ durch das Minimum; $\mu = 1$: durchzogene Linie, $\mu = 2$: grob stichliert, $\mu = 3$: fein stichliert, $\mu = 4$: punktiert

Natürlich wird $\tau$ durch das Stichprobenminimum überschätzt. Mit wachsendem Variationskoeffizienten und $\mu$ (bezügl. der korrespondierenden Normalverteilung) wird der Bias geringer.

Um die Auswirkung des Bias $E\left(X_{(1)}\right) - \tau$ auf die Schätzungen der restlichen Verteilungsparameter zu untersuchen, wurde eine Monte-Carlo-Simulation durchgeführt. Es wurden je 100 Stichproben vom Umfang 50 aus lognormalverteilten Populationen mit $\mu = 1$ generiert. Dabei wurde dieselbe Parameterkonstellation mit $VC \in (0, 1.5]$ wie oben gewählt und aus den resultierenden Populationen Zufallsstichproben gezogen.

In Abbildung 2.25 sind die Ergebnisse der Monte-Carlo Studien zusammengefaßt, die wie folgt abliefen:

1. Für festes $\mu = 1$ wurden Stichproben aus lognormalverteilten Grundgesamtheiten mit den übrigen Verteilungsparametern $\tau = 0$ und $\sigma = 0.1, 0.2, ..., 1.5$ gezogen. Anmerkung: $\tau$ wurde mit 0 festgelegt, da der Wert von $\tau$ einerseits keinen Einfluß auf die Güte des 'Minimumschätzers' und andererseits auf die Güte der dreiparametrischen ML-Schätzer hat.

2. Es wurden also je 100 Stichproben aus 15 lognormalverteilten Populationen generiert. Aus den jeweils 100 resultierenden Schätzungen wurde das empirische Mittel gebildet und in den Graphiken eingetragen. Um die Konformität mit obigen Betrachtungen (Erwartungswert des Stichprobenminimums) zu gewährleisten, wurde als Skalierung der $x$-Achse wiederum $VC^2$ gewählt.

3. Die Daten wurden um das Minimum ($-\epsilon$) verschoben und die zweiparametrische ML-Methode zur Bestimmung der Schätzungen verwendet, die als strichlierte Linie gekennzeichnet sind.

4. Um die Schätzungen durch die 'Minimummethode' quantifizieren zu können, sind in beiden Abbildungen noch die Schätzungen aus der dreiparametrischen ML-Methode (punktierte Linie), sowie die exakten Populationswerte (durchgezogene Linie) eingetragen.

Das Ergebnis von zuvor wird bestätigt: Die Schätzungen werden mit wachsendem Variationskoeffizienten der zugrundeliegenden Normalverteilung zwar besser, jedoch die Abweichungen von den exakten Populationswerten bzw. den dreiparametrischen ML-Schätzungen bleiben erheblich. Berücksichtigt man weiters, daß geringe Abweichungen in der Schätzung von $\sigma$ große Abweichungen in den Schätzungen der Varianz nach sich ziehen, so ist von der Verwendung dieser 'Vereinfachung' abzuraten.

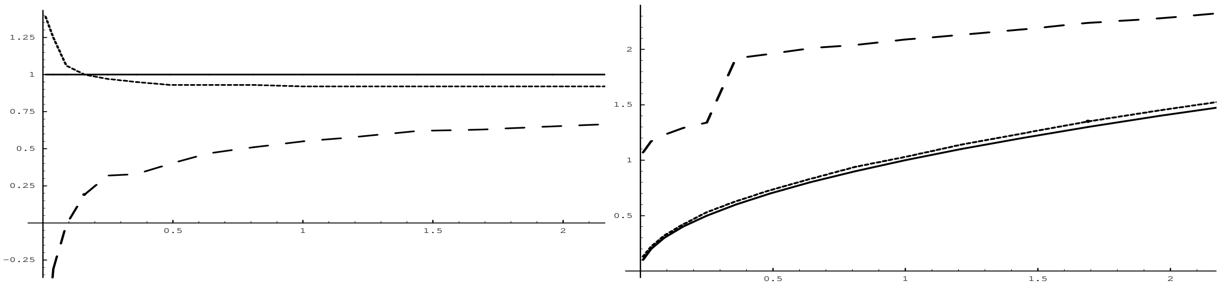



Abbildung 2.25: Auswirkungen des Bias der Schätzung von $\tau$ durch das Stichprobenminimum; Populationswerte: durchgezogene Linien; ML-Schätzungen nach Minimumtransformation: strichlierte Linien; dreiparametr. ML-Schätzungen: punktierte Linien; links: Schätzungen für $\mu$, rechts: Schätzungen für $\sigma$

### 2.4.2 Vorschlag zur empirischen Datenmodellierung

Wie bereits erwähnt, bereitet der threshold-Parameter $\tau$ bei weiteren statistischen Analysen (Bestimmung eines Konfidenzintervalls für einen Schätzer, Hypothesentests, Funktionen lognormalverteilter Zufallsvariabler) Probleme.

Daher folgender Vorschlag zur empirischen Datenmodellierung: Zeigt sich durch die explorative Datenanalyse, daß die Daten durch ein Lognormalverteilungsmodell beschrieben werden können, so soll zunächst der threshold-Parameter $\tau$ durch die vorgestellten Methoden (je nach 'Datenbeschaffenheit': ME, MME, MLE, MMLE) geschätzt werden. Um die weitere Analyse möglichst einfach zu halten, wird dann eine Beschreibung der um $\hat{\tau}$ geshifteten Daten mit dem zweiparametrischen Modell empfohlen, d.h. $\hat{\tau}$ wird in der weiteren Analyse nicht mehr als Verteilungsparameter sondern als Konstante betrachtet. Diese Vorgehensweise kann folgenderweise gerechtfertigt werden:

1. Bei den Simulationsstudien zeigte sich, daß der Schätzer des Shift-Parameters $\tau$ zuverlässig ist, wenn man die Schätzmethode in Abhängigkeit von der empirischen Schiefe und vom Stichprobenumfang wählt; d.h. auch wenn kein Shift 'nötig' ist, wird das durch den Schätzer von $\tau$ erkannt.

2. Bei der weiteren Analyse für die transformierten Werte kann ein zweiparametrisches Modell angesetzt werden, da in der Praxis bei gegebenen Datenmaterial weniger von dem Standpunkt 'Die Daten stammen aus dem speziellen Verteilungstyp', sondern vielmehr von dem Standpunkt 'Die Daten können durch ein Modell gut beschrieben werden' ausgegangen wird.

# Kapitel 3

# Hypothesentests und Konfidenzintervalle

## 3.1 Exakte Methode

Bei der Lognormalverteilung lassen sich alle 'interessanten' Größen durch exponentielle Funktionen von $\mu$ und $\sigma$ darstellen. In diesem Abschnitt werden Hypothesentests und Konfidenzintervalle für Funktionen $\theta(\mu, \sigma)$ der Form

1. $\theta(\mu, \sigma) = \mu$
2. $\theta(\mu, \sigma) = \sigma$
3. $\theta(\mu, \sigma) = \mu + k\sigma$
4. $\theta(\mu, \sigma) = \mu + k\sigma^2$

angeführt. LAND (1971) beschäftigte sich intensiv mit der Bestimmung von Konfidenzintervallen beziehungsweise kritischen Bereichen bei Hypothesentests von linearer Funktionen in den Normalverteilungsparametern $\mu$ und $\sigma$ (bzw. $\mu$ und $\sigma^2$).

Bei der Theorie von Tests auf lineare Funktionen von Verteilungsparametern $\mu$ und $\sigma$ einer Lognormalverteilung geht man von einer Verallgemeinerung der einfachen Zufallsstichprobe aus: $(Z, S)$ sei eine vollständig suffiziente Statistik für $(\mu, \sigma)$, wobei für $Z \sim N(\mu, \sigma^2/\gamma^2)$, $\gamma$ bekannt, gilt und $\nu S^2/\sigma^2 \sim \chi^2_\nu$ verteilt ist. Es wird vorausgesetzt, daß $Z$ und $S$ unabhängig sind.

In weiterer Folge werden die Ergebnisse für die vier zu Beginn angeführten linearen Funktionen in $\mu$ und $\sigma$ bzw. $\sigma^2$ angeführt. Die Bestimmung der kritischen Bereiche und der Konfidenzintervalle können bei LAND (1971) bzw. CROW and SHIMIZU (1988) nachgeschlagen werden.

Für jeden Punkt $\theta_0$ von $\theta$ sei

1. $H_1(\theta_0) := \{(\mu, \sigma) : \theta \leq \theta_0\}$ und $K_1(\theta_0) := \{(\mu, \sigma) : \theta > \theta_0\}$
2. $H_2(\theta_0) := \{(\mu, \sigma) : \theta \geq \theta_0\}$ und $K_2(\theta_0) := \{(\mu, \sigma) : \theta < \theta_0\}$
3. $H_3(\theta_0) := \{(\mu, \sigma) : \theta = \theta_0\}$ und $K_3(\theta_0) := \{(\mu, \sigma) : \theta \neq \theta_0\}$

### 3.1.1 Funktionen der Form $\theta(\mu, \sigma) = \mu$

**$\sigma$ bekannt:**

Für die einseitigen Hypothesentests (Test der Nullhypothese $H_1(\theta_0)$ gegen die Alternativhypothese $K_1(\theta_0)$ beziehungsweise der Hypothese $H_2(\theta_0)$ gegen $K_2(\theta_0)$) existieren UMP-kritische Bereiche (Uniformly Most Powerfull) zum Niveau $1 - \alpha$ (Definition siehe LEHMANN (1959)) der Form

$$\begin{aligned} C_1(\theta_0) &= \left\{ z : z > \frac{\sigma z_{1-\alpha}}{\gamma} + \theta_0 \right\} \\ C_2(\theta_0) &= \left\{ z : z < \frac{\sigma z_{\alpha}}{\gamma} + \theta_0 \right\}, \end{aligned}$$

wobei $z_\alpha$ das $\alpha$-Quantil der Standardnormalverteilung ist. Für den zweiseitigen Hypothesentest $H_3(\theta_0)$ gegen $K_3(\theta_0)$ existiert zwar kein UMP-kritischer Bereich, es kann aber gezeigt werden, daß ein UMP-unbiased-kritischer Bereich der Form

$$C_3(\theta_0) = \left\{ z : \frac{\gamma}{\sigma}(z - \theta_0) \notin \left[ z_{\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}} \right] \right\}$$

existiert. Für die zu $C_i(\theta_0)$, $i = 1, \ldots, 3$, korrespondierenden Konfidenzintervalle $U_i(z)$ folgt:

$$\begin{aligned} U_1(z) &= \left[ z - \frac{\sigma}{\gamma} z_{1-\alpha}, \infty \right) \\ U_2(z) &= \left( -\infty, z - \frac{\sigma}{\gamma} z_{\alpha} \right] \\ U_3(z) &= \left[ z - \frac{\sigma}{\gamma} z_{1-\frac{\alpha}{2}}, z - \frac{\sigma}{\gamma} z_{\frac{\alpha}{2}} \right] \end{aligned}$$

Für die Konfidenzintervalle $U_i$ gelten die zu $C_i$ korrespondierenden Eigenschaften.

**$\sigma$ unbekannt:**

Die optimalen kritischen Bereiche und Konfidenzintervallprozeduren werden anolog zu dem vorangegangenen Fall mit $\sigma$ bekannt bestimmt, wobei der exakte Wert $\sigma$ durch die Schätzung $s$ und, dadurch bedingt, die Quantile $z_\alpha$ der Standardnormalverteilung durch die Quantile $t_{\nu;\alpha}$ der Student-t-Verteilung mit $\nu$ Freiheitsgraden ersetzt werden:

$$\begin{aligned} C_1(\theta_0) &= \left\{ z : z > \frac{s t_{\nu;1-\alpha}}{\gamma} + \theta_0 \right\} \\ C_2(\theta_0) &= \left\{ z : z < \frac{s t_{\nu;\alpha}}{\gamma} + \theta_0 \right\} \\ C_3(\theta_0) &= \left\{ z : \frac{\gamma(z - \theta_0)}{s} \notin \left[ t_{\nu;\frac{\alpha}{2}}, t_{\nu;1-\frac{\alpha}{2}} \right] \right\} \end{aligned}$$

Die kritischen Bereiche $C_i(\theta_0)$ sind UMP-unbiased zum Niveau $\alpha$ für den Hypothesentest $H_i(\theta_0)$ gegen $K_i(\theta_0)$. Es existieren jedoch keine UMP-kritischen Bereiche zum Niveau

$\alpha$, da durch die Methode mit bekanntem $\sigma$ die 'besten' kritischen Bereiche bestimmt werden. Weiters gilt: $C_1(\theta_0)$ und $C_2(\theta_0)$ sind invariant unter $G_1$, wobei $G_1$ die Menge der Funktionen der Form $g(z,s) = (az, as), a > 0$, bezeichnet. $C_3(\theta_0)$ ist invariant unter $G_2 = \{g(z,s) : g(z,s) = (bz, \mid b \mid s), b \neq 0\}$.

Die zu $C_i$ korrespondierenden Konfidenzintervallprozeduren zum Niveau $1-\alpha$ mit der korrespondierenden Eigenschaft UMA-unbiased (bzw. UMA-invariant) - Uniformly Most Accurate (Definition siehe LEHMANN (1959))- sind definiert durch:

$$\begin{aligned}
U_1(z,s) &= \left[z - \frac{s}{\gamma} t_{\nu;1-\alpha}, \infty\right) \\
U_2(z,s) &= \left(-\infty, z - \frac{s}{\gamma} t_{\nu;\alpha}\right] \\
U_3(z,s) &= \left[z - \frac{s}{\gamma} t_{\nu;1-\frac{\alpha}{2}}, z - \frac{s}{\gamma} t_{\nu;\frac{\alpha}{2}}\right]
\end{aligned}$$

**Anwendung auf die zweiparametrische Lognormalverteilung**

Hypothesen über den Median von $\Lambda(\mu, \sigma^2)$, $e^\mu$, können in Termen des Logarithmus des Medians, $\mu$, spezifiziert werden. So kann die Nullhypothese $H_1(\theta_0) : e^\mu \leq \theta_0$ als $H_1(\theta_0) : \mu \leq \ln(\theta_0)$ geschrieben werden. Exponentiert man die Endpunkte des optimalen Konfidenzintervalls für $\mu$, so ist dadurch das optimale Konfidenzintervall für den Median $e^\mu$ definiert.

### 3.1.2 Funktionen der Form $\theta(\mu, \sigma) = \sigma$

**$\mu$ bekannt:**

Die Teststatistik

$$V = \sqrt{\nu S^2 + \gamma^2 (z-\mu)^2}$$

ist suffizient für $\sigma$. $V^2/\sigma^2$ ist $\chi^2_{\nu+1}$-verteilt. Für die einseitigen Tests existieren, analog zu obigem Abschnitt, UMP-kritische Bereiche:

$$\begin{aligned}
C_1(\theta_0) &= \left\{(z,s) : \frac{v^2}{\theta_0^2} > \chi^2_{\nu+1;1-\alpha}\right\} \\
C_2(\theta_0) &= \left\{(z,s) : \frac{v^2}{\theta_0^2} < \chi^2_{\nu+1;\alpha}\right\}
\end{aligned}$$

Für den zweiseitigen Test existiert kein UMP-kritischer, jedoch ein UMP-unbiased kritischer Bereich:

$$C_3(\theta_0) = \left\{(z,s) : \frac{v^2}{\theta_0^2} \notin \left[\chi^2_{\nu+1;\alpha_1}, \chi^2_{\nu+1;1-\alpha_2}\right]\right\}$$

Da die Referenzverteilung nun nicht mehr symmetrisch ist, müßten die Wahrscheinlichkeiten $\alpha_1$ und $\alpha_2$ aus Integralgleichungen über der Dichte der Chiquadrat-Verteilung bestimmt werden, ist $\nu$ jedoch nicht zu klein, so ist $\alpha/2$ eine gute Approximation für $\alpha_1$ und $\alpha_2$.

Für die Konfidenzintervallprozeduren ergibt sich:

$$
\begin{aligned}
U_1(z,s) &= \left[\frac{v}{\chi_{\nu+1;1-\alpha}}, \infty\right) \\
U_2(z,s) &= \left(0, \frac{v}{\chi_{\nu+1;\alpha}}\right] \\
U_3(z,s) &= \left[\frac{v}{\chi_{\nu+1;1-\alpha_2}}, \frac{v}{\chi_{\nu+1;\alpha_1}}\right]
\end{aligned}
$$

**$\mu$ unbekannt:**

Die Teststatistik $\frac{\nu S^2}{\sigma_0^2}$ ist $\chi^2_\nu$. Der Unterschied zu dem Fall mit bekanntem $\mu$ besteht darin, daß die Information des bekannten $\mu$ nicht mehr ausgenutzt werden kann (der Ausdruck $\gamma^2(z-\mu)^2$). Die Verteilung der Teststatistiken $\frac{\nu S^2}{\sigma^2}$ und $V^2/\sigma^2$ unterscheiden sich jedoch nur in einem Freiheitsgrad.

Für die Hypothesentests existieren UMP-unbiased kritische Bereiche:

$$
\begin{aligned}
C_1(\theta_0) &= \left\{(z,s) : \frac{\nu s^2}{\theta_0^2} > \chi^2_{\nu;1-\alpha}\right\} \\
C_2(\theta_0) &= \left\{(z,s) : \frac{\nu s^2}{\theta_0^2} < \chi^2_{\nu;\alpha}\right\} \\
C_3(\theta_0) &= \left\{(z,s) : \frac{\nu s^2}{\theta_0^2} \notin \left[\chi^2_{\nu;\alpha_1}, \chi^2_{\nu;1-\alpha_2}\right]\right\}
\end{aligned}
$$

Für die Wahrscheinlichkeiten $\alpha_1$ und $\alpha_2$ gilt selbes wie im vorherigen Abschnitt. Für die Konfidenzintervallprozeduren ergibt sich:

$$
\begin{aligned}
U_1(z,s) &= \left[\frac{\sqrt{\nu}s}{\chi_{\nu;1-\alpha}}, \infty\right) \\
U_2(z,s) &= \left(0, \frac{\sqrt{\nu}s}{\chi_{\nu;\alpha}}\right] \\
U_3(z,s) &= \left[\frac{\sqrt{\nu}s}{\chi_{\nu;1-\alpha_2}}, \frac{\sqrt{\nu}s}{\chi_{\nu;\alpha_1}}\right]
\end{aligned}
$$

Die kritischen Bereiche und Konfidenzintervallprozeduren sind invariant unter der Gruppe der Transformationen $g(z,s) = (z+a,s), -\infty < a < \infty$.

### Anwendung auf die zweiparametrische Lognormalverteilung

Anwendungen ergeben sich zum Beispiel bei der Untersuchung des Variationskoeffizienten, der Schiefe und Kurtosis. Die Methodik ist dieselbe wie bei Funktionen der Form $\theta(\mu,\sigma) = \mu$.

### 3.1.3 Funktionen der Form $\theta(\mu, \sigma) = \mu + k\sigma$

Die Hypothesen $H_1(\theta_0) : \mu + k\sigma \leq \theta_0$, $K_1(\theta_0) : \mu + k\sigma > \theta_0$, ect., sind invariant unter $G = \{g(z, s) : g(z - \theta_0, s) = g(a(z - \theta_0), as), a > 0\}$. Für die Teststatistik

$$T = \frac{\gamma(Z - \theta_0)}{S}$$

gilt:

1. $T$ ist maximal invariant unter G.
2. $T$ ist nicht-zentral t-verteilt mit $\nu$ Freiheitsgraden und dem Parameter $\delta = \gamma(\mu - \theta_0)/\sigma$.

Quantile der t-Verteilung für verschiedene Parameterkonstellationen können bei PEARSON and HARTLEY (1972) gefunden werden. Die Tests $H_i(\theta_0)$ gegen $K_i(\theta_0)$ können daher in Termen des Parameters $\delta$ angeschrieben werden: $H_1(\theta_0) : \delta \leq \eta, K_1(\theta_0) : \delta > \eta$ mit $\eta = -\gamma k$. Für die einseitigen Tests auf $\mu + k\sigma, k \neq 0$ ergeben sich somit folgende UMP-invarianten kritischen Bereiche:

$$\begin{aligned} C_1(\theta_0) &= \left\{(z, s) : \frac{\gamma(z - \theta_0)}{s} > t_{\nu,\eta;1-\alpha}\right\} \\ C_2(\theta_0) &= \left\{(z, s) : \frac{\gamma(z - \theta_0)}{s} < t_{\nu,\eta;\alpha}\right\} \end{aligned}$$

Für den zweiseitigen Test erhält man einen UMP-unbiased-invarianten kritischen Bereich

$$C_3(\theta_0) = \left\{(z, s) : \frac{\gamma(z - \theta_0)}{s} \notin [t_{\nu,\eta;\alpha_1}, t_{\nu,\eta;1-\alpha_2}]\right\}$$

Analog zum $\chi^2$-Test müßten die Wahrscheinlichkeiten $\alpha_1$ und $\alpha_2$ über Integralgleichungen über der Dichte ermittelt werden, in der Praxis wird jedoch meistens die Approximation $\alpha/2$ verwendet. Für die Konfidenzintervallprozeduren ergibt sich:

$$\begin{aligned} U_1(z, s) &= \left[z - \frac{s}{\gamma} t_{\nu,\eta;1-\alpha}, \infty\right) \\ U_2(z, s) &= \left(-\infty, z - \frac{s}{\gamma} t_{\nu,\eta;\alpha}\right] \\ U_3(z, s) &= \left[z - \frac{s}{\gamma} t_{\nu,\eta;1-\alpha_2}, z - \frac{s}{\gamma} t_{\nu,\eta;\alpha_1}\right] \end{aligned}$$

#### Anwendung auf die zweiparametrische Lognormalverteilung

Bei der Untersuchung von Quantilen der Lognormalverteilung wird diese Theorie verwendet, wobei $k$ durch das entsprechende Quantil der Standardnormalverteilung ersetzt wird. Wie auch schon in den Fällen zuvor, werden die Ausgangsdaten logarithmiert und für die transformierten Werte werden kritische Bereiche und optimale Konfidenzintervallprozeduren bestimmt. Führt man eine Rücktransformation der Konfidenzintervallgrenzen durch, so erhält man optimale Konfidenzbereiche für die Quantile der Lognormalverteilung.

### 3.1.4 Funktionen der Form $\theta(\mu, \sigma) = \mu + \lambda\sigma^2$

Dieser Fall ist für die Untersuchung lognormalverteilter Zufallsvariabler interessant, da alle Momente um Null in diese Kategorie passen. Dieser Fall ist aber auch der theoretisch schwierigste. Im folgenden werden die Ergebnisse beschrieben, genauere Ausführungen findet man bei LAND (1971) und CROW and SHIMIZU (1988).

Die UMP-unbiased kritischen Bereiche für den Test (ein- und zweiseitig) auf $\mu + \lambda\sigma^2, \lambda \neq 0$, basieren auf der bedingten Verteilung von $T$ unter $V$

$$\begin{aligned} T &= \frac{\gamma(Z - \theta_0)}{S} \\ V^2 &= \nu S^2 + \gamma^2(Z - \theta_0), \end{aligned}$$

beides Teststatistiken, die schon in früheren Abschnitten verwendet wurden. Diese bedingte Verteilung ist in geschlossener Form darstellbar, ihre Quantile werden mit $t_{\nu,q|V;\alpha}$ für einseitige und mit $t^1_{\nu,q|V;\alpha}$ bzw. $t^2_{\nu,q|V;\alpha}$ für zweiseitige Tests bezeichnet, wobei $q = \frac{-\gamma v\lambda}{\nu+1}$. Es gelten folgende Symmetrieeigenschaften:

1. $t_{\nu,q|V;\alpha} = -t_{\nu,-q|V;1-\alpha}$
2. $t^1_{\nu,q|V;\alpha} = -t^2_{\nu,-q|V;\alpha}$

Es ergeben sich folgende UMP-unbiased kritischen Bereiche:

$$\begin{aligned} C_1(\theta_0) &= \left\{(z, s) : \frac{\gamma(z - \theta_0)}{s} > t_{\nu,q|V;1-\alpha}\right\} \\ C_2(\theta_0) &= \left\{(z, s) : \frac{\gamma(z - \theta_0)}{s} < t_{\nu,q|V;\alpha}\right\} \\ C_3(\theta_0) &= \left\{(z, s) : \frac{\gamma(z - \theta_0)}{s} \notin \left[t^1_{\nu,q|V;\alpha}, t^2_{\nu,q|V;\alpha}\right]\right\} \end{aligned}$$

Die Bestimmung der optimalen Konfidenzbereiche ist nicht mehr so einfach wie in den Fällen zuvor, denn $\theta_0$ beeinflußt neben der Teststatistik auch die Bestimmung des kritischen Wertes, da $q$ von $\theta_0$ abhängig ist.

### 3.1.5 Nichtlineare Funktionen in $\mu$ und $\sigma$

Einige Charakteristiken der Lognormalverteilung lassen sich nicht durch die vorgestellten Methoden beschreiben, wie zum Beispiel die zentralen Momente. Eine Methode, die ebenfalls von Land entwickelt wurde, ermöglicht es, die Konfidenzintervalle von Ausdrücken der Form $E(f(Z))$ zumindest approximativ zu bestimmen, wobei $Z \sim N(\mu, \sigma^2)$ und $f$ eine nichtlineare, zweimal differenzierbare reelle Funktion (siehe CROW nad SHIMIZU (1988)). Land schlug vor, die parametrische Funktion $\theta(\mu, \sigma) = E(f(Z))$ durch eine lineare Funktion $\mu + \lambda\sigma^2$ zu approximieren, wobei $\lambda$ das Verhältnis der partiellen Ableitungen von $\theta(\mu, \sigma)$ nach $\mu$ und $\sigma^2$, also $\lambda = \frac{\theta_{\sigma^2}}{\theta_\mu}$, an der Stelle $(\hat{\mu}, \hat{\sigma}^2) = (z, s^2)$ ist. Erhält man $u(z, s)$ als Konfidenzgrenze für die lineare Funktion $\mu + \lambda\sigma^2$, so ist die korrespondierende Intervallgrenze für $\theta(\mu, \sigma)$ gleich $\theta(\tilde{\mu}, \tilde{\sigma})$, wobei $(\tilde{\mu}, \tilde{\sigma}^2)$ der Punkt auf der Geraden $\mu + \lambda\sigma^2 = u(z, s)$ ist, für den die Likelihoodfunktion maximal wird.

## 3.2 Approximative Verfahren

Neben der exakten Methode existieren noch approximative Methoden zur Bestimmung von Konfidenzintervallen für Kenngrößen einer lognormalverteilten Zufallsvariablen. Der Grund für die Anwendung von Approximationsmethoden liegt vor allem in der Komplexität der exakten Methode.

In diesem Abschnitt werden drei Approximationsmethoden zur Bestimmung eines (approximativen) Konfidenzintervalls für den Erwartungswert einer lognormalverteilten Zufallsvariablen vorgestellt:

1. Methode von Cox (LAND (1972))

2. Konservative Methode von Angus (ANGUS (1988))

3. Parametrische Bootstrapmethode von Angus (ANGUS (1994))

Sei $X$ eine lognormalverteilte Zufallsvariable mit

$Z = \ln(X) \sim N(\mu, \sigma^2)$,

$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ und

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2$ (womit gilt: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$).

### 3.2.1 Methode von Cox

Da die Statistik $\left(\overline{Z}, S^2\right)$ vollständig suffizient für $(\mu, \sigma^2)$ ist (siehe auch exakte Methode), kann man Schätzungen von $\theta = e^{\mu+\frac{\sigma^2}{2}}$ bzw. $\ln\theta = \mu + \frac{\sigma^2}{2}$ über $\left(\overline{Z}, S^2\right)$ treffen. Da $\overline{Z} + \frac{S^2}{2}$ der UMVU-Schätzer für $\ln\theta$ ist und seine Varianz die Form $\frac{S^2}{n} + \frac{S^4}{2(n-1)}$ hat (Herleitung über die Verteilungen von $\overline{Z}$ und $S^2$), ist es möglich, ein approximatives Konfidenzintervall für $\ln\theta$ wie folgt zu konstruieren:

$$
\begin{aligned}
L_{1-\alpha}(\overline{Z}, S^2) &= \overline{Z} + \frac{S^2}{2} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{S^2}{n} + \frac{S^4}{2(n-1)}} \\
U_{1-\alpha}(\overline{Z}, S^2) &= \overline{Z} + \frac{S^2}{2} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{S^2}{n} + \frac{S^4}{2(n-1)}}
\end{aligned} \tag{3.1}
$$

### 3.2.2 Konservative Methode von Angus

Angus geht bei seinem Konstruktionsvorschlag von folgender (approximativen) Statistik für $\ln\theta$ aus:

$$V(\theta) = \frac{\left(\overline{Z} + \frac{S^2}{2} - \ln\theta\right)}{\sqrt{\frac{S^2}{n} + \frac{S^4}{2n}}} \tag{3.2}$$

Da $\overline{Z} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ und $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ gilt, hat obige Teststatistik die selbe Verteilung wie

$$T(\sigma) = \frac{U + \frac{\sigma\sqrt{n}}{2}\left(\frac{\chi^2_{n-1}}{n-1} - 1\right)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}\left(1 + \frac{\sigma^2}{2}\frac{\chi^2_{n-1}}{n-1}\right)}}, \tag{3.3}$$

wobei $U$ die Standardnormalverteilung und $\chi^2_{n-1}$ die $\chi^2$-Verteilung mit $n-1$ Freiheitsgraden ist. $U$ und $\chi^2_{n-1}$ sind unabhängig, da $\overline{Z}$ und $S^2$ unabhängig sind. Bezeichnet man mit $F(x;\sigma) = P(T(\sigma) \leq x)$ die Verteilungsfunktion von $T(\sigma)$. Dann gilt: Für ein fixiertes $x$ und $n > 2$ ist $F(x;\sigma)$ eine monoton steigende Funktion in $\sigma$.

Für $\sigma^2 \to 0$ gilt:

$$T(\sigma) \overset{\sigma^2 \to 0}{\to} T_0 = \frac{U}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

$T_o$ ist t-verteilt mit $n-1$ Freiheitsgraden.

Für $\sigma^2 \to \infty$ gilt

$$T(\sigma) \overset{\sigma^2 \to \infty}{\to} T_\infty = \sqrt{\frac{n}{2}}\left(1 - \frac{n-1}{\chi^2_{n-1}}\right)$$

Bezeichnet man mit $t_{1-\alpha;n-1}$ das $(1-\alpha)$-Quantil einer t-Verteilung mit $n-1$ Freiheitsgraden und

$$q_{\alpha;n-1} = \sqrt{\frac{n}{2}}\left(\frac{n-1}{\chi^2_{\alpha;n-1}} - 1\right)$$

wobei $\chi^2_{\alpha;n-1}$ das $\alpha$-Quantil einer $\chi^2$- Verteilung mit $n-1$ Freiheitsgraden ist, so lassen sich die obere und untere Schranke des Konfidenzintervalls folgend angeben:

$$\begin{aligned} L_{1-\alpha}(\overline{Z}, S^2) &= \overline{Z} + \frac{S^2}{2} - \frac{t_{1-\frac{\alpha}{2};(n-1)}}{\sqrt{n}}\sqrt{S^2\left(1 + \frac{S^2}{2}\right)} \\ U_{1-\alpha}(\overline{Z}, S^2) &= \overline{Z} + \frac{S^2}{2} + \frac{q_{\frac{\alpha}{2};(n-1)}}{\sqrt{n}}\sqrt{S^2\left(1 + \frac{S^2}{2}\right)} \end{aligned} \tag{3.4}$$

Das Verfahren von Angus wird konservativ genannt, da die Wahrscheinlichkeit, daß $\ln\theta$ in das approximative Konfidenzintervall fällt, $\geq 1-\alpha$ ist.

### 3.2.3 Parametrische Bootstrap-Methode von Angus

Angus wandte auf die approximative Statistik $V(\theta)$ in (3.2.2) die parametrische $t$ - Perzentil-Bootstrapmethode an, um ein Bootstrap-Konfidenzintervall für den Erwartungswert einer lognormalverteilten Zufallsvariablen zu erhalten. Seien $t_0$ und $t_1$ die $\alpha/2$- bzw.

$1-\alpha/2$-Perzentile von $V(\theta)$. Ein theoretisches Konfidenzintervall für $\ln\theta$ zum Niveau $1-\alpha$ ist daher

$$I = \left[\overline{Z} + \frac{S^2}{2} - \frac{t_1}{\sqrt{n}}\sqrt{S^2\left(1+\frac{S^2}{2}\right)}, \overline{Z} + \frac{S^2}{2} - \frac{t_0}{\sqrt{n}}\sqrt{S^2\left(1+\frac{S^2}{2}\right)}\right]$$

Die unbekannten Quantile $t_0$ und $t_1$ können über eine parametrische Bootstrap-Stichprobe geschätzt werden. Da die Statistik $V(\theta)$ die selbe Verteilung wie $T(\sigma)$ in (3.2.2) hat, können $t_0$ und $t_1$ über folgenden Algorithmus geschätzt werden:

1. Man generiere $2*B$ Stichproben ($B$...Anzahl der Replikationen) derart, daß $N_i^* \sim N(0,1)$ und $\chi_i^{2*} \sim \chi_{n-1}^2$, $i = 1,\ldots,B$.

2. Man ersetze in der Formel (3.2.2) für $T(\sigma)$ die Größen $\sigma, N, \chi_{n-1}^2$ durch $S, N_i^*, \chi_i^{2*}$ und erhält somit $T_i^*$.

3. Sortiert man $T_i^*$ , $i = 1,\ldots,B$, derart, daß $T_{(1)}^* < T_{(2)}^* < \ldots < T_{(B)}^*$, so wird $t_1$ durch $t_1^* = T_{[(1-\alpha/2)B]}^*$ und $t_0$ durch $t_0^* = T_{[\alpha/2*B]}^*$ geschätzt, wobei $[v]$ die größte ganze Zahl kleiner oder gleich $v$.

Man erhält damit folgende Intervallgrenzen:

$$\begin{aligned} L_{1-\alpha}^*(\overline{Z}, S^2) &= \overline{Z} + \frac{S^2}{2} - \frac{t_1^*}{\sqrt{n}}\sqrt{S^2\left(1+\frac{S^2}{2}\right)} \\ U_{1-\alpha}^*(\overline{Z}, S^2) &= \overline{Z} + \frac{S^2}{2} - \frac{t_0^*}{\sqrt{n}}\sqrt{S^2\left(1+\frac{S^2}{2}\right)} \end{aligned} \tag{3.5}$$

### 3.2.4 Vergleich der Approximationsmethoden

ZHOU und GAO (1997) führten eine Simulationsstudie zum Vergleich der drei vorgestellten Methoden durch. Sie trafen folgende Annahmen: Es werden zweiseitige Konfidenzintervalle mit gleichen Restwahrscheinlichkeit ($\alpha_1 = \alpha_2 = \alpha/2$) berechnet. Es wurden Stichproben vom Umfang $n = 11$, $n = 101$ und $n = 400$ untersucht; die Varianz wurde aus dem Intervall $[0.1, 20.0]$ genommen. Da bei der Berechnung der Konfidenzintervalle eine Änderung des Parameters von $\mu$ auf $\mu_1$ bei gleichem Stichprobenumfang $n$ und festem $S^2$ nur eine Verschiebung des Intervalls um $\mu_1 - \mu$ bewirkt, wurde, ohne Beschränkung der Allgemeinheit, $\mu = -\frac{\sigma^2}{2}$ gesetzt, womit $\ln\theta = 0$ folgt. Für jede Parameterkonstellation wurden 5000 Stichproben aus der jeweiligen Lognormalverteilung generiert. Die Anzahl der Replikationen $B$ für das parametrische Bootstrapverfahren wurde mit 2000 festgelegt.

Die Vergleichskriterien waren:

1. Der Überdeckungsfehler, definiert als die absolute Differenz zwischen der Überdeckungswahrscheinlichkeit und dem vorgegebenen Niveau.

2. Die durchschnittliche Intervalllänge.

3. Die Art des Fehlers: Prozentsatz der KIV, die den Wert $\ln\theta = 0$ nicht überdecken. ($\%KIV < \ln\theta$ bzw. $\%KIV > \ln\theta$)

4. Der relative Bias, definiert als $\frac{|(\%KIV<\ln\theta)-(\%KIV>\ln\theta)|}{(\%KIV<\ln\theta)+(\%KIV>\ln\theta)}$.

### 3.2.5 Ergebnisse

Die konservative Methode von Angus liefert stets Konfidenzintervalle mit einer größeren Überdeckungswahrscheinlichkeit als die vorgegebene. Daher sind deren durchschnittliche Intervalllängen größer als jene der anderen Methoden. Die Methode von Cox ist für $n > 50$ zu empfehlen, die parametrische Bootstrapmethode für kleinere Stichprobenumfänge.

# Kapitel 4

# Funktionen von lognormalverteilten Zufallsvariablen

Funktionen von lognormalverteilten Zufallsvariablen spielen in der Praxis eine große Rolle. In diesem Abschnitt werden (i) das Produkt und der Quotient und (ii) die Summe von lognormalverteilten Zufallsvariablen besprochen.

## 4.1 Das Produkt und der Quotient

Seien $X_i \sim \Lambda(\mu_i, \sigma_i^2)$, und $X_i$ unabhängig, dann gilt für das Produkt

$$L = \prod_{i=1}^{N} X_i = \exp\left(\ln \prod_{i=1}^{N} X_i\right) = \exp\left(\sum_{i=1}^{N} \ln X_i\right) \sim \Lambda\left(\sum_{i=1}^{N} \mu_i, \sum_{i=1}^{N} \sigma_i^2\right),$$

wegen $\sum \ln X_i \overset{iid}{\sim} N\left(\sum_{i=1}^{N} \mu_i, \sum_{i=1}^{N} \sigma_i^2\right)$.

Diese Aussage läßt sich noch verallgemeinern zu

$$a \prod_{i=1}^{N} X_i^{b_i} \sim \Lambda\left(\ln a + \sum_{i=1}^{N} b_i \mu_i, \sum_{i=1}^{N} b_i^2 \sigma_i^2\right), \qquad \text{mit} \quad a > 0, b_i \in \mathbb{R} \tag{4.1}$$

Für die Verteilung des Quotienten von lognormalverteilten Zufallsvariablen ergibt sich mit Konstanten $c > 0$, $d_j \in \mathbb{R}$:

$$\frac{a}{c} \frac{\prod_{i=1}^{N} X_i^{b_i}}{\prod_{j=1}^{M} X_j^{d_j}} \sim \Lambda\left(\ln a - \ln c + \sum_{i=1}^{N} b_i \mu_i - \sum_{j=1}^{M} d_j \mu_j, \sum_{i=1}^{N} b_i^2 \sigma_i^2 + \sum_{j=1}^{M} d_j^2 \sigma_j^2\right) \tag{4.2}$$

## 4.2 Die Summe

Die Verteilung der Summe von lognormalverteilten Zufallsvariablen ist nicht mehr in geschlossener, mit vorigen Ausführungen vergleichbarer, Form darstellbar. Nach dem zentralen Grenzwertsatz gilt zwar, daß die Verteilung der standardisierten Summe von $X_i \sim \Lambda(\mu, \sigma^2)$ für $N \to \infty$ gegen die Normalverteilung strebt, es soll aber in diesem Abschnitt auch untersucht werden, welche Form die Summenverteilung für eine kleine Anzahl von Summanden hat.

### 4.2.1 Idee

Es wurde folgende Vorgehensweise gewählt:

1. Es wird versucht, die Summenverteilung durch drei Verteilungstypen zu approximieren:

   (a) Normalverteilung

   (b) Lognormalverteilung

   (c) Gammaverteilung

2. Bei der Approximation soll gelten: Die Summenverteilung stimmt mit den Verteilungen, durch die sie approximiert wird, im Erwartungswert und der Varianz (damit natürlich auch im zweiten Moment um Null) überein.

3. Als Kriterium für die Güte der Approximation dient die Abweichung der Schiefekoeffizienten (Schiefe der Summenverteilung – Schiefe der Approximationsverteilung).

4. Die Anzahl der Summanden $N$ wird im Bereich $[2, \ldots, 2000]$ betrachtet.

5. Folgende Werte von $\sigma$ wurden betrachtet: $0.1(0.1), \ldots, 2.0$.

6. Die Güte der Approximation wird in Abhängigkeit von der Anzahl der Summanden und dem Schiefekoeffizienten der Summanden untersucht ($\sigma$ ist der einzige Parameter, der auf die Schiefe lognormalverteilter Zufallsvariabler Einfluß hat).

### 4.2.2 Aufstellen der Gleichungssysteme

Die Momente der Summenverteilung lassen sich wie folgt angeben: Seien $X_i \overset{iid}{\sim} \Lambda\left(\mu, \sigma^2\right)$, $i = 1, \ldots, N$, $S = \sum_i^N X_i$:

$$\begin{aligned}
E(S) &= E\left(\sum X_i\right) = Ne^{\mu+\frac{\sigma^2}{2}} \qquad (4.3)\\
Var(S) &= Var\left(\sum X_i\right) = Ne^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)\\
E(S^3) &= E\left(\left(\sum X_i\right)^3\right)\\
&= Ne^{3\mu+\frac{9\sigma^2}{2}} + 3N(N-1)e^{3\mu+\frac{5\sigma^2}{2}} + N(N-1)(N-2)e^{3\mu+\frac{3\sigma}{2}}\\
\gamma_1(S) &= \frac{E\left((S-E(S))^3\right)}{(Var(S))^{3/2}} = \frac{e^{3\sigma^2} - 3e^{\sigma^2} + 2}{\sqrt{N\left(e^{\sigma^2}-1\right)^3}}
\end{aligned}$$

Anmerkung: Die Gleichungen für das 3. Moment um Null erhält man durch Anwendung der Aussagen über das Produkt von lognormalverteilten Zufallsvariablen.

Sei $Z \sim N\left(\mu_N, \sigma_N^2\right)$, $L \sim \Lambda\left(\mu_L, \sigma_L^2\right)$ und $G \sim \Gamma(a, \lambda)$. Es gilt:

1. Momente der Normalverteilung:

$$E(Z) = \mu_Z \qquad Var(Z) = \sigma_Z^2 \qquad \gamma_1(Z) = 0$$

2. Momente der Lognormalverteilung:

$$\begin{aligned} E(L) &= e^{\mu_L + \frac{\sigma_L^2}{2}} \qquad Var(L) = e^{2\mu_L + \sigma_L^2}\left(e^{\sigma_L^2} - 1\right) \\ \gamma_1(L) &= \left(e^{\sigma_L^2} - 1\right)^{1/2}\left(e^{\sigma_L^2} + 2\right) \end{aligned}$$

3. Momente der Gammaverteilung:

$$E(G) = \frac{a}{\lambda} \qquad Var(G) = \frac{a}{\lambda^2} \qquad \gamma_1(G) = \frac{2}{\sqrt{a}}$$

1. Normalverteilung: Mit
   (a) $\mu_Z = Ne^{\mu + \frac{\sigma^2}{2}}$ und
   (b) $\sigma_Z^2 = Ne^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$

   folgt für die Abweichung in der Schiefe $\gamma_1(S) - \gamma_1(Z)$

$$d(Z) = \gamma_1(S) \tag{4.4}$$

2. Lognormalverteilung: Mit
   (a) $e^{\mu_L + \frac{\sigma_L^2}{2}} = Ne^{\mu + \frac{\sigma^2}{2}}$ und
   (b) $e^{2\mu_L + \sigma_L^2}\left(\sigma_L^2 - 1\right) = Ne^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$

   folgt $\sigma_L^2 = \ln\left(e^{\sigma^2} + N - 1\right) - \ln(N)$.

$$d(L) = \gamma_1(S) - \gamma_1(L) \tag{4.5}$$

3. Gammaverteilung: Mit
   (a) $\frac{a}{\lambda} = Ne^{\mu + \frac{\sigma^2}{2}}$ und
   (b) $\frac{a}{\lambda^2} = Ne^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$

   folgt $a = \frac{N}{e^{\sigma^2} - 1}$

$$d(G) = \gamma_1(S) - \gamma_1(G) \tag{4.6}$$

### 4.2.3 Ergebnisse

Wie bereits erwähnt, ergab sich bei allen drei Approximationsversuchen, daß die Abweichnung der Schiefekoeffizienten unabhängig von der Größe $\mu$ der Summanden ist und als Funktion in $\sigma$ darstellbar ist. Die Ergebnisse sind einigermaßen überraschend: Selbst bei 2000 Summanden erreicht man durch die Lognormalverteilung die 'beste' Approximation, wenngleich die Unterschiede in den Abweichungen nur mehr gering sind. Die Abweichung der Gammaapproximation ist ebenfalls für alle untersuchten Variationen geringer als jene von der Normalapproximation. Dieses Ergebnis ist nicht überraschend, da die Gammaverteilung eine rechtsschiefe Verteilung ist.

In Abbildung 4.1 sind die Abweichungen der Schiefekoeffizienten für 2, 10, 50, 100, 500 und 2000 Summanden eingetragen (Schiefe der Summenverteilung – Schiefe der Approximationsverteilung).

Aus der Abbildung kann abgeleitet werden:

1. Mit wachsendem $N$ nähern sich die drei Approximationsverteilungen.

2. Für geringe Schiefekoeffizienten (bzw. geringem $\sigma$) der Summanden erreicht man mit allen drei Verteilungen gute Approximationen.

3. Der Punkt, ab dem die Abweichung groß wird, verschiebt sich mit wachsendem $N$ nach rechts.

4. Durch die Lognormalverteilung wird stets die 'beste' Approximation erreicht.

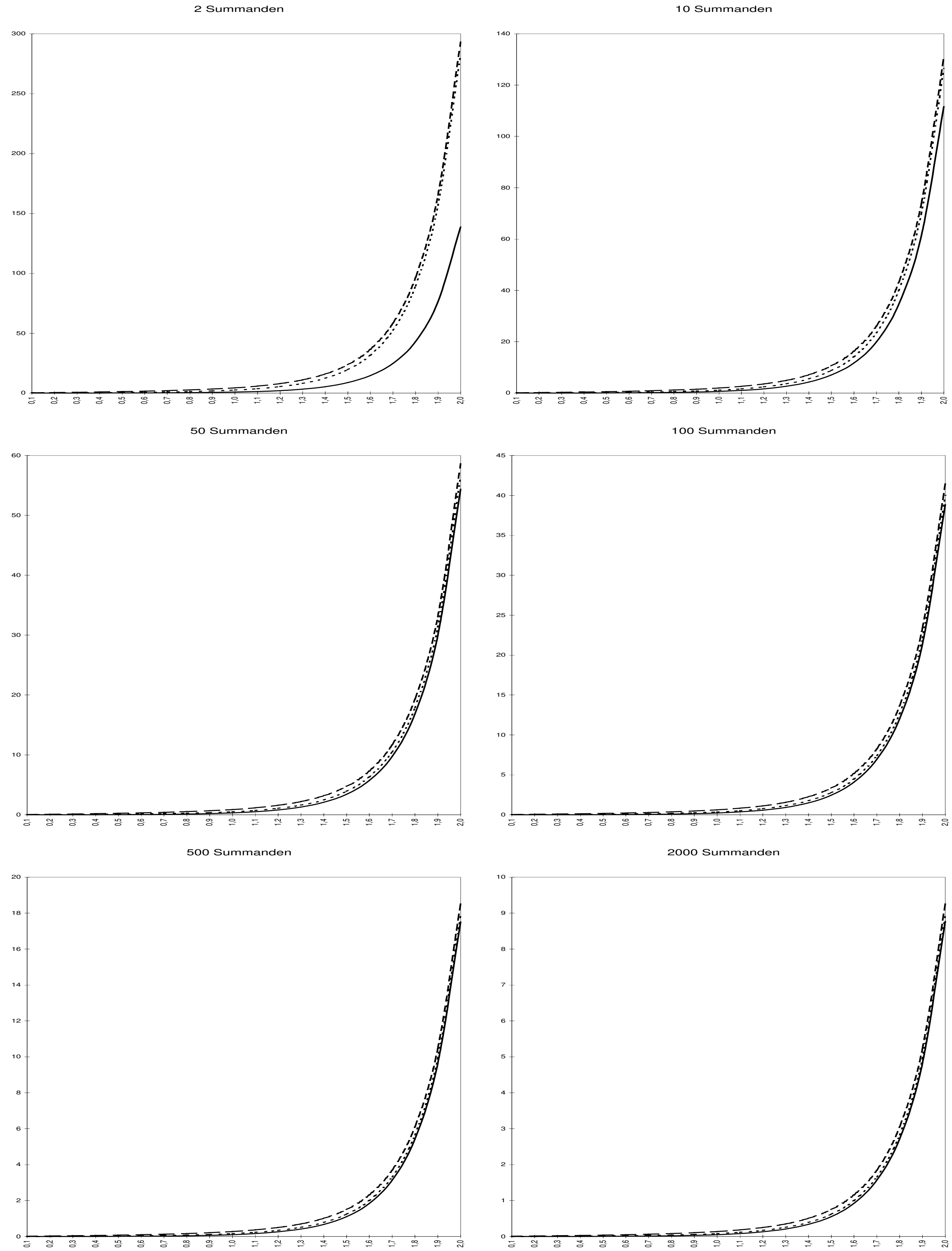



Abbildung 4.1: Vergleich der Abweichungen der Schiefekoeffizienten; Lognormalverteilung: durchgezogen; Normalverteilung: strichliert; Gammaverteilung: punktiert. Auf der $x$-Achse ist $\sigma$ aufgetragen, auf der $y$-Achse die Abweichung der Schiefekoeffizienten ( d aus (4.4)- (4.6))

# Kapitel 5

# Auswertung von Energieverbrauchs- und Emissionsdaten

## 5.1 Einleitung

Im Zuge der Erstellung eines Umwelt-Informations-Sytems für die Stadt Graz (kurz UIS) wurde das Institut für Statistik mit der Analyse von Energieverbrauchs- und Emissionsfaktoren für den Bereich 'Betriebe' beauftragt.

Primärer Zweck der Analyse war ein Vergleich der aus den erhobenen Daten berechneten empirischen Werte mit den Literaturwerten, die vom Institut für Verfahrenstechnik zur Bestimmung von Gesamtemissionsfaktoren verwendet wurden. Als Quelle diente dem Institut für Verfahrenstechnik dabei eine schweizer Studie aus dem Jahr 1995, die von dem Bundesamt für Umwelt, Wald und Landschaft (BUWAL) durchgeführt wurde (Emissionsfaktoren für stationäre Quellen, 1995).

Die Analyse der Energieverbrauchs- und Emissionsfaktoren wurde im Rahmen von zwei Diplomarbeiten durchgeführt. Während in dieser Arbeit das theoretische Modell zur Beschreibung von Energieverbrauchs- und Emissionsfaktoren behandelt wird (es zeigte sich, daß die relevanten Kenngrößen durch ein Lognormalverteilungsmodell beschrieben werden können) und eine Energieverbrauchsanalyse durchgeführt wird, analysiert G. GLAWAR (1997) in seiner Diplomarbeit die Emissionsfaktoren und führte einen Vergleich der Schätzwerte mit den Literaturwerten durch.

## 5.2 Beschreibung des Datenmaterials

Ausgehend von einer von der Wirtschaftskammer vorgegebenen branchenspezifischen Unterteilung der Unternehmungen in sogenannte Fachgruppen wurden vom Umweltamt des Magistrates Graz und dem Institut für Verfahrenstechnik jene Gruppen ausgewählt, die nach deren Ansicht als emissionsrelevant einzustufen waren.

Flußdiagramm 5.1 zeigt die formale Unterteilung der gesamten Betriebe in branchenspezifische Untergruppen:

Gesamt → Sektionen → Fachgruppen (FG) → Berufsgruppen (BG)

Flußdiagramm 5.1: Branchenspezifische Unterteilung der Betriebe

Um die Gruppenidentifikation zu erleichtern, wurden sämtliche Untergruppen numerisch kodiert:

1. Die Unterteilung Sektion umfaßt fünf Untergruppen:
   Sektion 1: Gewerbe und Handwerk
   Sektion 2: Industrie
   Sektion 3: Handel
   Sektion 5: Verkehr
   Sektion 6: Tourismus und Freizeitwirtschaft

2. Jede Sektion wird durch eine sogenannte Fachguppenunterteilung feiner gegliedert. Diese Unterteilung wird durch eine dreistellige Zahl kodiert, wobei die erste Stelle die übergeordnete Sektion angibt (Ein Betrieb mit der Fachgruppennummer 501 stammt zum Beispiel aus der Sektion 5).

3. Die Berufsgruppenunterteilung ist eine sehr feine branchenspezifische Zuordnung von Betrieben zu einzelnen Gruppen. Numerisch wird die Berufsgruppe durch eine zweistellige Zahl kodiert, wobei aus der Berufsgruppenkodierung kein Rückschluß auf die übergeordnete Fachgruppe (bzw. Sektion) möglich ist.

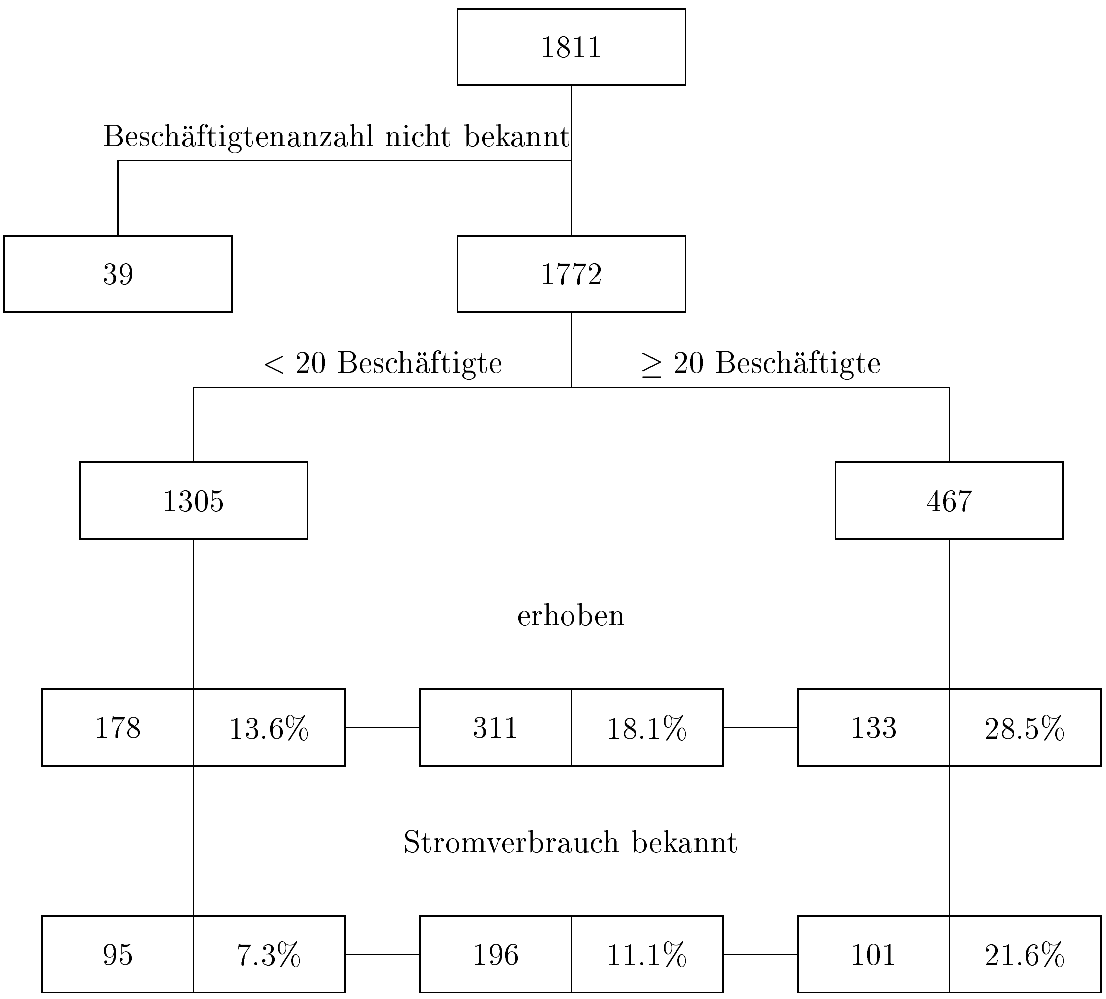
Beispiel:

| FG | Fachgruppenbezeichnung | BG | Berufsgruppenbezeichnung |
|---|---|---|---|
| 106 | Landesinnung der Maler, | 01 | Maler und Anstreicher |
| | Lackierer und Schriftenhersteller | 02 | Lackierer |
| | | 06 | Schildermaler und -hersteller |
| | | 10 | Sandstrahler |

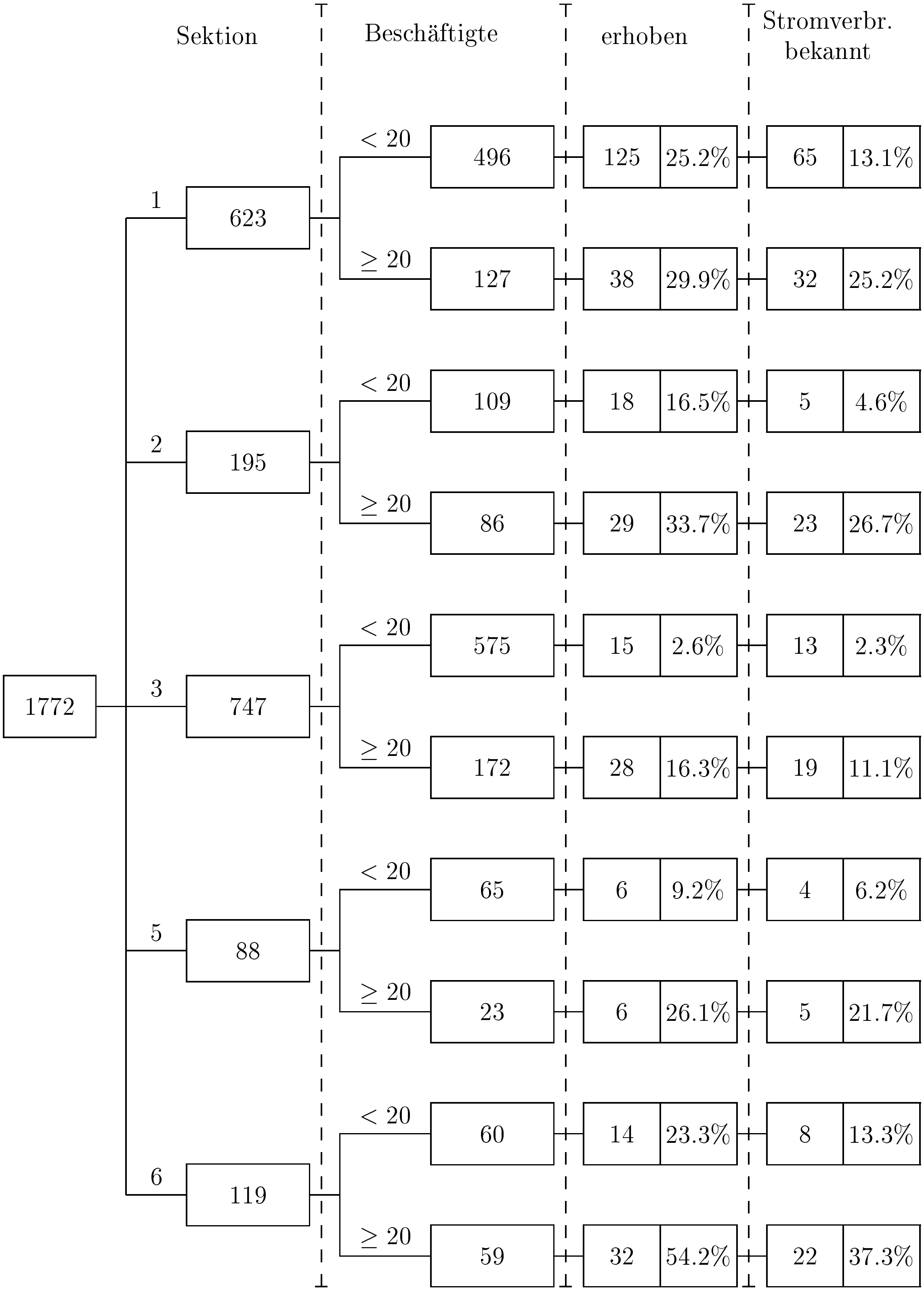
Anmerkung: Als feinste Unterteilung für die Emissionsrelevanz wurde die Fachgruppe gewählt, da für eine Aufteilung in Berufsgruppen keine ausreichenden Stichprobengrößen zur Verfügung standen.

Die Datenerhebung wurde größtenteils durch Aussendung von Fragebögen durchgeführt. Die konkrete Vorgangsweise, dabei aufgetretene Probleme (Nichtbeantwortung) und Problemlösungsvorschläge werden in der Diplomarbeit von G. Glawar (1997) geschildert.

Ausgehend von der Zuordnung der emissionsrelevanten Betriebe (1881 der insgesamt über 9000 Unternehmungen im Stadtgebiet von Graz wurden als emissionsrelevant eingestuft) lag nach der Befragung folgender Erhebungsstand vor:




Flußdiagramm 5.2: Anzahl erhobener Betriebe




Flußdiagramm 5.3: Anzahl erhobener Betriebe nach Sektionen aufgeteilt

Im folgenden werden die Variablen, die in dieser Arbeit in die Analyse einfließen, angegeben. Die emissionsbezogenen Variablen, die nicht zur Analyse der Energieverbrauchsfaktoren verwendet werden, sind in Glawar (1997) zu finden. Neben den betriebsspezifischen Angaben (Standort, Betreiber, etc.) wurden folgende Daten ermittelt:

| | |
|---|---|
| 1) FG-Nummer | 5) Stromverbrauch Sommer |
| 2) BG-Nummer | 6) Stromverbrauch Winter |
| 3) Mitarbeiteranzahl (gesamt) | 7) Fernwärmeverbrauch Sommer |
| 4) Mitarbeiteranzahl im Verwaltungsbereich | 8) Fernwärmeverbrauch Winter |

In den Flußdiagrammen 5.2 und 5.3 wurde neben der Anzahl der erhobenen Betriebe auch die Anzahl der Betriebe ausgewiesen, von denen der Stromverbrauch bekannt ist. Die Differenz zwischen den beiden Angaben läßt sich dadurch erklären, daß von einigen Betrieben der Fragebogen nicht vollständig ausgefüllt wurde. Da keine gesetzliche Verpflichtung zum Ausfüllen der Bögen bestand, konnte die teilweise geringe Kooperationsbereitschaft der Betriebe nur marginal beeinflußt werden (näheres dazu in der Diplomarbeit von G. Glawar (1997)).

Allgemein konnte festgestellt werden, daß die Variable Stromverbrauch auch ein Indikator dafür war, wie sorgfältig die restlichen Fragen beantwortet wurden. Dies war ein Grund, die entsprechenden Anzahlen in den Flußdiagrammen anzuführen.

Die Angaben zur Anzahl der Mitarbeiter im Verwaltungsbereich waren für eine Analyse unbrauchbar, da dies nur bei Vorort-Erhebungen erfragt wurde aber auf den Fragebögen gar nicht aufschien.

## 5.3 Aufgabenstellung

Primärer Zweck der Analyse war der Vergleich der empirischen Werte mit Literaturdaten, die im UIS zur Berechnung der Gesamtemissionen der Grazer Betriebe verwendet werden (siehe auch Glawar (1997)). Hierbei sind die Größen

1. Energieverbrauchsfaktor pro Mitarbeiter
2. Emissionsfaktor pro Mitarbeiter
3. Gesamtenergieverbrauch bzw. Gesamtemission aller Betriebe

von besonderem Interesse.

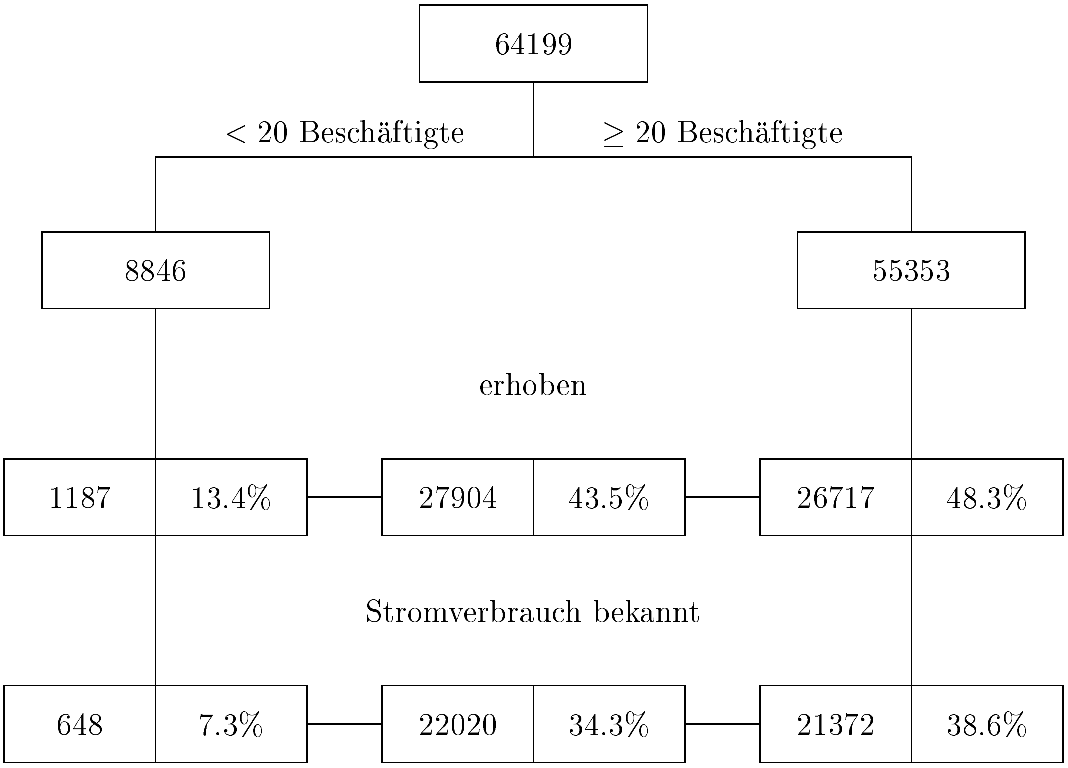
Eine weitere Aufgabe der Analyse war die Erstellung eines detaillierten Stichprobenplans, der für spätere Untersuchungen auf diesem Gebiet herangezogen werden soll.

## 5.4 Die Variable Stromverbrauch

Ziel unserer Untersuchungen ist es, Schätzungen über den Stromverbrauch pro Mitarbeiter und den Gesamtenergieverbrauch (aller emissionsrelevanten Betriebe) zu bekommen.

Es ist daher neben der Anzahl der erhobenen Betriebe auch die damit erfaßte Mitarbeiteranzahl von Relevanz, da zur Gewinnung von Aussagen über den Gesamtenergieverbrauch über alle Betriebe gewichtet nach der Mitarbeiteranzahl zu summieren ist.

In den Flußdiagrammen 5.4 und 5.5 wird der Stichprobenbestand bezüglich der erhobenen Mitarbeiter dargestellt:

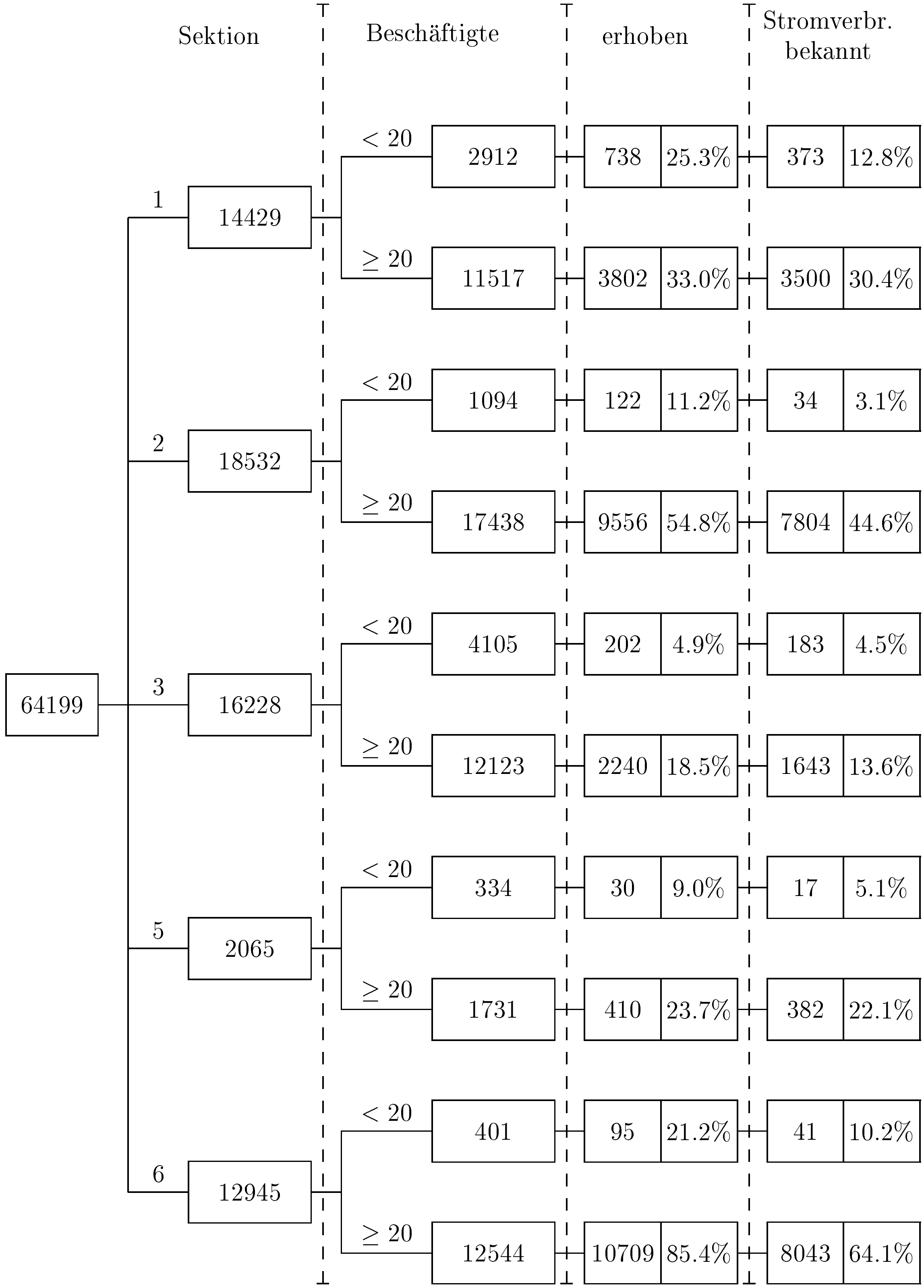



Flußdiagramm 5.4: Anzahl erhobener Mitarbeiter

Vergleicht man die Flußdiagramme 5.2 und 5.4, so erkennt man, daß die Relationen bzgl. erhobenener Mitarbeiter günstiger sind.

| | gesamt | < 20MA | ≥ 20MA |
|---|---|---|---|
| Erhobene Betriebe | 18.1% | 13.6% | 28.5% |
| Erhobene MA | 43.5% | 13.4% | 48.3% |
| Str. bekannt Betriebe | 11.1% | 7.3% | 21.6% |
| Str. bekannt MA | 34.3% | 7.3% | 38.6% |

Tabelle 5.1: Vergleich der Relation: erhoben Betriebe - erhobene Mitarbeiter




Flußdiagramm 5.5: Anzahl erhobener Mitarbeiter nach Sektionen aufgeteilt

Anmerkung: Die Unterteilung der Betriebe in die 2 Gruppen (weniger als 20 Mitarbeiter, 20 oder mehr Mitarbeiter) war uns vom Umweltamt vorgegeben.

Die Variable Stromverbrauch pro Mitarbeiter und Jahr (im folgenden kurz STR/MA) ist rechtsschief verteilt (siehe Abbildung 5.1). Es hat sich gezeigt, daß solche Daten mit Lognormalverteilungsmodellen beschrieben werden können.

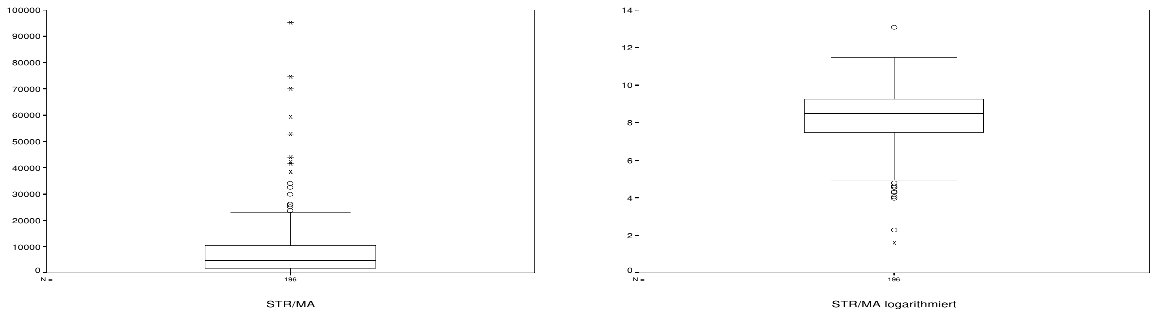



Abbildung 5.1: links: Boxplot von STR/MA (ohne Extremwert STR/MA=475250 kWh); rechts: Boxplot der logarithmierten Werte von STR/MA (inklusive des links ausgeschlossenen Extremwertes)

Eliminiert man die 12 Datensätze, die bei der Boxplotanalyse der logarithmierten Werte als Extremwerte ausgewiesen werden, so erhält man folgenden Boxplot der restlichen 184 Beobachtungen:

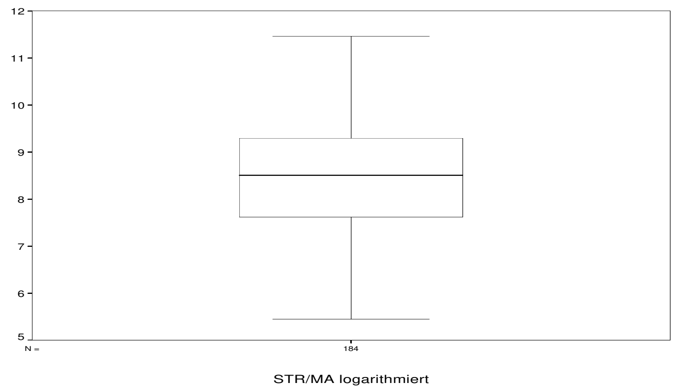



Abbildung 5.2: Boxplot der logarithmierten Werte von STR/MA ohne 12 Extremwerte)

Der Ausschluß der 12 Beobachtungen ist auch sachlich gerechtfertigt, da es sich entweder um offensichtliche Falschangaben (zum Beispiel Jahresstromverbrauch=100 kWh bei 20 Mitarbeitern), oder um für die ausgewählte Grundgesamtheit nicht repräsentative Betriebe handelt (z.B. eine Tankstelle mit einer SB-Waschstraße: Jahresstromverbrauch=475250 kWh bei einem Mitarbeiter).

Der Boxplot der logarithmierten Werte (Abbildung 5.2) deutet ebenfalls darauf hin, daß die beobachteten Werte durch ein (zweiparametrisches) Lognormalverteilungsmodell beschrieben werden können. Ein Lilliefors-Anpassungstest der logarithmierten Werte auf eine Normalverteilung (zum Niveau $\alpha = 0.05$) ergab, daß die Nullhypothese (=die logarithmierten Werte stammen aus einer Normalverteilung) nicht verworfen werden kann.

Anmerkung: Der Lilliefors-Anpassungtest auf Normalverteilung ist das Analogon zum Kolmogorov-Smirnov-Anpassungstest und wird verwendet, wenn die Verteilungsparameter der Normalverteilung aus der Stichprobe zu schätzen sind.

Es soll nun untersucht werden, ob die Verteilung der Variable STR/MA durch Gruppeneffekte beeinflußt wird. Die Werte in Abbildung 5.2 werden unabhängig von der Branche und der Mitarbeiteranzahl betrachtet. Wenn die Rechtsschiefe der empirischen Verteilung ein Resultat von Gruppeneffekten (bzgl. der Branche und/oder der Mitarbeiteranzahl) wäre, so würde man mit einem Lognormalverteilungsmodell zu falschen Schlüssen kommen, da sich in den einzelnen Sektionen und Mitarbeitergruppen einerseits die Grundgesamtheiten und andererseits die Stichprobenumfänge stark unterscheiden (siehe Flußdiagramm 5.3 und 5.5).

Es wird daher untersucht, ob die Variable Stromverbrauch pro Mitarbeiter und Jahr einerseits von der Betriebsgröße und andererseits von der Branche anhängig ist. Als Gruppierungsvariable für die Branchenzugehörigkeit wird die Sektion gewählt, da die Stichprobenumfänge bei einer feineren Unterteilung zu gering wären (siehe Flußdiagramm 5.3). Bezüglich der Betriebsgröße wird zwischen Betrieben mit weniger als 20 Mitarbeitern und solchen mit zumindest 20 Mitarbeitern unterschieden (diese Unterteilung wurde - wie schon erwähnt - vom Umweltamt vorgeschlagen und auch vom Institut für Verfahrentechnik verwendet).

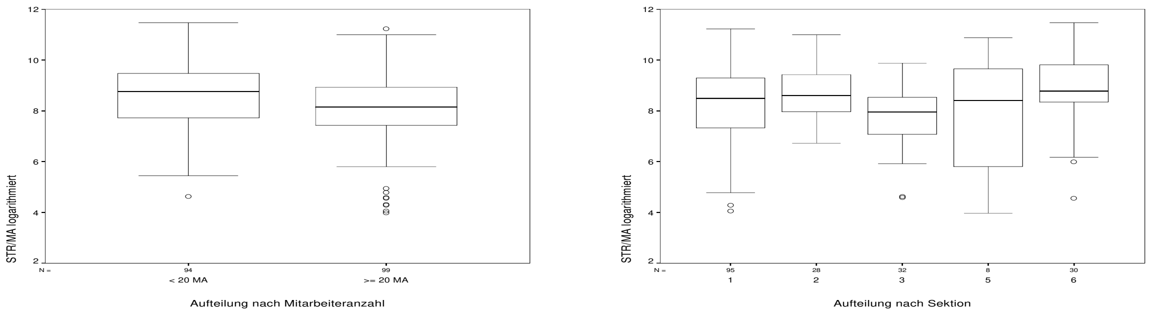



Abbildung 5.3: links: Boxplot der logarithmierten Werte von STR/MA augeteilt nach der Mitarbeiteranzahl; rechts: Boxplot der logarithmierten Werte von STR/MA aufgeteilt nach Sektionen

Anmerkung: In den Boxplots aus Abbildung 5.3 wurden nur 3 der 12 Extremwerte ausgeschlossen. Bei den 3 ausgeschlossenen Beobachtungen handelt es sich um 2 offensichtliche Falschangaben (Jahresstromverbrauch = 100 kWh bei 20 Mitarbeitern bzw. Jahrestromverbrauch = 2450 kWh bei 250 Mitarbeitern) und um einen Extremwert (STR/MA = 475250 kWh, wobei der zweitgrößte Wert bei STR/MA = 95167 kWh liegt).

Teilt man die Beobachtungen nach Sektionen oder Mitarbeiteranzahl auf, so scheint es keine starken Lokationsunterschiede zwischen den einzelnen Gruppen zu geben (t-Test auf Lokationsunterschied). Verwendet man allerdings die Sektionszugehörigkeit und die Mitarbeiteranzahl als Gruppierungsvariable, so erhält man folgendes Resultat:

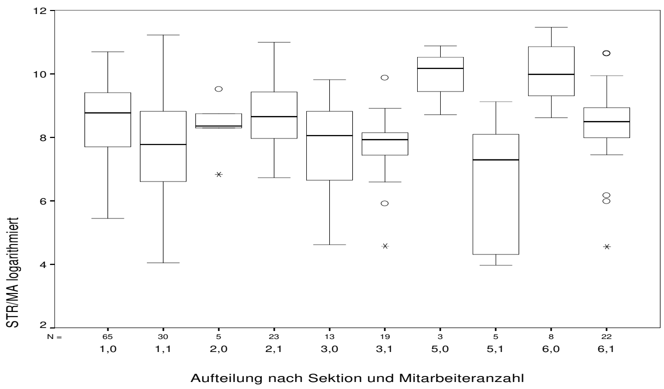



Abbildung 5.4: Boxplot der logarithmierten Werte von STR/MA nach Mitarbeiteranzahl und Sektionszugehörigkeit gruppiert

Die Gruppierungen sind wie folgt zu interpretieren: Die Einerstelle bezeichnet die übergeordnete Sektion und die Dezimalstelle bezieht sich auf die Mitarbeiteranzahl, wobei '0' Betriebe mit weniger als 20 Mitarbeitern und '1' Betriebe mit mindestens 20 Mitarbeitern kennzeichnet.

In Tabelle 5.2 ist die Zusammensetzung der Gruppenstichproben aus Betrieben bzgl. der Fachgruppenzugehörigkeit dargestellt. Beinahe die Hälfte der Beobachtungen aus der Gruppe 1.0 sind Betriebe aus der Fachgruppe 145 (Chem. Reinigung, Putzerei). Bei dieser Fachgruppe wurden 32 der 38 Betriebe über Interviews erhoben. Um eine Verfälschung der Ergebnisse für die anderen Betriebe aus der Gruppe 1.0 auszuschließen, wird die Fachgruppe 145 getrennt analysiert. Die Gruppen 5.0 und 5.1 setzen sich jeweils nur aus Beobachtungen einer Fachgruppe zusammen. Es wird angenommen, daß diese Beobachtungen für die Betriebe aus der Sektion 5 repräsentativ sind. Ähnliches gilt für die Beobachtungen aus Gruppe 6.0, wo 6 der 8 Beobachtungen aus der Fachgruppe 602 stammen. Die beiden restlichen Beobachtungen liegen jedoch nahe bei den Beobachtungen aus der FG 602 (die Extremwerte in beiden Richtungen sind Betriebe aus der FG 602). Es wird auch in dieser Gruppe angenommen, daß die Stichprobe die übergeordnete Sektion repräsentiert.

In Abbildung 5.5 ist der Boxplot der Beobachtungen aus der Gruppe 1.0 mit der Aufteilung FG 145 und Gruppe 1.0 ohne Beobachtungen aus der FG 145 dargestellt. Es ist ein Lokationsunterschied zwischen den beiden Untergruppen erkennbar, der die getrennte Analyse der Beobachtungen aus FG 145 zusätzlich rechtfertigt.

| Gruppe | Anzahl | Fachgruppe | Gruppe | Anzahl | Fachgruppe |
|---|---|---|---|---|---|
| 1.0 | 1 | 105, 115, 139 | 1.1 | 1 | 108, 111, 115, 134 |
| | 2 | 106, 120, 128, 142 | | | 135, 136, 139, 142 |
| | 4 | 109, 114B, 136, 140 | | | 153, 155 |
| | 8 | 121 | | 2 | 105, 120, 145 |
| | 30 | 145 | | 3 | 114B |
| | | | | 6 | 101, 121 |
| 2.0 | 1 | 203, 205, 209 | 2.1 | 1 | 219, 222, 223 |
| | 2 | 216 | | 2 | 203, 211, 218 |
| | | | | 3 | 207, 217 |
| | | | | 8 | 216 |
| 3.0 | 1 | 301A, 301B, 317 | 3.1 | 1 | 307, 313, 317, 319 |
| | | 323, 325 | | | 325, 331 |
| | 3 | 318 | | 2 | 308, 323, 326 |
| | 4 | 308 | | 3 | 316 |
| | | | | 4 | 318 |
| 5.0 | 3 | 510 | 3.1 | 5 | 507 |
| 6.0 | 1 | 601, 604 | 6.1 | 1 | 607, 609 |
| | 6 | 602 | | 5 | 602 |
| | | | | 6 | 601 |
| | | | | 9 | 603 |

Tabelle 5.2: Zusammensetzung der Gruppenstichproben

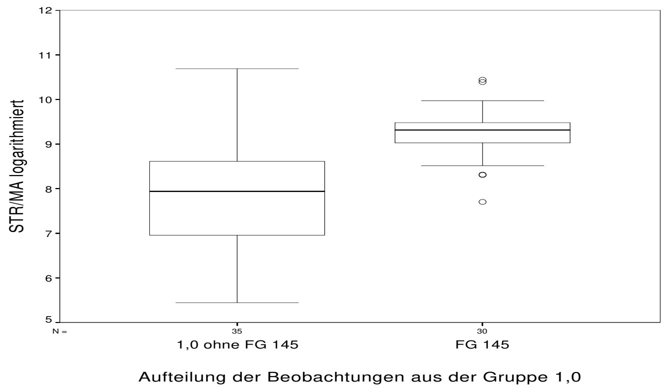



Abbildung 5.5: Boxplot der logarithmierten Werte von STR/MA aus der Gruppe 1.0 mit Aufteilung der Beobachtungen nach FG 145

Bei den Beobachtungen aus der Gruppe 1.1 werden jene der FG 145 ebenfalls nicht berücksichtigt.
Anmerkung: Da nur zwei Betriebe aus der Fachgruppe 145 mehr als 20 Mitarbeiter haben, beide aber erhoben wurden, ist eine Aufteilung der Beobachtungen aus der FG 145 mit $< 20$ bzw. $\geq 20$ Mitarbeiter nicht nötig. Diese beiden Beobachtungen werden daher von folgenden Berechnungen ausgeschlossen.

Folgende Analysen basieren auf der Annahme, daß die Beobachtungen für die entsprechenden Gruppen repräsentativ sind. Es resultieren folgende Gruppen, die analysiert werden (mit 0.0 sind die Beobachtungen aus der FG 145 gekennzeichnet):

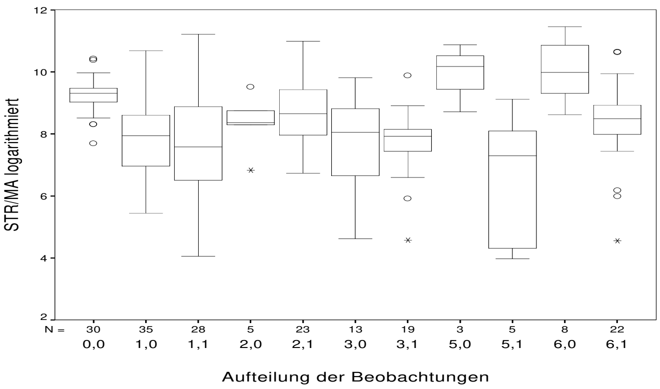



Abbildung 5.6: Boxplot der logarithmierten Werte von STR/MA nach Mitarbeiteranzahl, Sektionszugehörigkeit und FG 145 aufgeteilt

Zwischen einigen Gruppen lassen sich deutliche Lokationsunterschiede erkennen. Um die Entscheidung, ob zwischen jeweils zwei Gruppen ein signifikanter Unterschied besteht, nicht nur aufgrund der Boxplotanalyse zu treffen, wird ein t-Test auf Lokationsunterschiede durchgeführt. Mit dem t-Test wird geprüft, ob in einer Sektion ein signifikanter Lokationsunterschied zwischen Betrieben mit $< 20$ bzw. $\geq 20$ Mitarbeitern festzustellen ist, wobei aufgrund des Ergebnisses des Levene-Tests auf Dispersionsunterschied entweder gleiche oder ungleiche Varianzen der zu testenden Guppen vorausgesetzt werden. Ein Lilliefors-Anpassungstest der logarithmierten Daten auf Normalverteilung zum Niveau $\alpha = 0.05$ brachte für alle Gruppen das Ergebnis, daß die Nullhypothese (= die logarithmierten Daten stammen aus einer Normalverteilung) nicht zu verwerfen ist.

In Tabelle 5.3 sind die Ergebnisse der t-Tests angeführt. Die 2-seitigen Signifikanzwerte geben an, zu welcher Irrtumwahrscheinlichkeit die Nullhypothese (= keine Lokationsunterschiede zwischen den Gruppen) gerade noch abgelehnt werden könnte (meist wird ein Niveau $\alpha = 0.05$ verwendet).

| | 1.1 | 2.1 | 3.1 | 5.1 | 6.1 |
|---|---|---|---|---|---|
| 1.0 | 0.66 | | | | |
| 2.0 | | 0.43 | | | |
| 3.0 | | | 0.79 | | |
| 5.0 | | | | 0.06 | |
| 6.0 | | | | | 0.004 |

Tabelle 5.3: Ergebnisse der paarweisen t-Tests auf Lokationsunterschiede

Ausgehend von den Ergebnissen der t-Tests und des Boxplots aus Abbildung 5.6 wird eine Gruppierung nach der Variablen Sektion vorgenommen; Diagnostiziert der t-Test einen signifikanten Lokationsunterschied zwischen den Betrieben ($<$ 20 MA und $\geq$ 20 MA), wird innerhalb der Sektion noch einmal gruppiert. Es ergeben sich daher folgende Gruppen:

1. Gruppe A : 0.0 Putzereien mit $<$ 20 Mitarbeiter
2. Gruppe B : 1.0 und 1.1
3. Gruppe C : 2.0 und 2.1
4. Gruppe D : 3.0 und 3.1
5. Gruppe E : 5.0
6. Gruppe F : 5.1
7. Gruppe G : 6.0
8. Gruppe H : 6.1

Aufgrund der geringen Anzahl von Beobachtungen in Sektion 5 wird diese Sektion (Gruppen E und F) auf andere Gruppen aufgeteilt. Als Kriterium wird der geringste Lokationsunterschied zu anderen Gruppen (unter Berücksichtigung, daß keine Verletzung der Annahme der Varianzgleichheit vorliegt) herangezogen. Damit ergibt sich die neue Aufteilung in folgende 6 Gruppen:

1. Gruppe 1 : 0.0
2. Gruppe 2 : 1.0, 1.1 und 5.1
3. Gruppe 3 : 2.0 und 2.1
4. Gruppe 4 : 3.0 und 3.1
5. Gruppe 5 : 5.0 und 6.0
6. Gruppe 6 : 6.1

Es soll nun geprüft werden, ob die Variable STR/MA innerhalb der Gruppierung von der Mitarbeiteranzahl abhängig ist, d.h. ob die Variablen STR/MA und Mitarbeiteranzahl korreliert sind. In Tabelle 5.4 sind die Pearson-Korrelationskoeffizienten (P) und die Spearman-Rangkorrelationskoeffizienten (S) für die nach Gruppenzugehörigkeit zusammengefaßten Daten angegeben:

| | Anzahl | P | S |
|---|---|---|---|
| Gruppe 1 | 30 | −0.12 | −0.19 |
| Gruppe 2 | 68 | 0.29 | −0.07 |
| Gruppe 3 | 28 | 0.20 | 0.13 |
| Gruppe 4 | 32 | −0.22 | −0.09 |
| Gruppe 5 | 11 | −0.29 | −0.30 |
| Gruppe 6 | 22 | −0.01 | −0.01 |

Tabelle 5.4: Pearson- und Spearman-Korrelationskoeffizienten für die Variablen STR/MA und Mitarbeiteranzahl

In allen Gruppen ist der Korrelationskoeffizient klein, es kann daher auf Unkorreliertheit der beiden Variablen geschlossen werden. Auffällig ist, daß das Vorzeichen der Korrelationskoeffizienten nicht für alle Gruppen einheitlich ist, was ebenfalls auf die Unkorreliertheit der beiden Variablen hindeutet.

Es wird daher die Variable STR/MA getrennt nach den Gruppen 1 bis 6 über ein Verteilungsmodell beschrieben. Für jede Gruppe werden Konfidenzintervalle für den Erwartungswert und den Median konstruiert, sowie Bereiche für den erwarteten Gesamtenergieverbrauch pro Gruppe und für alle emissionsrelevanten Betriebe angegeben.

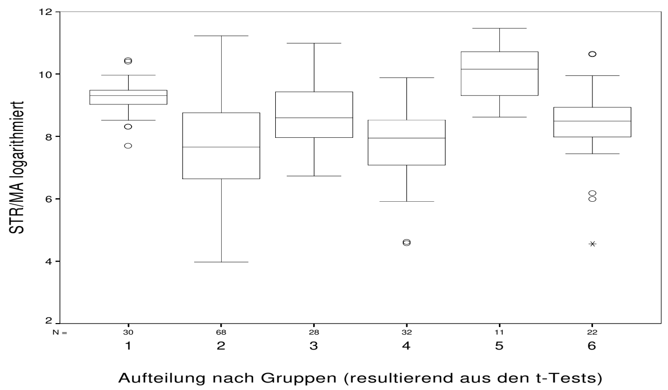



Abbildung 5.7: Boxplot der logarithmierten Werte von STR/MA nach Gruppen 1-6 aufgeteilt

Abbildung 5.7 zeigt den Boxplot der logarithmierten Werte von STR/MA aufgeteilt nach den Gruppen 1 bis 6. Aufgrund der Boxplotanalyse und einer darauf aufbauenden Datenbegutachtung werden 8 weitere Beobachtungen von den Berechnungen ausgeschlossen, da sie für die jeweilige Gruppe nicht repräsentativ sind; zwei Beobachtungen aus

Gruppe 4 (Firmenidenifikationsnummern 1082 und 1220), eine Beobachtung aus Gruppe 6 (Firmen-ID 1222) und fünf Beobachtungen aus Gruppe 1 (Firmen-ID 1773, 1782, 1784, 1797 und 1798) werden von den folgenden Berechnungen ausgeschlossen. Die Anzahl der gesamten Beobachtungen reduziert sich daher auf 183.

Für jede Gruppe wird wiederum ein Lilliefors-Anpassungstest auf Normalverteilung zum Niveau $\alpha = 0.05$ durchgeführt, wobei in allen Fällen die Nullhypothese (= die logarithmierten Daten stammen aus einer Normalverteilung) nicht verworfen wird.

Dies rechtfertigt eine Beschreibung der gruppierten Daten mit Lognormalverteilungsmodellen. Aufgrund der empirischen Schiefe in den jeweiligen Gruppen werden die Verteilungsparameter mit der (zweiparametrischen) Maximum-Likelihood-Methode geschätzt (siehe Kapitel 2), wobei bei der Schätzung von $\sigma^2$ nicht der Skalierungsfaktor $\frac{1}{n}$ sondern der Faktor $\frac{1}{n-1}$ verwendet wird.

In Tabelle 5.5 sind die Schätzer für die Normalverteilungsparameter $\mu$ und $\sigma$ der logarithmierten Werte der Variablen Stromverbrauch pro Mitarbeiter und Jahr angegeben:

| Gruppe | Betriebe | $\hat{\mu}$ | $\hat{\sigma}$ | n |
|---|---|---|---|---|
| 1 | Putzereien mit $<$ 20 Mitarbeitern | 9.3134 | 0.3687 | 25 |
| 2 | Sektion 1 und 5 mit $\geq$ 20 Mitarbeitern | 7.5897 | 1.5879 | 68 |
| 3 | Sektion 2 | 8.7084 | 1.0960 | 28 |
| 4 | Sektion 3 | 7.9682 | 0.9967 | 30 |
| 5 | Sektion 5 und 6 mit je $<$ 20 Mitarbeitern | 10.1025 | 0.9685 | 11 |
| 6 | Sektion 6 mit $\geq$ 20 Mitarbeitern | 8.5271 | 1.1423 | 21 |

Tabelle 5.5: Schätzungen der Verteilungsparameter für ln STR/MA pro Gruppe

In Tabelle 5.6 sind die Schätzer für den Erwartungswert (E), die Standardabweichung (SD) und den Variationskoeffizienten (SD/E) der Variablen STR/MA gruppenweise zusammengefaßt:

| Gruppe | E | SD | SD/E |
|---|---|---|---|
| 1 | 11865 | 4528 | 0.3816 |
| 2 | 6977 | 23605 | 3.3833 |
| 3 | 11037 | 16826 | 1.5245 |
| 4 | 4745 | 6188 | 1.3041 |
| 5 | 35650 | 44453 | 1.2469 |
| 6 | 9697 | 15895 | 1.6392 |

Tabelle 5.6: Schätzungen für den Erwartungswert, die Standardabweichung und den Variationskoeffizienten der Variablen STR/MA pro Gruppe (Angaben in kWh)

Unter Annahme einer lognormalverteilten Zufallsvariablen STR/MA kann ein approximatives Konfidenzintervall für den Erwartungswert nach einer der folgenden 3 Methoden konstruiert werden (siehe Kapitel 3): Methode von Cox, konservative Methode von Angus oder parametrische Bootstrapmethode von Angus.

In Tabelle 5.7 sind die Schätzungen und die approximativen Konfidenzintervalle für die Gruppenerwartungswerte der Variablen STR/MA, nach obigen drei Methoden, angeführt. Alle Konfidenzintervalle sind Intervalle zum Niveau $1 - \alpha = 0.95$, für die parametrische Bootstrapmethode wurde die Anzahl der Replikationen mit 2000 festgelegt. Die zur Konstruktion benötigten Zufallszahlen (aus Normal- und $\chi^2$-Verteilungen) wurden mit `Mathematica` generiert:

| Gruppe | n | E | | Bootstrap | % E |
|---|---|---|---|---|---|
| 1 | 25 | 11865 | | [10365; 13983] | [ 87%;118% ] |
| 2 | 68 | 6977 | | [ 4313; 13573] | [ 62%;195% ] |
| 3 | 28 | 11037 | | [ 7191; 20343] | [ 65%;184% ] |
| 4 | 30 | 4745 | | [ 3276; 7911] | [ 69%;167% ] |
| 5 | 11 | 35650 | | [20614; 92465] | [ 58%;259% ] |
| 6 | 21 | 9697 | | [ 5824; 21198] | [ 60%;219% ] |
| Gruppe | n | Methode von Cox | % E | konserv. Meth. | % E |
| 1 | 25 | [10217;13779] | [ 86%;116% ] | [10138; 15265] | [ 85%;129% ] |
| 2 | 68 | [ 3946;12336] | [ 57%;177% ] | [ 3914; 14872] | [ 56%;213% ] |
| 3 | 28 | [ 6580;18512] | [ 60%;168% ] | [ 6447; 25465] | [ 58%;231% ] |
| 4 | 30 | [ 3059; 7359] | [ 64%;155% ] | [ 3010; 9517] | [ 63%;201% ] |
| 5 | 11 | [17620;72126] | [ 49%;202% ] | [16202;200345] | [ 45%;562% ] |
| 6 | 21 | [ 5143;18283] | [ 53%;189% ] | [ 4970; 29924] | [ 51%;309% ] |

Tabelle 5.7: Schätzung für den Erwartungswert der Variable STR/MA; approximative Konfidenzintervalle für den Erwartungswert ($1 - \alpha = 0.95$) (Angaben in kWh)

Aufgrund des geringen Stichprobenumfanges in den jeweiligen Gruppen ist nach Zhou and Gao (1997) die parametrische Bootstrapmethode von Angus zu empfehlen. Die beiden anderen Methoden wurden zu Vergleichszwecken angegeben. Die konservative Methode von Angus liefert stets Konfidenzintervalle, welche die Intervalle aus den beiden anderen Methoden überdecken.

In Tabelle 5.8 sind Schätzungen und 95%- Konfidenzintervalle für den Median der Zufallsvariablen STR/MA angegeben. Die 95%-Konfidenzintervalle für den Median wurden mit der exakten Methode von Land konstruiert (siehe Kapitel 3):

| Gruppe | Schätzung für den Median | KIV |
|---|---|---|
| 1 | 11085 | [ 9521;12908] |
| 2 | 1978 | [ 1347; 2905] |
| 3 | 6054 | [ 3958; 9259] |
| 4 | 2888 | [ 1990; 4190] |
| 5 | 22304 | [11636;42751] |
| 6 | 5050 | [ 3002; 8494] |

Tabelle 5.8: Schätzung für den Median der Variable STR/MA; Konfidenzintervall für den Median ($\alpha = 0.05$) (Angaben in kWh)

Es sind deutliche Unterschiede in der Lokation der einzelnen Gruppen erkennbar. Gruppe 5 zeichnet sich durch einen besonders hohen Jahresstromverbrauch pro Mitarbeiter aus, die Gruppen 2 und 4 durch einen besonders geringen.

Eine Analyse des Gesamtstromverbrauchs ist äquivalent zur Analyse der Variablen 'Summe der Jahresstromverbrauchsfaktoren aller Betriebe'. Da vom Großteil der betrachteten Betriebe zwar die Mitarbeiteranzahl, aber nicht die Jahresstromverbrauchsfaktoren (bei 1576 (88.9%) der 1772 Betriebe sind nur die Mitarbeiteranzahlen bekannt) gegeben sind, muß der Jahresstromverbrauch dieser Betriebe über die Variable STR/MA und der Mitarbeiteranzahl geschätzt werden. In Tabelle 5.9 ist die Anzahl der Betriebe und die daraus resultierende Anzahl der Mitarbeiter angeführt, von der keine Angaben bezüglich des Jahresstromverbrauches bekannt sind:

| Gruppe | Anzahl der Betriebe | Mitarbeiteranzahl |
|---|---|---|
| 1 | 6 | 9 |
| 2 | 538 | 11896 |
| 3 | 167 | 10694 |
| 4 | 715 | 14402 |
| 5 | 113 | 677 |
| 6 | 37 | 4501 |

Tabelle 5.9: Betriebe ohne Stromverbrauchsangaben

Bezeichne $X$ die Variable STR/MA. Unter der Annahme, daß $X_{i,j} \overset{iid}{\sim} \Lambda\left(\mu_j, \sigma_j^2\right)$, wobei $j = 1, \ldots, 6$ der Guppenindex ist, gilt (siehe Kapitel 4):

$$a_{i,j} X_{i,j} \sim \Lambda\left(\ln(a_{i,j}) + \mu_j, \sigma_j^2\right),$$

wobei $a_{i,j}$ die Mitarbeiteranzahl des $i$-ten Betriebes in der $j$-ten Gruppe bezeichnet.

Summiert man nun über die Betriebe in der jeweiligen Gruppe auf, so gilt nach den Aussagen aus Kapitel 4 und Fenton (1960): Die Summe von unabhängigen lognormalverteilten Zufallsvariablen ist approximativ lognormalverteilt. Die Summenverteilung wird durch eine Lognormalverteilung approximiert, die in den ersten beiden Momenten mit der Summenverteilung übereinstimmt. Die Verteilungsparameter werden aus den Schätzungen der Verteilungsparameter der einzelnen Summanden berechnet. Es ergeben sich folgende Schätzungen der Verteilungsmomente der Summenverteilung (ohne Betriebe, bei denen der Jahresstromverbrauch bekannt ist):

| Gruppe | n | E | SD | SD/E |
|---|---|---|---|---|
| 1 | 6 | 106 | 18 | 0.1698 |
| 2 | 538 | 82813 | 10723 | 0.1295 |
| 3 | 167 | 118030 | 34790 | 0.2948 |
| 4 | 715 | 68342 | 9090 | 0.1330 |
| 5 | 113 | 24135 | 3666 | 0.1519 |
| 6 | 37 | 43644 | 22094 | 0.5063 |

Tabelle 5.10: Schätzungen für den Erwartungswert (E), die Standardabweichung (SD) und den Variationskoeffizienten des Gesamtstromverbrauches pro Gruppe (Angaben in MWh); Betriebe, bei denen der Jahresstromverbrauch bekannt ist, sind nicht enthalten.

Unter der Annahme, daß die Summenverteilung approximativ lognormalverteilt mit den geschätzten Parametern ist, kann folgender approximative 95%-Bereich (2.5%-Quantil und 97.5%-Quantil) für den Gesamtenergieverbrauch pro Gruppe angegeben werden (bekannte Jahresstromverbrauchsfaktoren inkludiert):

| Gruppe | n | Betriebe | untere Gr. | obere Gr. |
|---|---|---|---|---|
| 1 | 38 | Betriebe aus der FG 145 | 1181 | 1250 |
| 2 | 608 | Sektion 1 und 5 mit $\geq$ 20 Mitarbeitern | 84424 | 126378 |
| 3 | 195 | Sektion 2 | 169507 | 304547 |
| 4 | 747 | Sektion 3 | 60640 | 96203 |
| 5 | 125 | Sektion 5 und 6 mit je $<$ 20 Mitarbeitern | 21067 | 35401 |
| 6 | 59 | Sektion 6 mit $\geq$ 20 Mitarbeitern | 55603 | 139637 |

Tabelle 5.11: 95%-Bereich für den Gesamtstromverbrauch pro Gruppe (Angaben in MWh); Betriebe, bei denen der Jahresstromverbrauch bekannt ist, sind enthalten.

Um Aussagen über den Gesamtenergieverbrauch ohne Gruppenunterteilung treffen zu können, werden wiederum die Ergebnisse der Untersuchungen aus Kapitel 4 und Fenton (1960) verwendet: Falls die Verteilung der Zufallsvariablen Jahresstromverbrauch in jeder Gruppe approximativ lognormalverteilt ist, ist auch die Summe der sechs unabhängigen Zufallsvariablen approximativ lognormalverteilt. Unter diesen Voraussetzungen ergeben sich folgende Schätzungen für die Verteilungsmomente der Summenverteilung.

Daraus ergibt sich folgender 95%-Bereich für den Jahresstromverbrauch aller 1772 Betriebe (Konstruktion aus dem 2.5%- und 97.5%-Quantil):

| Erwartungswert | Standardabweichung |
|---|---|
| 337070 | 43698 |

Tabelle 5.12: Schätzungen für den Erwartungswert und die Standardabweichung des Gesamtstromverbrauches (Angaben in MWh); Betriebe, bei denen der Jahresstromverbrauch bekannt ist, sind nicht enthalten.

| untere Grenze | obere Grenze |
|---|---|
| 438528 | 609505 |

Tabelle 5.13: 95%-Bereich für den Gesamtstromverbrauch (Angaben in MWh); Betriebe, bei denen der Jahresstromverbrauch bekannt ist, sind enthalten.

### 5.4.1 Zusammenhang zwischen Stromverbrauch-Sommer und -Winter

Die Analyse eines Zusammenhangs zwischen den Variablen Stromverbrauch-Sommer und Stromverbrauch-Winter konnte nicht angestellt werden, da bei der Erhebung in einigen Fällen nur der Jahresstromverbrauch angegeben wurde. Dieser Wert wurde bei der Dateneingabe auf die beiden Perioden aufgeteilt. Auch bei Rücksendungen, die den Verbrauch auf die beiden Perioden aufgeteilt enthielten, wurde allem Anschein nach oft der Jahresverbrauch auf die Perioden aufgeteilt (meist halbiert). Da im Nachhinein nicht mehr nachvollziehbar ist, bei welchen Betrieben die Angaben Stromverbrauch-Sommer und -Winter exakt sind, würde eine Analyse mit dem vollständigen Datensatz wahrscheinlich zu verfälschten Ergebnissen führen und daher wenig Aussagekraft besitzen.

### 5.4.2 Bestimmung des benötigten Stichprobenumfanges

In diesem Abschnitt soll untersucht werden, inwieweit sich einerseits die Konfidenzintervalle der Variablen STR/MA und andererseits der geschätzte Realisationsbereich für die Variable 'Gesamtenergieverbrauch aller Betriebe' mit wachsendem Stichprobenumfang ändert.

Der Auftraggeber, das Umweltamt der Stadt Graz, hat uns keine Genauigkeitsgrenzen für die beiden Variablen vorgegeben, daher wird die Analyse des benötigten Stichprobenumfanges in einer etwas atypischen Weise durchgeführt:

#### Die Variable Stromverbrauch pro Mitarbeiter und Jahr (STR/MA)

Aufbauend auf den Beobachtungen, den zugrundegelegten Verteilungsannahmen und den dadurch gewonnenen Schätzungen und Konfidenzintervallen wird gruppenweise untersucht, in welcher Weise sich das Konfidenzintervall für den Erwartungswert unter der Annahme, daß die Schätzungen der Verteilungsparameter für alle betrachteten Stichprobenumfänge gleich bleiben, ändert. Die resultierenden Konfidenzgrenzen werden graphisch in Abhängigkeit vom Stichprobenumfang dargestellt. Anhand der Steigung der Grenzgeraden kann entschieden werden, ob eine Erhöhung des Stichprobenumfanges einen Genauigkeitsgewinn bringt, der in einem akzeptablen Verhältnis zum dadurch entstehenden Aufwand (Zeitaufwand, zusätzliche Kosten) steht.

Zur Konstruktion eines approximativen Konfidenzintervalles für den Erwartungswert der Zufallsvariablen STR/MA wird nicht die parametrische Bootstrapmethode von Angus, sondern die von Zhou and Gao (1997) empfohlene Methode von Cox verwendet. Bei der parametrischen Bootstrapmethode sind die Quantile der Statistik, die zur Konstruktion eines approximativen Konfidenzintervalles herangezogen werden, vom Stichprobenumfang abhängig. Dies ist bei der Methode von Cox nicht der Fall, was die Berechnung der hypothetischen approximativen Konfidenzintervalle erheblich vereinfacht.

In Abbildung 5.8 sind für jede Gruppe die approximativen Konfidenzintervalle in Abhängigkeit vom Stichprobenumfang abgebildet. Diese Graphiken sind als Entscheidungshilfen für die Wahl des Stichprobenumfanges (bzw. für die Entscheidung, ob der Stichprobenumfang erhöht werden soll) zu verstehen. Es kann natürlich zu Verschiebungen der prognostizierten Intervalle kommen, wenn sich die Schätzungen der Verteilungsparameter für größerer Stichprobenumfänge gegenüber den hier angenommenen Werten ändern sollten. Man hat aber jedenfalls qualitative Anhaltspunkte über die Änderung der Schwankungsbereiche.

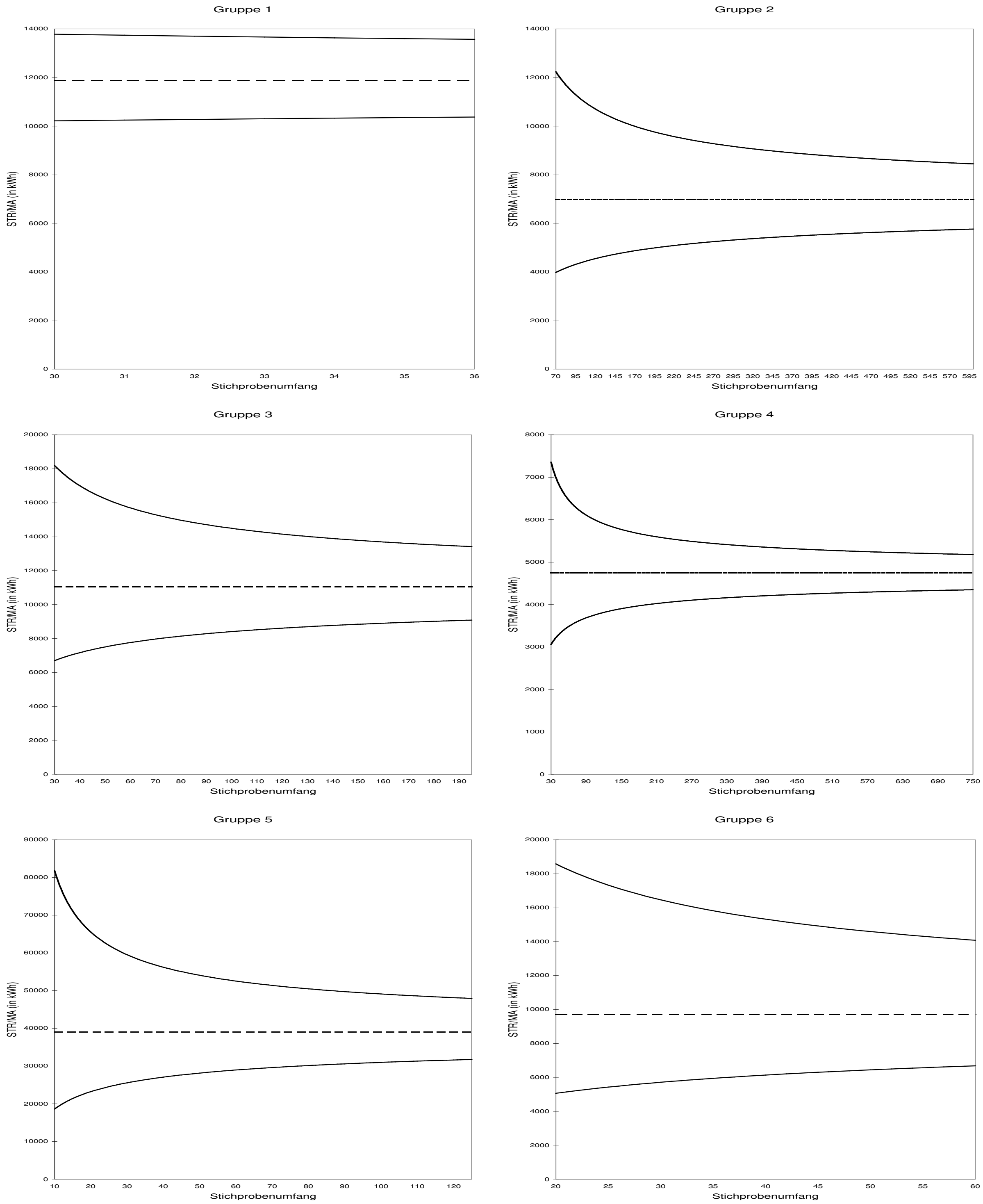



Abbildung 5.8: Veränderung der Konfidenzintervalle für den Erwartungswert der Variablen STR/MA (gruppenweise) mit wachsendem Stichprobenumfang (bezüglich erhobener Betriebe) unter den getroffenen Annahmen; der geschätzte Erwartungswert ist strichliert gekennzeichnet.

Anmerkung zu Abbildung 5.8: Der Anfangspunkt der $x$-Achse (Stichprobenumfang) ist der aktuelle Erhebungsstand, der Endpunkt der $x$-Achse ist die Anzahl aller Grazer Betriebe (in der jeweiligen Gruppe). Der geschätzte Erwartungswert der Zufallsvariablen STR/MA ist durch die strichlierte Linie gekennzeichnet; die durchgezogenen Linien bezeichnen unter den getroffenen Annahmen die prognostizierte untere bzw. obere Intervallgrenze. Der Umstand, daß die Konfidenzintervalllänge selbst bei einer Vollerhebung größer als Null ist, ist dadurch erklärbar, daß die Grundgesamtheit nicht auf Betriebe aus dem Raum Graz eingeschränkt ist.

In den Gruppen 2 bis 6 ist der Abfall der (oberen) Grenzgeraden gravierend, d.h. eine Erhöhung des Stichprobenumfanges würde hier einen markanten Genauigkeitsgewinn bringen. Man beachte, daß die Skalierungen der $x$- und $y$-Achsen in den Plots verschieden sind, da diese von der Schätzung des Gruppenerwartungswertes und dem verbleibenden Stichprobenumfang der jeweiligen Gruppe abhängig sind. Am stärksten ist die Veränderung des prognostizierten Intervalles mit wachsendem Stichprobenumfang in der Gruppe 5 (nur 11 erhobene Betriebe).

### Jahresstromverbrauch aller n=1772 relevanten Betriebe

Aufbauend auf den Beobachtungen, den Verteilungsannahmen und den dadurch gewonnenen Schätzungen wird gruppenweise untersucht, in welcher Weise sich der angegebene 95%-Bereich für den Jahresstromverbrauch aller Betriebe der jeweiligen Gruppe ändert. Wir gehen dabei wieder von der Annahme aus, daß die Schätzungen der Verteilungsparameter der Variablen STR/MA für alle betrachteten Stichprobenumfänge gleich bleiben.

Die 95%-Bereiche für den Jahresstromverbrauch aller Betriebe pro Gruppe werden über die Verteilungsannahmen der Variablen STR/MA konstruiert. Dabei wird von folgenden Voraussetzungen ausgegangen:

1. Die Variable STR/MA ist für jede Gruppe lognormalverteilt mit den aus den erhobenen Stichproben geschätzten Verteilungsparametern: Wenn man mit $X$ die Variable STR/MA bezeichnet, gilt daher:
   $X_{i,j} \overset{iid}{\sim} \Lambda\left(\mu_j, \sigma_j^2\right)$, wobei $j = 1, \ldots, 6$ der Gruppenindex ist.
2. Ist $a_{i,j}$ die Mitarbeiteranzahl des $i$-ten Betriebes in der $j$-ten Gruppe, dann gilt unter 1.:
   $a_{i,j} X_{i,j} \sim \Lambda\left(\ln(a_{i,j}) + \mu_j, \sigma_j^2\right)$
3. Aus 1. und 2. folgt:
   $\sum_i a_{i,j} X_{i,j}$ ist approximativ lognormalverteilt, wobei die Approximationsverteilung mit der Summenverteilung in den ersten beiden Momenten übereinstimmt.
4. Der Bereich, in dem der Wert 'Jahresstromverbrauch aller Betriebe' (der jeweiligen Gruppe) zu 95% realisiert, kann über das 0.025- und das 0.975-Quantil der approximierten Lognormalverteilung angegeben werden.

Es sollen nun Aussagen darüber getroffen werden, inwieweit sich der 95%-Bereich mit wachsendem Stichprobenumfang (bezüglich der Anzahl der erhobenen Betriebe) ändert. Da neben der Anzahl der Summanden auch die Gewichtung der Summanden (= die Mitarbeiteranzahl) von Bedeutung ist, wird folgende Vorgehensweise gewählt.

1. Die Veränderung des 95%-Bereiches mit wachsendem Stichprobenumfang wird graphisch analog zum vorigen Abschnitt dargestellt.

2. Auf der $x$-Achse wird die Anzahl der noch nicht erhobenen Betriebe (= Anzahl der zu berechnenden Betriebe) aufgetragen, wobei der Startpunkt wiederum der derzeitige Erhebungsstand und der Endpunkt die Vollerhebung ist.

3. Da, wie bereits erwähnt, neben der Anzahl der erhobenen Betriebe auch die Anzahl der dadurch erfassten Mitarbeiter von Bedeutung ist, wird folgendes festgelegt: Die nicht erhobenen Betriebe werden gruppenweise nach der Mitarbeiteranzahl absteigend geordnet und in dieser Reihenfolge fiktiv erhoben. Es wird also untersucht, inwieweit sich der 95%-Bereich für den Gesamtjahresstromverbrauch pro Gruppe verändert, wenn die Betriebe ihrer Größe nach erhoben werden.

4. Auf der $y$-Achse der linken Graphiken werden die Schätzungen für den Erwartungswert zusammen mit den 0.025- und 0.975-Quantilen der Variablen Jahresstromverbrauch aller relevanter Betriebe aufgetragen.

5. An den rechten Graphiken wird gezeigt, wie sich der Erhebungsstand bezüglich erhobener Mitarbeiter ändert.

In den Abbildungen 5.9 und 5.10 sind die Ergebnisse für die sechs Gruppen zusammengefaßt.

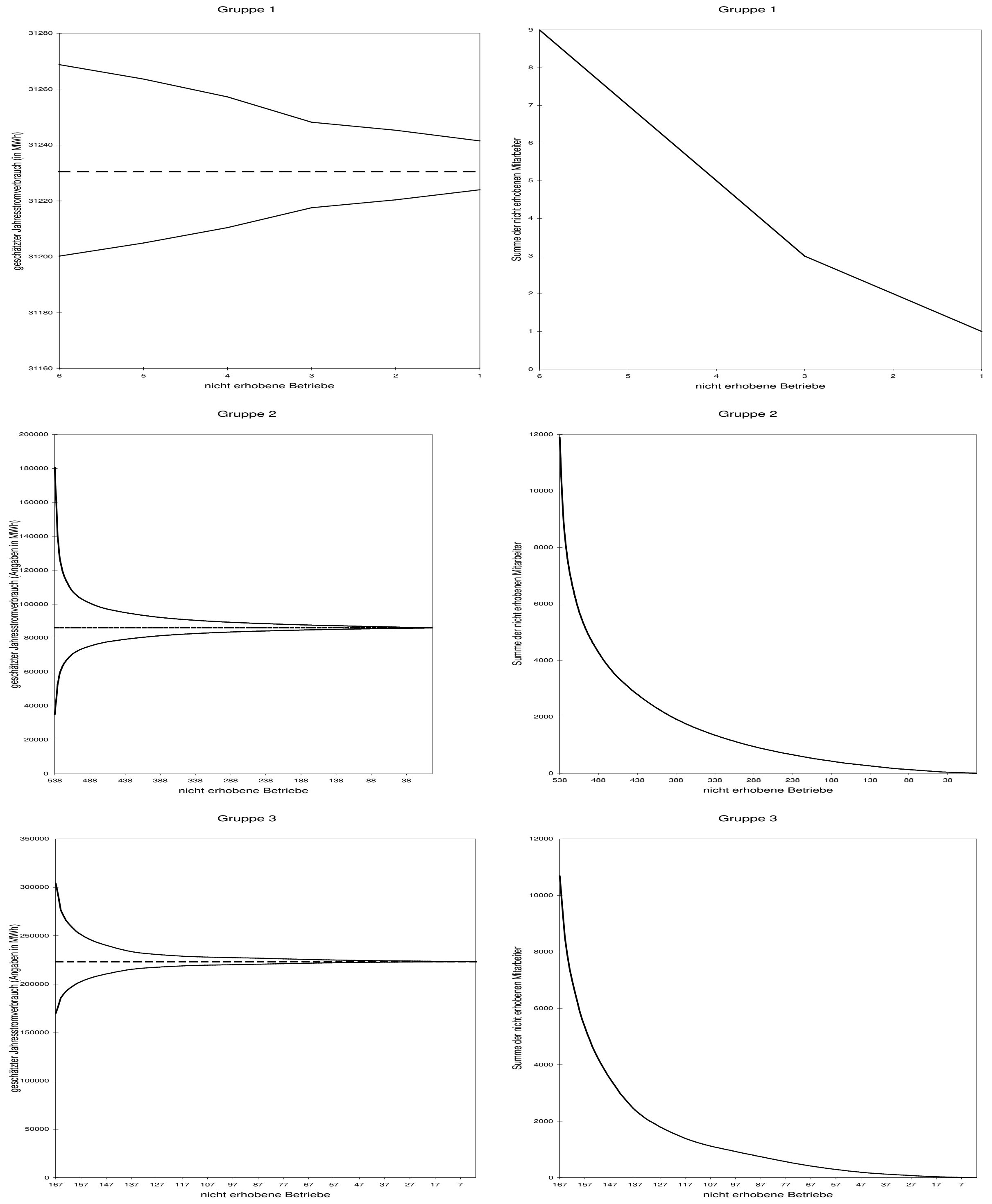



Abbildung 5.9: links: Veränderung der 95%-Bereiche für den Jahresstromverbrauch mit wachsendem Stichprobenumfang (bezüglich erhobener Betriebe) unter den getroffenen Annahmen für die Gruppen 1 bis 3; der geschätzte Erwartungswert ist strichliert gekennzeichnet; rechts: Darstellung der Mitarbeiteranzahl bei den nicht erhobenen Betrieben.

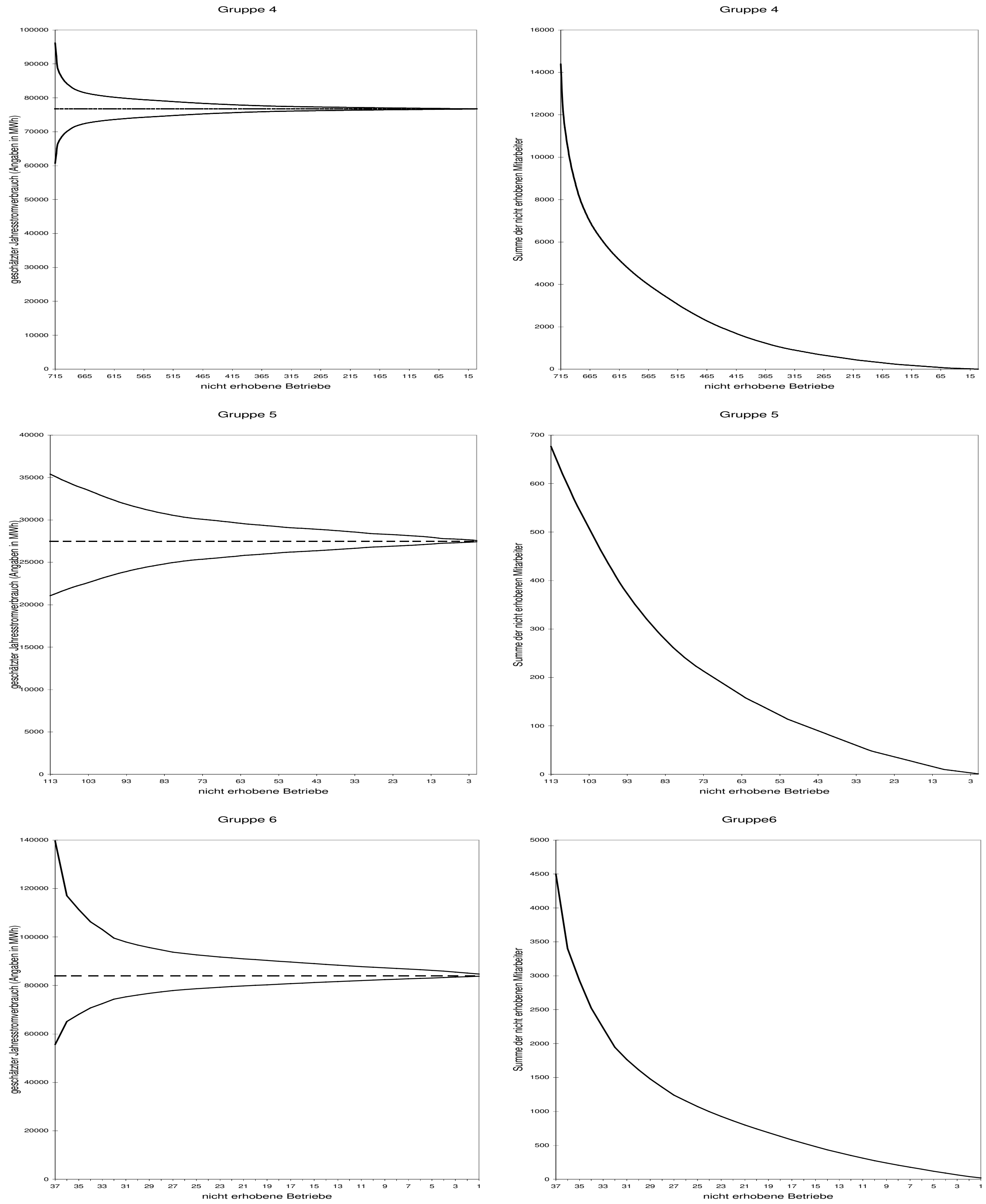



Abbildung 5.10: links: Veränderung der 95%-Bereiche für den Jahresstromverbrauch mit wachsendem Stichprobenumfang (bezüglich erhobener Betriebe) unter den getroffenen Annahmen für die Gruppen 4 bis 6; der geschätzte Erwartungswert ist strichliert gekennzeichnet; rechts: Darstellung der Mitarbeiteranzahl bei den nicht erhobenen Betrieben.

Aus den Graphiken läßt sich deutlich erkennen, daß vor allem in den Gruppen mit großen noch nicht erhobenen Betrieben durch eine Hinzunahme von wenigen Großbetrieben schon enorme Genauigkeitsgewinne erzielbar wären.

## 5.5 Die Variable Fernwärmeverbrauch

Von 44 (=14.1%) der 311 erhobenen Betriebe sind auch Angaben über den Jahresfernwärmeverbrauch bekannt. Eine Entscheidung, ob bei den restlichen 267 Betrieben (i) kein Fernwärmeanschluß besteht oder (ii) der Verbrauch nicht angegeben wurde (Missing-Value) war nicht möglich, da bei den Fragebögen eine Abfrage der Form 'Fernwärme Ja/Nein' fehlte. Daher kann man aus dem vorliegenden Datensatz keine Schlüsse über den Anteil der an das Fernwärmenetz angeschlossenen Betriebe treffen. Aus Abbildung

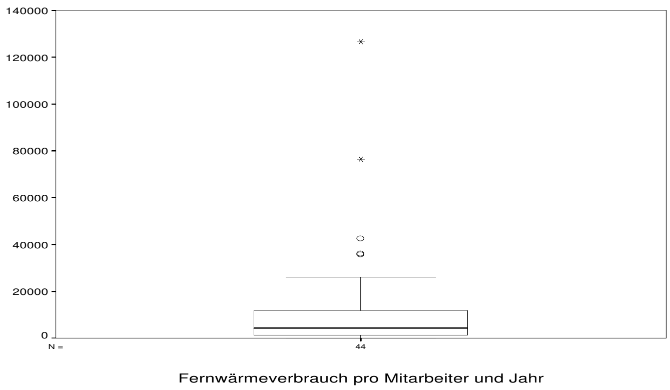



Abbildung 5.11: Boxplot der Variable Fernwärmeverbrauch pro Mitarbeiter und Jahr (FW/MA), Angaben in kWh

5.11 kann man erkennen, daß die empirische Verteilung der Variablen FW/MA extrem rechtsschief ist. Es soll nun geprüft werden, ob die Variablen FW/MA und Mitarbeiteranzahl korreliert sind. Aufgrung des geringen Stichprobenumfanges werden die Berechnungen von Extremwerten stark beeinflußt. Zur Berechnung der Korrelationskoeffizienten wird deshalb der Spearman-Rangkorrelationskoeffizient verwendet. In Tabelle 5.14 sind die Spearman-Rangkorrelationskoeffizienten für die gesamten Beobachtungen und für die nach Sektionszugehörigkeit gruppierten Beobachtungen zusammengefaßt.

Bei den Beobachtungen aus Sektion 1 ist eine Korrelation zwischen den beiden Variablen festzustellen, während in den übrigen Sektionen der Korrelationskoeffizient zwischen dem beobachteten Fernwärmeverbrauch pro Mitarbeiter und der Mitarbeiteranzahl so klein ist, daß nur auf einen schwachen Zusammenhang der beiden Variablen geschlossen werden kann.

Teilt man die Beobachtungen aus der Sektion 1 nach der Betriebsgröße in 2 Gruppen (Betrieb mit $< 20$ und$\geq 20$ Mitarbeitern), so erhält man die Ergebnisse aus Tabelle 5.15.

Aus obigen Untersuchungen kann geschlossen werden, daß der Fernwärmeverbrauch pro Mitarbeiter und Jahr bei vorliegenden Daten kaum von der Mitarbeiteranzahl abhängt.

| | Anzahl | Korr.-Koeff. |
|---|---|---|
| Gesamt | 44 | $-0.09$ |
| Sektion 1 | 10 | $-0.70$ |
| Sektion 2 | 5 | 0.40 |
| Sektion 3 | 12 | $-0.37$ |
| Sektion 5 | 2 | – |
| Sektion 6 | 15 | 0.31 |

Tabelle 5.14: Spearman-Rangkorrelationskoeffizienten für die Variablen FW/MA und Mitarbeiteranzahl

| | Anzahl | Korr.-Koeff. |
|---|---|---|
| $< 20$ MA | 6 | 0.2 |
| $\geq 20$ MA | 4 | $-0.2$ |

Tabelle 5.15: Spearman-Rangkorrelationskoeffizienten für die Variablen FW/MA und Mitarbeiteranzahl für Beobachtungen aus der Sektion 1, aufgeteilt nach der Mitarbeiteranzahl

Es wird daher versucht, die Variable FW/MA durch ein Verteilungsmodell zu beschreiben.

Aus Abbildung 5.11 ist die Rechtsschiefe der empirischen Verteilung der Variablen FW/MA erkennbar. Um zu prüfen, ob die Rechtsschiefe eine Resultat von Gruppeneffekten bezüglich der Mitarbeiteranzahl und der Sektionszugehörigkeit ist, werden die Boxplots der nach Sektion und Mitarbeitanzahl gruppierten Beobachtungen verglichen.

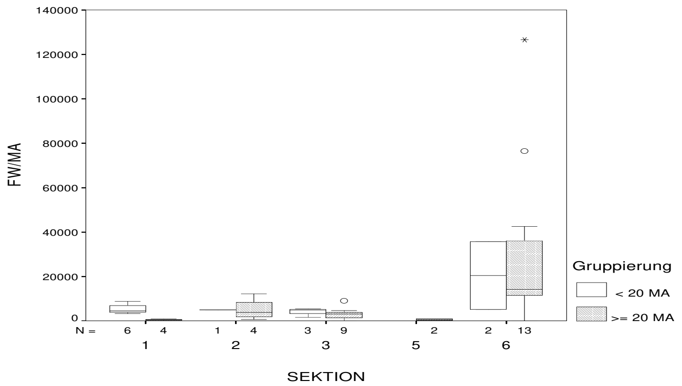



Abbildung 5.12: Boxplot der Variable FW/MA, aufgeteilt nach Sektion und Mitarbeiteranzahl (Angaben in kWh)

Die Beobachtungen aus der Sektion 6 weichen von den übrigen Beobachtungen ab, wobei in dieser Sektion kein Unterschied zwischen der Gruppe der Betriebe mit $<$ 20 und der mit $\geq$ 20 Mitarbeitern feststellbar ist. Die Sektion 6 bezeichnet Betriebe aus der Gruppe 'Tourismus und Freizeitwirtschaft', wodurch der höhere Fernwärmeverbrauch pro Mitarbeiter und Jahr erklärbar ist. Folglich sind die Beobachtungen aus der Sektion 6 getrennt zu analysieren.

Bei den restlichen Gruppen ist kein Unterschied bei einer branchenspezifischen Unterteilung feststellbar. Es ergeben sich jedoch Unterschiede zwischen der Gruppe von Betrieben mit $< 20$ bzw $\geq 20$ Mitarbeitern (sektionsübergreifend).

Die 44 Beobachtungen werden daher in 3 Gruppen unterteilt:

1. Gruppe 1: Betriebe aus den Sektionen 1, 2, 3 und 5 mit $< 20$ Mitarbeitern.

2. Gruppe 2: Betriebe aus den Sektionen 1, 2, 3 und 5 mit $\geq 20$ Mitarbeitern.

3. Gruppe 3: Betriebe aus der Sektion 6.

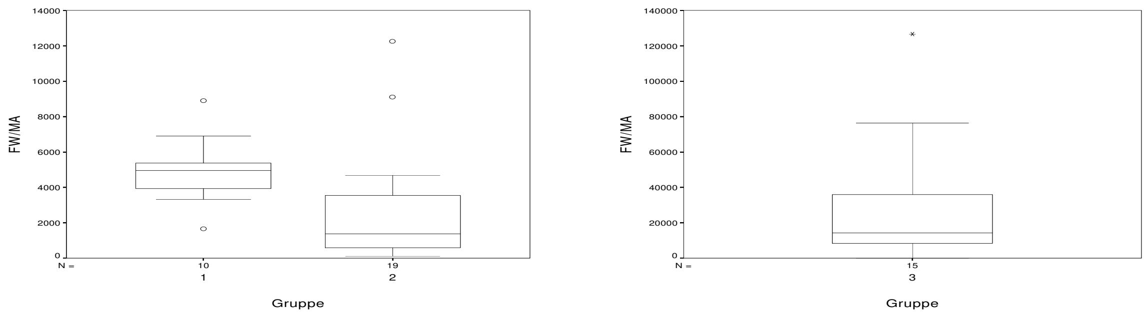



Abbildung 5.13: Boxplot der auf die 3 Gruppen aufgeteilten Werte der Variable FW/MA

Während die Daten aus der Gruppe 1 annähernd symmetrisch verteilt sind, sind die Beobachtungen aus den Gruppen 2 und 3 eindeutig rechtsschief verteilt (siehe Abbildung 5.13). Es wird daher versucht, die Variable FW/MA für die Gruppen 2 und 3 über ein Lognormalverteilungsmodell zu beschreiben.

Der Lilliefors-Anpassungtest auf Normalverteilung (Niveau $\alpha = 0.05$) ergab, daß bei dem Test der logarithmierten Daten aus Gruppe 2 und 3 die Nullhypothese (=logarithmierten Daten stammen aus einer Normalverteilung) nicht verworfen wird. Es ist daher gerechtfertigt, die Variable FW/MA aus den Gruppen 2 und 3 mit einem (zweiparametrischen) Lognormalverteilungsmodell zu beschreiben. Die Verteilungsparameter werden (analog zur Analyse der Variable STR/MA) mit der Maximum-Likelihood-Methode geschätzt, wobei bei der Schätzung des Parameters $\sigma^2$ der Skalierungsfaktor $1/(n-1)$ verwendet wird. Die Beobachtung aus Gruppe 3 mit der Firmenidentifikationsnummer 1222 wurde bei der Boxplotanalyse der logarithmierten Werte als Extremwert ausgewiesen. Der Jahresfernwärmeverbrauch von 1000 kWh weicht deutlich von den restlichen Werten ab, es handelt sich wahrscheinlich um eine Falschangabe. Diese Beobachtung wird daher von den folgenden Berechnungen ausgeschlossen.

In Tabelle 5.16 sind die Schätzungen für die Normalverteilungsparameter $\mu$ und $\sigma$ der logarithmierten Werte der Variablen FW/MA zusammengefaßt:

| Gruppe | Betriebe | n | $\hat{\mu}$ | $\hat{\sigma}$ |
|---|---|---|---|---|
| 2 | Sektion 1, 2, 3 und 5 mit $\geq$ 20 MA | 19 | 7.2282 | 1.2951 |
| 3 | Sektion 6 | 14 | 9.6772 | 1.3146 |

Tabelle 5.16: Schätzungen der Verteilungsparameter für die Gruppen 2 und 3

In Gruppe 1 werden die Momente mit der Momentenmethode geschätzt. Für die Schätzungen des Erwartungswertes (E), der Standardabweichung (SD) und des Variationskoeffizienten (SD/E) folgt:

| Gruppe | n | E | SD | SD/E |
|---|---|---|---|---|
| 1 | 10 | 4916 | 1973 | 0.40 |
| 2 | 19 | 3187 | 6647 | 2.09 |
| 3 | 14 | 37846 | 81438 | 2.15 |

Tabelle 5.17: Schätzungen für den Erwartungswert, die Standardabweichung und den Variationskoeffizienten der Variable FW/MA pro Gruppe (Angaben in kWh)

Ein approximatives 95%-Konfidenzintervall für die Erwartungswerte der Gruppen 2 und 3 wird durch die parametrische Bootstrapmethode von Angus ermittelt, wobei die Anzahl der Replikationen B mit 2000 festgelegt wird. Die zur Konstruktion benötigten Zufallszahlen wurden mit `Mathematica` generiert:

| Gruppe | E | (approx.) KIV |
|---|---|---|
| 1 | 4916 | [ 3504; 6327 ] |
| 2 | 3187 | [ 1714; 9226 ] |
| 3 | 37846 | [ 18465;152437 ] |

Tabelle 5.18: Schätzung für den Erwartungswert der Variable FW/MA; Konfidenzintervall für den Erwartungswert der Gruppe 1; approximative Konfidenzintervalle für die Erwartungswerte der Gruppen 2 und 3 ($1 - \alpha = 0.95$) (Angaben in kWh)

Für die Beobachtungen aus den Gruppen 2 und 3, die rechtsschief verteilt sind, sind in Tabelle 5.19 die Schätzungen für den Median und 95%-Konfidenzintervalle für den Median angegeben (bei Beobachtungen aus der Gruppe 1 wird angenommen, daß sie symmetrisch verteilt sind und somit Erwartungswert und Median ident sind.)

Da uns nicht bekannt ist, welche der als emissionsrelevant eingestuften Betriebe an das Fernwärmenetz angeschlossen sind, kann keine Aussage über den Gesamtfernwärmeverbrauch aller Betriebe getroffen werden.

Man kann daher nur Aussagen über den Stichprobenumfang bezüglich der Variablen FW/MA zu treffen.

| Gruppe | Median | KIV |
|---|---|---|
| 2 | 1378 | [ 738; 2572 ] |
| 3 | 15950 | [ 7466;34072 ] |

Tabelle 5.19: Schätzung für den Median der Variable FW/MA für die Gruppen 2 und 3; Konfidenzintervall für den Median ($1 - \alpha = 0.95$) (Angaben in kWh)

### 5.5.1 Bestimmung des benötigten Stichprobenumfanges

Aufbauend auf den Beobachtungen, den Verteilungsannahmen und den dadurch gewonnenen Schätzungen und Konfidenzintervallen wird gruppenweise untersucht, inwieweit sich das Konfidenzintervall für den Erwartungswert unter der Annahme, daß die Schätzung der Verteilungsparameter für alle betrachteten Stichprobenumfänge ident bleiben, mit wachsendem Stichprobenumfang ändert. Die Darstellung der resultierenden Konfidenzintervalle erfolgt in analoger Weise zu den obigen Analysen.

Zur Berechnung der approximativen Konfidenzintervalle für die Erwartungswerte der Gruppen 2 und 3 wird wie zuvor die Methode von Cox verwendet.

Unter den in Abschnitt 5.4.2 getroffenen Annahmen gilt daher:

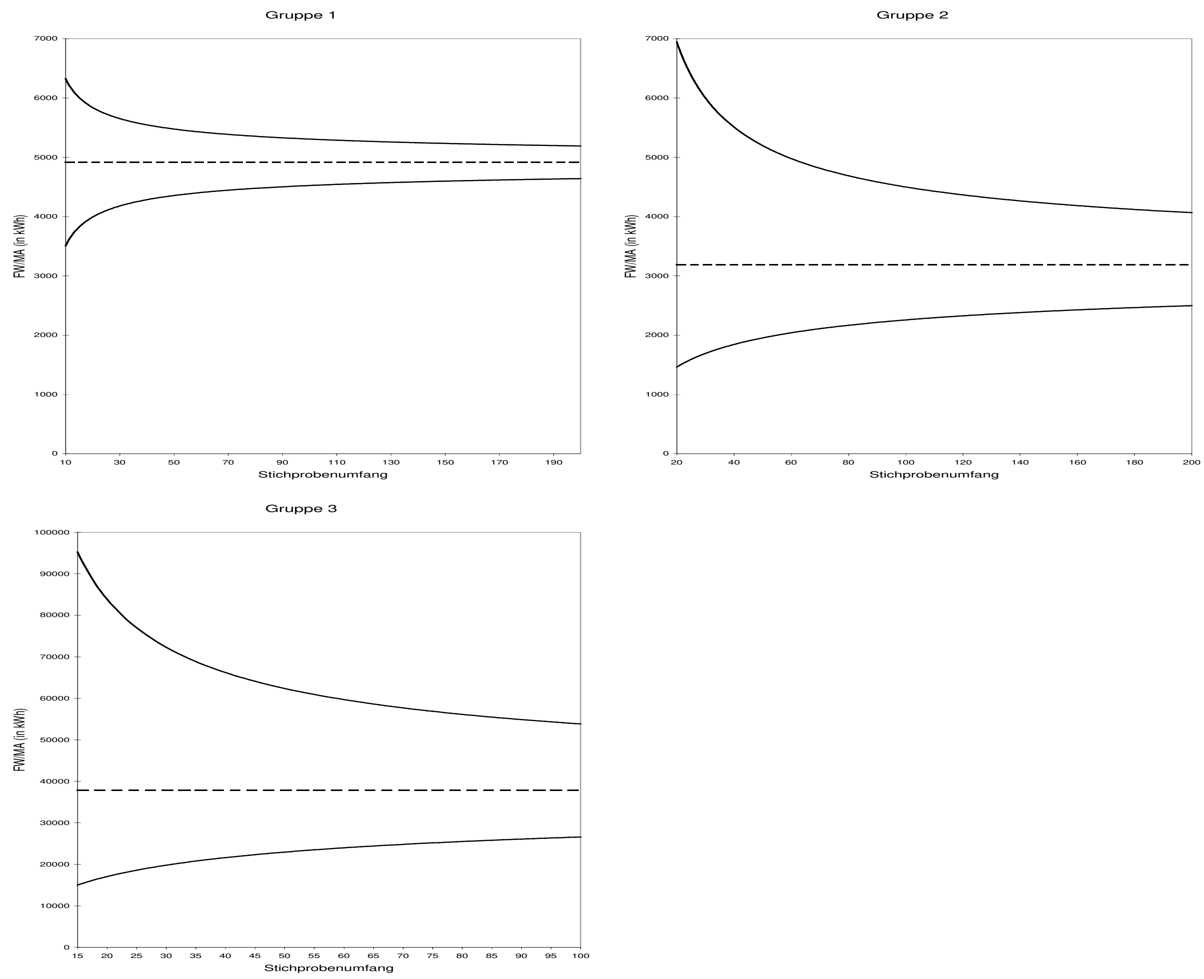



Abbildung 5.14: Veränderung der Konfidenzintervalle für den Erwartungswert der Variablen FW/MA (gruppenweise) mit wachsendem Stichprobenumfang (bezüglich erhobener Betriebe) unter den getroffenen Annahmen; der geschätzte Erwartungswert ist strichliniert gekennzeichnet.

Anmerkung zu Abbildung 5.14: Der Endpunkt der $x$-Achse ist nicht wie bei der Analyse der Variable STR/MA der Vollerhebungszustand, da bei der Variablen FW/MA die Anzahl der Betriebe mit Fernwärmeanschluß nicht bekannt ist. Es wurde daher ein fiktiver Wert gewählt.

## 5.6 Betrachtung einiger spezieller Fachgruppen

Neben der Globalanalyse werden noch jene Branchen getrennt betrachtet, in denen zumindest 4 Betriebe erhoben wurden (bzgl. Stromverbrauchsangaben). Für die Fachgruppen mit $\geq 10$ Beobachtungen ist eine Analyse sinnvoll, bei geringeren Stichprobenumfängen werden nur die erhobenen Werte angegeben und diese gegebenfalls graphisch dargestellt.

| Bezeichnung | FG | UIS (MA bekannt) | erhoben | Stromverbr. bekannt |
|---|---|---|---|---|
| Baugewerbe | 101 | 40 (40) | 7 | 6 (15%) |
| Tischler | 109 | 94 (93) | 8 | 4 ( 4%) |
| Schlosser u. Schmied | 114B | 74 (74) | 12 | 8 (11%) |
| Mechatroniker | 120 | 36 (36) | 7 | 4 (11%) |
| KFZ-Mechaniker | 121 | 66 (66) | 18 | 14 (21%) |
| Gärtner | 140 | 35 (26) | 35 | 4 (11%) |
| Chemischreiniger | 145 | 39 (38) | 39 | 32 (82%) |
| Maschinen- und Stahlbauindustrie | 216 | 53 (51) | 17 | 10 (18%) |
| Textilhandel | 308 | 44 (44) | 7 | 6 (14%) |
| Fahrzeughandel | 318 | 22 (22) | 8 | 7 (32%) |
| Güterbeförderung | 507 | 17 (17) | 6 | 5 (29%) |
| Gastronomie | 601 | 33 (33) | 7 | 7 (21%) |
| Hotellerie | 602 | 21 (20) | 16 | 11 (52%) |
| Krankenanstalten | 603 | 21 (21) | 19 | 9 (43%) |

Tabelle 5.20: Fachgruppen, die getrennt betrachtet werden

Für ein dem Grazer UIS ähnlichen Projekt der Salzburger Landesregierung, dem Salzburger Energie- und Emissionskataster (SEMIKAT), wurde für die FG-en 109, 114B, 121, 140 und 145 ebenfalls eine Auswertung von Energieverbrauchsdaten durchgeführt (siehe KAISER (1996), HITZL (1996)). Die Ergebnisse der Salzburger und Grazer Auswertungen werden in diesem Abschnitt miteinander verglichen.

In Tabelle 5.21 ist die Summe der Mitarbeiter pro FG angegeben. Neben der Summe der Mitarbeiter ist noch die Anzahl der Mitarbeiter pro Betrieb von Interesse. In Tabelle 5.22 ist die Anzahl der Betriebe pro Mitarbeiterklasse angeführt. Die Angabe ohne Klammer bezieht sich auf die im UIS eingetragenen Betriebe, in den Klammern steht der derzeitige Erhebungsstand bzw. die Anzahl der Betriebe, von denen der Stromverbrauch bekannt ist (erhoben; Stromverbrauch bekannt).

| FG | Summe MA im UIS | erhoben | Stromverbr. bekannt |
|---|---|---|---|
| 101 | 4170 | 454 | 451 (11%) |
| 109 | 610 | 69 | 40 ( 7%) |
| 114B | 1644 | 400 | 376 (23%) |
| 120 | 1656 | 1278 | 1247 (75%) |
| 121 | 969 | 445 | 347 (36%) |
| 140 | 271 | 271 | 7 ( 3%) |
| 145 | 497 | 497 | 488 (98%) |
| 216 | 4047 | 2487 | 1166 (29%) |
| 308 | 1136 | 131 | 125 (11%) |
| 318 | 923 | 544 | 430 (47%) |
| 507 | 907 | 392 | 382 (42%) |
| 601 | 1602 | 523 | 523 (33%) |
| 602 | 488 | 259 | 235 (48%) |
| 603 | 10241 | 9810 | 7144 (70%) |

Tabelle 5.21: Summe der Mitarbeiter pro FG

| FG | 0-5 | 6-10 | 11-20 | 21-50 | 51-100 | 101-500 | 501-1000 | >1000 |
|---|---|---|---|---|---|---|---|---|
| 101 | 5 ( 1; 0) | – | 2 (1;1) | 15 (0;0) | 10 (4;4) | 5 (1;1) | 3 (0;0) | – |
| 109 | 54 ( 4; 1) | 25 (1;1) | 11 (3;2) | 3 (0;0) | – | – | – | – |
| 114B | 31 ( 3; 2) | 20 (4;1) | 13 (1;1) | 7 (3;3) | – | 2 (1;1) | 1 (0;0) | – |
| 120 | 18 ( 3; 1) | 4 (1;1) | 6 (0;0) | 6 (2;1) | – | 1 (0;0) | – | 1 (1;1) |
| 121 | 34 ( 3; 2) | 9 (1;1) | 10 (8;6) | 9 (4;4) | 3 (2;1) | 1 (0;0) | – | – |
| 140 | 17 (17; 4) | 3 (3;0) | 5 (5;0) | – | – | 1 (1;0) | – | – |
| 145 | 33 (33;27) | 3 (3;3) | – | – | – | 2 (2;2) | – | – |
| 216 | 8 ( 2; 0) | 6 (2;1) | 14 (3;2) | 8 (4;3) | 3 (0;0) | 10 (4;3) | 1 (1;1) | 1 (1;1) |
| 308 | 4 ( 0; 0) | 8 (2;1) | 9 (4;4) | 20 (1;1) | 1 (0;0) | 2 (0;0) | – | – |
| 318 | 2 ( 0; 0) | 2 (1;1) | 5 (3;3) | 8 (1;1) | 3 (1;1) | 2 (2;1) | – | – |
| 507 | 2 ( 0; 0) | 1 (1;0) | 3 (0;0) | 5 (3;3) | 3 (1;1) | 3 (1;1) | – | – |
| 601 | 4 ( 0; 0) | 2 (0;0) | 4 (1;1) | 11 (3;3) | 7 (0;0) | 5 (3;3) | – | – |
| 602 | 6 ( 6; 4) | 4 (4;2) | 3 (0;0) | 3 (3;3) | 3 (2;2) | 1 (0;0) | – | – |
| 603 | – | – | – | 6 (5;2) | 4 (4;3) | 9 (8;3) | – | 2 (2;1) |

Tabelle 5.22: Betriebe pro FG und Mitarbeiterklasse

### 5.6.1 Baugewerbe (FG 101)

Von 40 im UIS eingetragenen Betrieben ist von 6 der Jahresstromverbrauch bekannt.

| #(MA) | Strom pro Jahr (in MWh) | STR/MA (in kWh) |
|---|---|---|
| 20 | 0.1 | 5 |
| 59 | 7.0 | 119 |
| 74 | 150.1 | 2029 |
| 77 | 28.1 | 365 |
| 87 | 28.8 | 331 |
| 134 | 7.7 | 58 |

Bei der ersten Beobachtung (20 MA) handelt es sich um eine offensichtliche Falschangabe (diese Beobachtung wurde auch bei der Globalanalyse eliminiert). Bei den restlichen 5 Beobachtungen ist kein Zusammenhang zwischen STR/MA und der Mitarbeiteranzahl feststellbar.

| Beob. | Mittelwert | Median | Min. | Max. | IQR |
|---|---|---|---|---|---|
| 5 | 580 | 331 | 58 | 2029 | 1108 |

Bei der Globalanalyse der Variablen STR/MA aller Betriebe aus der Sektion 1 wurde der Erwartungswert mit 6977 kWh und der Median mit 1978 kWh geschätzt. Die 5 Betriebe des Baugewerbes liegen deutlich unter diesen Werten. Der geringe Stromverbrauch pro Mitarbeiter ist dadurch erklärt, daß bei obigen Angaben der Baustrom (Vorortverbrauch auf Baustellen) nicht inkludiert ist.

### 5.6.2 Tischler (FG 109)

Bei 4 der 94 Betriebe aus der FG 109 ist der Jahresstromverbrauch bekannt.

| #(MA) | Strom pro Jahr (in MWh) | STR/MA (in kWh) |
|---|---|---|
| 5 | 32.4 | 6482 |
| 8 | 5.0 | 625 |
| 11 | 16.9 | 1540 |
| 16 | 36.0 | 2250 |

Diese Branche hat ebenfalls einen im Vergleich zur gesamten Sektion 1 geringen Jahresstromverbrauch pro Mitarbeiter. Im Vergleich zu den Ergebnissen der SEMIKAT liegen die registrierten Verbräuche unter den dort ermittelten Werten, jedoch ist die Diskrepanz gering.

| | Beob. | Mittelwert | Median | Min. | Max. |
|---|---|---|---|---|---|
| UIS | 4 | 2724 | 1895 | 625 | 6482 |
| SEMIKAT | 12 | 3800 | 3300 | 1800 | 9800 |

### 5.6.3 Schlosser und Schmiede (FG 114B)

Von 8 der 74 emissionsrelevanten Betriebe aus der FG 114B sind Angaben über den Jahresstromverbrauch bekannt.

| #(MA) | Strom pro Jahr (in MWh) | STR/MA (in kWh) |
|---:|---:|---:|
| 2 | 3.8 | 1878 |
| 2 | 6.0 | 3000 |
| 6 | 47.9 | 7986 |
| 18 | 32.6 | 1810 |
| 25 | 15.0 | 600 |
| 30 | 47.0 | 1567 |
| 43 | 289.7 | 6738 |
| 250 | 2.5 | 10 |

Die letzte Beobachtung (250 MA) weicht deutlich von den restlichen ab, es handelt sich hierbei wahrscheinlich um eine Falschangabe. Diese Beobachtung wurde auch bei der Globalanalyse als Extremwert identifiziert und von den Berechnungen ausgeschlossen. Bei den restlichen 7 Beobachtungen ist wiederum keine Korrelation zwischen STR/MA und der Mitarbeiteranzahl feststellbar (Pearson'scher Korrelationskoeffizient = 0.07, Spearman-Rangkorrelationskoeffizient = −0.23; siehe auch Abbildung 5.15). Als Schätzungen ergeben sich:

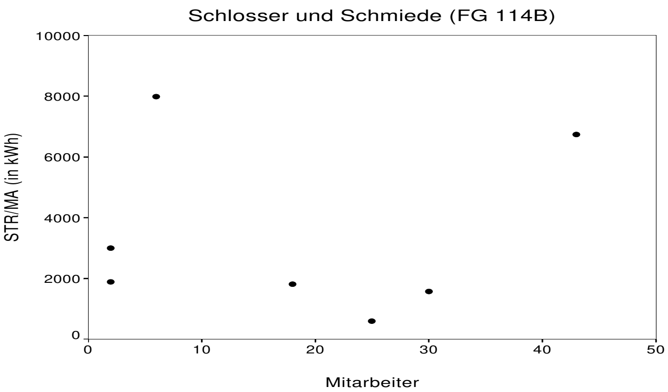



Abbildung 5.15: Scatterplot: STR/MA-Mitarbeiter; FG 114B

| | Beob. | Mittelwert | Median | Min. | Max. | IQR | SD | Schiefe |
|---|---:|---:|---:|---:|---:|---:|---:|---:|
| UIS | 7 | 3368 | 1878 | 600 | 7986 | 5171 | 2840 | 1.05 |
| SEMIKAT | 10 | 1961 | 1846 | 1067 | 3625 | – | – | – |

Bei der SEMIKAT-Auswertung sind der Mittelwert und der Median nahezu ident (= symmetrische Verteilung), während die Grazer Daten wiederum auf eine rechtsschiefe Verteilung hinweisen (Mittelwert > Median, rechte Extremwerte treten auf).

### 5.6.4 Mechatroniker (FG 120)

Aus der FG 120 wurden 7 der 36 Betriebe erhoben, wobei von 4 der Jahresstromverbrauch bekannt ist.

| #(MA) | Strom pro Jahr (in MWh) | STR/MA (in kWh) |
|---|---|---|
| 1 | 0.5 | 500 |
| 9 | 11.0 | 1222 |
| 23 | 1.7 | 72 |
| 1214 | 5872.8 | 4838 |

| Beob. | Mittelwert | Median | Min. | Max. |
|---|---|---|---|---|
| 4 | 1658 | 861 | 72 | 4838 |

### 5.6.5 KFZ-Mechaniker (FG 121)

Bei 14 der 66 Betriebe aus FG 121 wurden Stromverbrauchsdaten erhoben. Der Stromverbrauch pro Jahr ist wiederum in MWh, der Stromverbrauch pro Mitarbeiter und Jahr in kWh angegeben.

| #(MA) | Strom pro Jahr | STR/MA | #(MA) | Strom pro Jahr | STR/MA |
|---|---|---|---|---|---|
| 2 | 2.4 | 1200 | 18 | 50.7 | 2817 |
| 3 | 14.9 | 4963 | 20 | 33.0 | 1650 |
| 9 | 8.0 | 889 | 25 | 208.7 | 8348 |
| 14 | 40.7 | 2906 | 30 | 35.8 | 1193 |
| 14 | 46.8 | 3340 | 33 | 207.8 | 6296 |
| 16 | 5.3 | 331 | 40 | 45.7 | 1142 |
| 17 | 4.3 | 256 | 98 | 278.3 | 2840 |

Es soll nun geprüft werden, ob die Variable STR/MA von der Mitarbeiteranzahl abhängig ist. Der Korrelationskoeffizient zwischen den beiden Größen ist klein (Pearson = 0.08,

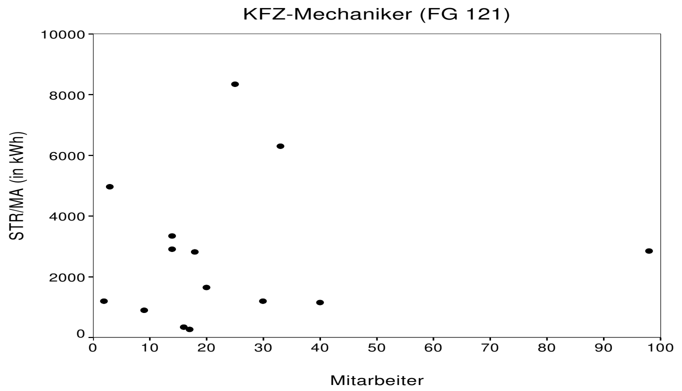



Abbildung 5.16: Scatterplot: STR/MA-Mitarbeiter; FG 121

Spearman = 0.09), aus dem Scatterplot kann ebenfalls keine Abhängigkeitsstruktur erkannt werden. Man kann daher auf Unkorreliertheit der beiden Variablen schließen. In

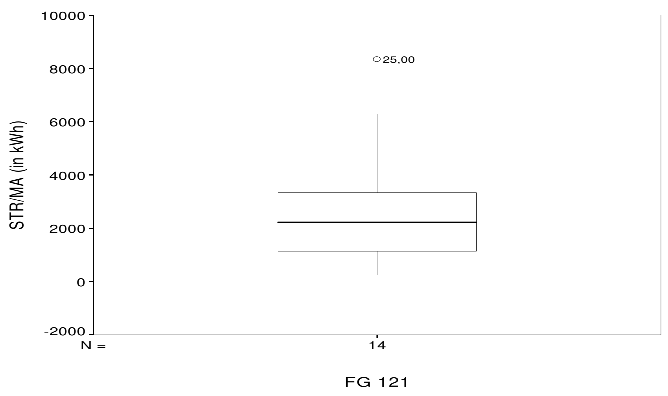



Abbildung 5.17: Boxplot der Variablen STR/MA; FG 121

Abbildung 5.17 ist der Boxplot der Variablen STR/MA dargestellt. Es ergeben sich folgende Schätzungen:

| | Beob. | Mittelwert | Median | Min. | Max. | IQR | SD | SD/E | Schiefe |
|---|---|---|---|---|---|---|---|---|---|
| UIS | 14 | 2727 | 2233 | 256 | 8348 | 2668 | 2375 | 0.87 | 0.87 |
| SEMIKAT | 13 | 3300 | 2400 | 1400 | 6500 | – | – | – | – |

Für das 95%-Konfidenzintervall für den Erwartungswert ($\overline{X} \pm \frac{SD}{\sqrt{n}} t_{n-1;0.975}$, wobei $t_{n-1;\alpha}$ das $\alpha$- Quantil der Student-t-Verteilung mit $n-1$ Freiheitsgraden bezeichnet) folgt: $[1356; 4098]$. Der Mittelwert der SEMIKAT-Beobachtungen liegt eindeutig in diesem Intervall. Vergleicht man die Schätzungen für die FG 121 mit jenen der Sektion 1, so fällt auf, daß die Medianschätzungen nahezu ident sind (2233 und 1978), während die Schätzungen für den Erwartungswert stark voneinander abweichen (2727 und 6977), d.h. die empirische Verteilung der Variablen STR/MA ist für die FG 121 nicht so rechtsschief wie für die Sektion 1.

### 5.6.6 Gärtner (FG 140)

Von 4 der 35 Gärtnereien ist der Stromverbrauch bekannt.

| #(MA) | Strom pro Jahr (in MWh) | STR/MA (in kWh) |
|---|---|---|
| 1 | 44.0 | 44000 |
| 1 | 4.8 | 4800 |
| 2 | 0.5 | 232 |
| 3 | 40.3 | 13446 |

Diese Gruppe ist bezüglich der Variablen STR/MA besonders inhomogen.

| | Beob. | Mittelwert | Median | Min. | Max. |
|---|---|---|---|---|---|
| UIS | 4 | 15620 | 9123 | 232 | 44000 |
| SEMIKAT | 5 | 4090 | 4230 | 2350 | 5560 |

Die SEMIKAT-Auswertungen weichen gravierend von unseren Ergebnissen ab. Es muß allerdings berücksichtigt werden, daß der Stichprobenumfang bei beiden Untersuchungen klein ist und somit auch die Ergebnisse von Extremwerten stark beeinflußt werden. Auffallend ist die Homogenität der Salzburger Beobachtungen, während die Spannweite der Grazer Daten sehr groß ist.

### 5.6.7 Chemischreiniger (FG 145)

Die 30 Beobachtungen aus der FG 145 wurden bereits bei der Globalanalyse in sorgfältiger Weise untersucht (Gruppe 1). Die detaillierten Ergebnisse werden an dieser Stelle nicht angeführt. Der Vergleich mit den SEMIKAT-Ergebnissen zeigt eine gute Übereinstimmung.

| | Beob. | Mittelwert | Median | Min. | Max. | IQR |
|---|---|---|---|---|---|---|
| UIS | 30 | 11865 | 11085 | 2200 | 34000 | 5670 |
| SEMIKAT | 8 | 11000 | – | – | – | – |

Der Mittelwert der SEMIKAT-Auswertung ist im 95%-Konfidenzintervall für den Erwartungswert der Variablen STR/MA enthalten.

### 5.6.8 Maschinen- und Stahlbauindustrie (FG 216)

Von 10 der 53 Betriebe aus der FG 216 ist der Stromverbrauch bekannt. Der Jahresstromverbrauch ist wiederum in MWh angeben, der Stromverbrauch pro Mitarbeiter und Jahr in kWh.

| #(MA) | Strom pro Jahr | STR/MA | #(MA) | Strom pro Jahr | STR/MA |
|---|---|---|---|---|---|
| 7 | 44.0 | 6287 | 50 | 599.7 | 11995 |
| 15 | 64.0 | 4267 | 116 | 163.4 | 1408 |
| 20 | 90.0 | 4500 | 119 | 360.3 | 3028 |
| 23 | 109.3 | 4752 | 148 | 272.6 | 1842 |
| 48 | 274.7 | 5723 | 620 | 5556.0 | 8961 |

Um zu prüfen, ob die Variablen STR/MA und Mitarbeiteranzahl korreliert sind, werden der Scatterplot (Abbildung 5.18) und der Korrelationskoeffizient angegeben. Der Korrelationskoeffizient ist klein (Pearson = 0.26, Spearman = −0.16). Betrachtet man jedoch Abbildung 5.18, so fallen zwei Extremwerte auf: (#(MA) = 50, STR/MA = 11994 kWh) und (#(MA) = 180, STR/MA = 8961 kWh). Diese Beobachtungen haben bei der Bestimmung der Korrelationskoeffizienten aufgrund des geringen Stichprobenumfanges einen großen Einfluß. Schließt man diese beiden Beobachtungen bei der Berechnung aus, so gilt: Pearson = −0.86, Spearman = −0.70. Die restlichen 8 Beobachtungen sind stark negativ korreliert (= STR/MA wird mit wachsender Mitarbeiteranzahl geringer). Wird diese Vermutung bei einem größeren Stichprobenumfang bestätigt, so ist ein Regressionsmodell eine geeignete Methode, um den Zusammenhang zwischen beiden Variablen zu beschreiben.

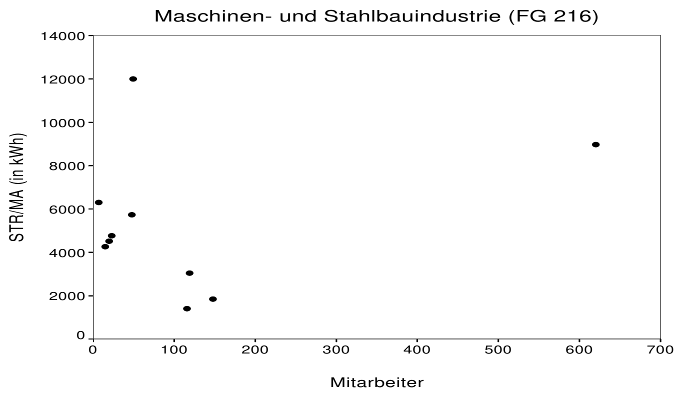



Abbildung 5.18: Scatterplot: STR/MA-Mitarbeiter; FG 216

Abbildung 5.19 zeigt den Boxplot der Variablen STR/MA

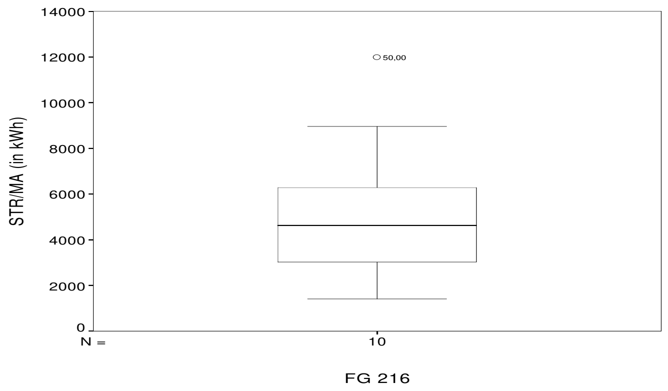



Abbildung 5.19: Boxplot der Variablen STR/MA; FG 216

Es ergeben sich folgende Schätzungen:

| Beob. | Mittelwert | Median | Min. | Max. | IQR | SD | SD/E | Schiefe |
|---|---|---|---|---|---|---|---|---|
| 10 | 5276 | 4626 | 1408 | 11994 | 4224 | 3225 | 0.61 | 1.00 |

Die Betriebe aus der FG 216 haben einen im Vergleich zur Sektion 2 niedrigen Stromverbrauch pro Mitarbeiter und Jahr (11037). Das 95%- Konfidenzintervall für den Erwartungswert der Variablen STR/MA ( $\overline{X} \pm \frac{SD}{\sqrt{n}} t_{n-1;0.975}$) lautet: $[2969; 7583]$.

### 5.6.9 Textilhandel (FG 308)

6 der 44 Textilhandelsbetriebe wurden bezüglich des Stromverbrauchs erhoben.

| #(MA) | Strom pro Jahr (in MWh) | STR/MA (in kWh) |
|---|---|---|
| 7 | 22.0 | 3143 |
| 15 | 101.2 | 6745 |
| 18 | 13.9 | 775 |
| 18 | 249.8 | 13878 |
| 20 | 95.5 | 4775 |
| 47 | 162.3 | 3453 |

| Beob. | Mittelwert | Median | Min. | Max. | IQR |
|---|---|---|---|---|---|
| 6 | 5461 | 4114 | 775 | 13878 | 5997 |

Die Durchschnittswerte der FG 308 weichen kaum von jenen der Globalanalyse der Sektion 3 (Mittelwert=4745) ab.

Anmerkung: Der Betrieb mit 18 Mitarbeitern und einem Jahresstromverbrauch von 13.9 MWh ist ein Großhandelsbetrieb, während alle übrigen Einzelhandelsbetriebe sind. Es scheint einen Unterschied zwischen Groß- und Einzelhandelsbetrieben zu geben, der jedoch aufgrund des geringen Stichprobenumfanges nicht geprüft werden kann. Die Schätzungen ohne den Großhandelsbetrieb lauten:

| Beob. | Mittelwert | Median | Min. | Max. | IQR |
|---|---|---|---|---|---|
| 5 | 6399 | 4775 | 3143 | 13878 | 7013 |

### 5.6.10 Fahrzeughandel (FG 318)

7 der 22 Fahrzeughändler wurden erhoben (Strom pro Jahr in MWh, STR/MA in kWh).

| #(MA) | Strom pro Jahr | STR/MA | #(MA) | Strom pro Jahr | STR/MA |
|---|---|---|---|---|---|
| 10 | 20.0 | 2000 | 43 | 101.8 | 2367 |
| 11 | 50.0 | 4545 | 62 | 194.4 | 3136 |
| 19 | 44.6 | 2347 | 256 | 733.1 | 2767 |
| 20 | 34.0 | 1700 | | | |

| Beob. | Mittelwert | Median | Min. | Max. | IQR |
|---|---|---|---|---|---|
| 7 | 2695 | 2367 | 1700 | 4545 | 1135 |

Die Beobachtungen der FG 318 sind annähernd symmetrisch verteilt, die Gruppe ist im Vergleich zu den übrigen Fachgruppen sehr homogen. Um einen möglichen Zusammenhang zwischen den Variablen STR/MA und MA festzustellen, werden wiederum der Scatterplot und der Korrelationskoeffizient angegeben. Es gilt: Pearson = 0.04, Spearman = 0.29. Ähnlich zur FG 216 gibt es eine Beobachtung (MA=256), die aufgrund der hohen Mitarbeiteranzahl einen großen Einfluß bei der Berechnung des (Pearson-) Korrelationskoeffizienten hat. Schließt man diesen Betrieb von den Berechnungen aus, so folgt: Pearson = 0.02 und Spearman = 0.20. Aufgrund dieser Ergebnisse kann auf Unkorreliertheit der beiden Variablen geschlossen werden.

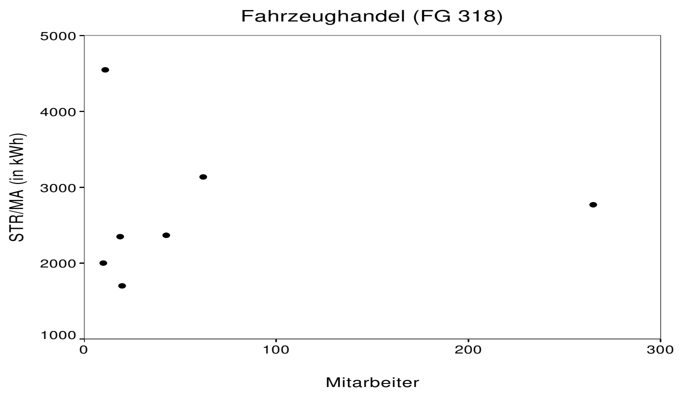



Abbildung 5.20: Scatterplot: STR/MA-Mitarbeiter; FG 318

Auffallend ist die Ähnlichkeit der Ergebnisse für FG 318 (Fahrzeughandel) und FG 121 (KFZ-Mechaniker). Der Grund dafür ist, daß eine Zuordnung der entsprechenden Betriebe zur FG in diesem Fall nicht eindeutig ist, da 'Mischformen' (=Handel und Reparatur) des öfteren auftreten.

### 5.6.11 Güterbeförderungsgewerbe (FG 507)

5 der 17 Betriebe der FG 507 wurden erhoben.

| #(MA) | Strom pro Jahr (in MWh) | STR/MA (in kWh) |
|---:|---:|---:|
| 31 | 2.3 | 75 |
| 32 | 47.1 | 1472 |
| 33 | 303.4 | 9196 |
| 85 | 278.0 | 3271 |
| 201 | 10.7 | 53 |

Unter den 5 Beobachtungen sind 2 mit extrem niedrigem und ein Betrieb mit relativ hohem STR/MA. Diese Gruppe ist - wie auch FG 140 - sehr inhomogen.

| Beob. | Mittelwert | Median | Min. | Max. | IQR |
|---:|---:|---:|---:|---:|---:|
| 5 | 2813 | 1472 | 53 | 9196 | 6170 |

### 5.6.12 Gastronomie (FG 601)

7 der 33 Betriebe aus der FG 601 wurden erhoben (Stromverbrauch pro Jahr in MWh, STR/MA in kWh).

| #(MA) | Strom pro Jahr | STR/MA | #(MA) | Strom pro Jahr | STR/MA |
|---:|---:|---:|---:|---:|---:|
| 15 | 118.8 | 7917 | 128 | 633.0 | 4945 |
| 21 | 2.0 | 95 | 140 | 675.0 | 4821 |
| 26 | 165.0 | 6346 | 148 | 6147.0 | 41534 |
| 45 | 339.4 | 7542 | | | |

Die zweite Beobachtung (MA = 21) weicht deutlich von den restlichen Werten ab. Diese Beobachtung wurde auch bei der Globalanalyse als Extremwert identifiziert (Firmen-ID 1222) und von den Berechnungen ausgeschlossen. Für die restlichen 6 Betriebe folgt:

| Beob. | Mittelwert | Median | Min. | Max. | IQR |
|---|---|---|---|---|---|
| 6 | 12184 | 6944 | 4821 | 41534 | 11407 |

Ohne den hohen Extremwert (MA = 148) ergibt sich:

| Beob. | Mittelwert | Median | Min. | Max. | IQR |
|---|---|---|---|---|---|
| 5 | 6314 | 6346 | 4821 | 7917 | 2846 |

### 5.6.13 Hotellerie (FG 602)

11 der 21 Hotelleriebetriebe wurden erhoben. Der Stromverbrauch pro Jahr ist wiederum in MWh, der Stromverbrauch pro Mitarbeiter und Jahr in kWh angegeben. In der Spalte FW sind jene Betriebe gekennzeichnet, von welchen bekannt ist, daß ein Fernwärmeanschluß besteht.

| #(MA) | Strom pro Jahr | STR/MA | FW | #(MA) | Strom pro Jahr | STR/MA | FW |
|---|---|---|---|---|---|---|---|
| 1 | 70.0 | 70000 | | 28 | 211.8 | 7565 | J |
| 2 | 76.9 | 38435 | | 35 | 60.0 | 1714 | |
| 3 | 16.3 | 5543 | J | 35 | 80.0 | 2285 | |
| 5 | 90.8 | 18158 | | 52 | 1085.2 | 20869 | J |
| 6 | 571.0 | 95166 | J | 60 | 517.7 | 8628 | J |
| 8 | 123.0 | 15375 | | | | | |

Bei der Globalanalyse wurden die Beobachtungen aus der Sektion 6 in 2 Gruppen aufgeteilt ($<20$ und $\geq 20$ MA). Dieser Ansatz wird durch obige Beobachtungen bestätigt: Betriebe mit $<20$ MA scheinen einen höheren STR/MA zu haben als jene mit $\geq 20$. Innerhalb der beiden Gruppen kann keine Abhängigkeit zwischen den Variablen STR/MA und MA festgestellt werden.

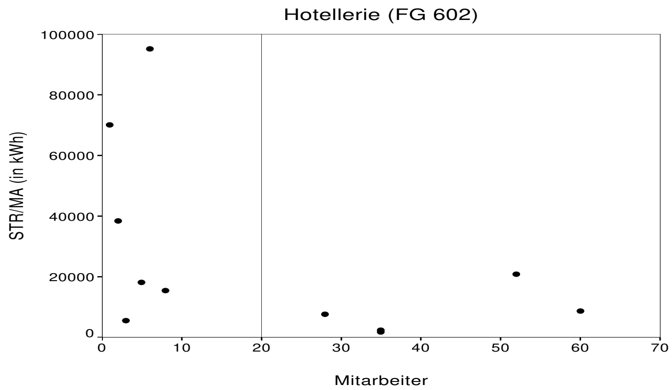



Abbildung 5.21: Scatterplot: STR/MA-Mitarbeiter; FG 602

Bei 5 der 11 Betriebe ist bekannt, daß ein Fernwärmeanschluß besteht. Ob bei den restlichen Betrieben kein Anschluß besteht oder der Eintrag fehlt, kann aus in Abschnitt 5.5 geschilderten Problemen nicht entschieden werden. Es besteht kein offensichtlicher Zusammenhang zwischen der Größe der Variablen STR/MA und einem Fernwärmeanschluß.

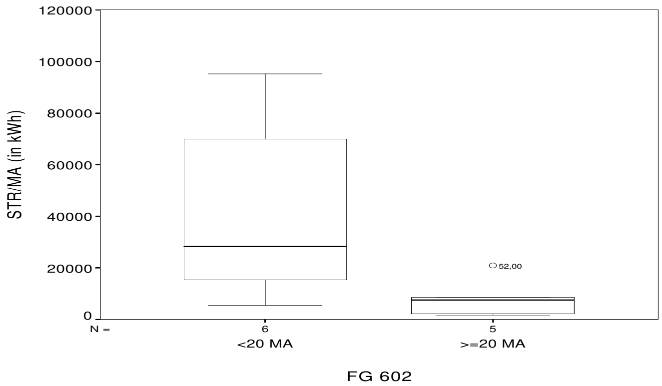



Abbildung 5.22: Boxplot der Variablen STR/MA; FG 602

Es ergeben sich folgende Schätzungen:

| | Beob. | Mittelwert | Median | Min. | Max. | IQR |
|---|---|---|---|---|---|---|
| $< 20$ MA | 6 | 40446 | 28296 | 5534 | 95167 | 63375 |
| $\geq 20$ MA | 5 | 8212 | 7565 | 1714 | 20869 | 12748 |

### 5.6.14 Krankenanstalten (FG 603)

Bei 9 der 21 Betriebe aus der FG 603 ist der Jahresstromverbrauch bekannt (Stromverbrauch pro Jahr in MWh, STR/MA in kWh)

| #(MA) | Strom pro Jahr | STR/MA | #(MA) | Strom pro Jahr | STR/MA |
|---|---|---|---|---|---|
| 41 | 172.7 | 4212 | 180 | 71.7 | 398 |
| 49 | 237.0 | 4837 | 290 | 1246.0 | 4297 |
| 51 | 24.5 | 480 | 301 | 1620.0 | 5382 |
| 66 | 429.6 | 6509 | 6080 | 24193.0 | 3979 |
| 86 | 252.5 | 2936 | | | |

Die Variablen STR/MA und MA sind nicht korreliert (Pearson = 0.06, Spearman = -0.02).

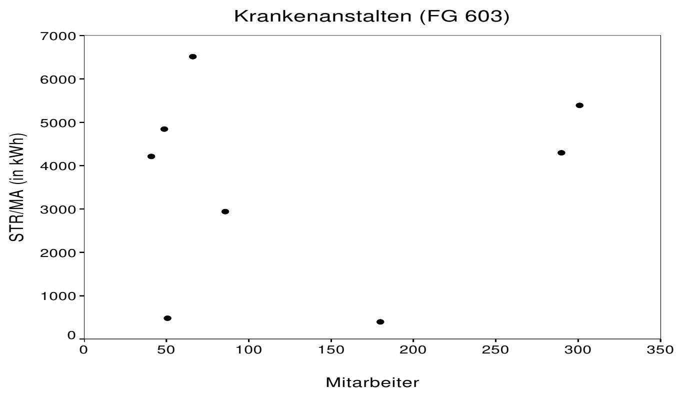



Abbildung 5.23: Scatterplot: STR/MA-Mitarbeiter ohne Beobachtung mit MA = 6080; FG 603

Anmerkung: In Abbildung 5.23 fallen zwei Betriebe durch einen geringen STR/MA auf. Bei diesen beiden Beobachtungen handelt es sich jedoch nicht um einen speziellen Typ von Krankenanstalten (ein Pflegeheim (51 MA) und ein Sanatorium (180 MA), wobei unter den restlichen Beobachtungen ebenfalls Pflegeanstalten und Sanatorien enthalten sind).

Abbildung 5.24 zeigt den Boxplot der Variablen STR/MA.

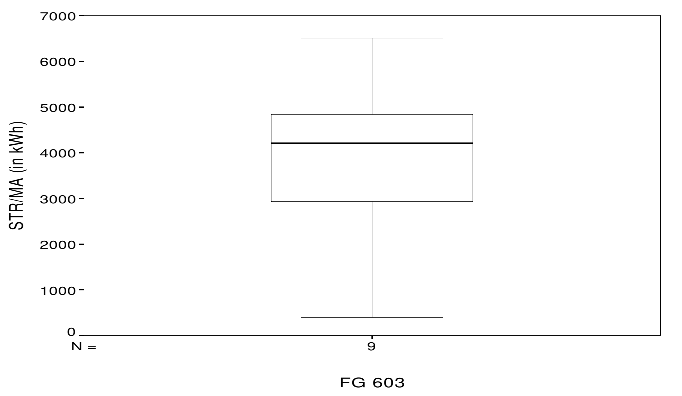



Abbildung 5.24: Boxplot der Variablen STR/MA; FG 603

Es ergeben sich folgende Schätzungen:

| Beob. | Mittelwert | Median | Min. | Max. | IQR | SD | SD/E | Schiefe |
|---|---|---|---|---|---|---|---|---|
| 9 | 3670 | 4212 | 398 | 6509 | 3401 | 2078 | 0.57 | -0.68 |

Für das 95%-Konfidenzintervall für den Erwartungswert ($\overline{X} \pm \frac{SD}{\sqrt{n}} t_{n-1;0.975}$) folgt: $[2073; 5249]$.

## 5.7 Schlußbemerkung

Analysiert man Abbildung 5.9 und 5.10 bzw. Tabelle 5.22, so sieht man, daß einige Großbetriebe noch nicht erhoben wurden bzw. der Stromverbrauch nicht bekannt ist. Hier würde ein geringer Mehraufwand (Erhebung weniger Großbetriebe) einen erheblichen Genauigkeitsgewinn bei Aussagen über den Gesamtstromverbrauch aller relevanten 1772 Betriebe bringen. Bei einem größeren Stichprobenumfang könnte die Gruppeneinteilung in der Globalanalyse spezifischer gestaltet werden und dadurch die Variabilität innerhalb einer Gruppe reduziert werden, wodurch sich die Konfidenzintervalllängen und -bereiche für die Lokationscharakteristiken ebenfalls verringern würden.

Betrachtet man sich die Auswertung einiger spezieller Fachgruppen, so fallen einige Punkte auf:

1. Bei den meisten Fachgruppen konnte keine bzw. nur eine sehr geringe Korrelation zwischen den Variablen STR/MA und Mitarbeiteranzahl festgestellt werden. Die Aussagekraft der berechneten Korrelationskoeffizienten ist jedoch zu relativieren, da Extremwerte aufgrund der geringen Stichprobenumfänge einen sehr großen Einfluß haben.

2. Bei Fachgruppen einer Sektion gibt es oft große Unterschiede bzgl. der Variablen STR/MA (siehe Abbildung 5.25).

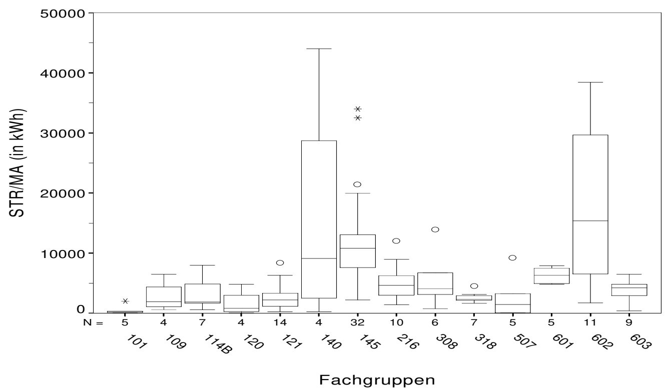



Abbildung 5.25: Boxplot der Variablen STR/MA; Vergleich der getrennt analysierten Fachgruppen

Anmerkung: In Abbildung 5.25 sind die Beobachtungen, die in Abschnitt 5.6 als Extremwerte ausgewiesen wurden, nicht enthalten. Aus der FG 602 sind die extremen Beobachtungen mit STR/MA = 95166 kWh und STR/MA = 70000 kWh nicht eingetragen, um die Graphik vergleichbarer gestalten zu können.

Wählt man daher die Sektion als Gruppierungsvariable (wie es bei der Globalanalyse aufgrund des geringen Stichprobenumfanges nötig war, um Aussagen über den Gesamtstromverbrauch treffen zu können), so resultieren Gruppen, die bzgl. der Variablen STR/MA sehr inhomogen sein können.

3. Einige Fachgruppen sind bzgl. der Variablen STR/MA sehr inhomogen (z. B. FG 140, 507). In diesen Fällen scheint selbst die Fachgruppe als Gruppierungsvariable zu grob zu sein, bei größerem Stichprobenumfang könnte man eine Varianzreduktion über die Berufsgruppenunterteilung erreichen.

4. Konträr dazu gibt es Fachgruppen (z. B. FG 121 und 318; sektionsübergreifend!), bei denen nahezu kein Unterschied zwischen den empirischen Verteilungen besteht.

Es können daher folgende Empfehlungen für einen auf der Kammersystematik aufbauenden Stichprobenplan gegeben werden.

1. **Planungsphase**:
   (a) Kausale Gruppenbildung aus der Fachgruppeneinteilung der Betriebe.
   (b) Großbetriebe sollen, sofern möglich, vollständig erhoben werden (vor Ort). Diese Betriebe haben daher auf die weitere Planung keinen Einfluß mehr.
   (c) Eruierung der jeweiligen Grundgesamtheiten. Prüfung, ob alle Berufsgruppen einer Fachgruppe bzgl. der Aufgabenstellung relevant sind ($\rightarrow$ Varianzreduktion).
   (d) Gestaltung eines, der Aufgabenstellung entsprechenden Fragebogens (siehe Glawar (1997)).

2. **Erhebungsphase**: Für Untersuchungen dieser Art scheint sich eine sequentielle Erhebungsmethode besonders zu eignen. Durch ständige Einbeziehung neuer Beobachtungen kann entschieden werden, ob der Stichprobenumfang schon ausreicht. Die Vorteile sind:
   (a) Da aus obigen Analysen ersichtlich ist, daß die Fachgruppenunterteilung nicht immer detailliert genug ist, um homogene Gruppen zu erhalten, kann bei einem sequentiellen Vorgehen flexibel entschieden werden, ob die gewählte Gruppierung noch zielführend ist. Dadurch kann auch die im Punkt 1.(a) empfohlene kausale Gruppenbildung verifiziert werden.
   (b) Der Stichprobenumfang wird, den Genauigkeitsanforderungen entsprechend, minimal gehalten. Somit minimiert sich auch der Erhebungsaufwand.

3. **Aufwand**: Der Nachteil eines sequentiellen Stichprobenplans liegt im höheren Organisationsaufwand, da unter Umständen der Zyklus 'Daten erheben und auswerten' mehrmals zu durchlaufen ist und jeweils Adaptierungen von Gruppen durchzuführen sind. Berücksichtigt man allerdings den Zeitaufwand für die Erhebung der vorliegenden 196 Betriebe mit Stromverbrauchsdaten, so hätte man mit einem durchdachten sequentiellen Plan (zusätzlich zu der zeitintensiven Befragung über den Postweg wäre ein 'elektronischer' Fragebogen (über e-mail, Fax) denkbar, ebenso Interviews durchgeführt von Ferialpraktikanten, Studenten,...) wahrscheinlich nicht nur Zeit und Geld gespart (siehe Glawar (1997)), sondern auch aussagekräftigere Resultate erhalten.

# Literaturverzeichnis


ANGUS, C. E. (1988). *Inferences on the lognormal mean for complete samples*, Commun. Statist. - Simulation and Computation, B 17, 1307-1331.

ANGUS, C. E. (1994). *Bootstrap one-sided confidence intervals for the lognormal mean*, Statistician, 43, 395-401.

COHEN, A. C. (1951). *Estimating Parameters of Logarithmic Normal Distributions by Maximum Likelihood*, JASA, 46, 206-212.

COHEN, A. C. AND WHITTEN, B. J. (1980). *Estimation in the Three-Parameter Lognormal Distribution*, JASA, 75, 399-404.

CROW, E. L. AND SHIMIZU, K. (1988). *Lognormal Distributions - Theory and Applications*, Statistics: Textbooks and Monographs, 88, Marcel Dekker, New York.

EFRON, B. (1979). *Bootstrap methods: another look at the jacknife*, The Annals of Statistics, 7, 1-26.

EFRON, B. (1987). *Better bootstrap confidence intervals*, JASA, 82, 171-200.

EFRON, B. (1990). *More efficient bootstrap computations*, JASA, 85, 79-89.

FENTON, L. F. (1960). *The sum of log-normal probability distributions in scatter transmission systems*, IRE Trans. Commun. Systems, CS-8, 57-67.

GLAWAR, G. (1997). *Statistische Evaluierung des Emissionskatasters der Stadt Graz*, Diplomarbeit, Institut für Statistik, TU-Graz.

HARTER, H. L. (1961). *Expected values of normal order statistics*, Biometrika, 48, 151-166.

HITZL, M. (1996). *Ermittlung branchenspezifischer Energiekennzahlen*, Diplomarbeit, Abteilung für Industrielehre und Innovationsforschung, TU-Graz.

KAISER, S. (1996). *Ermittlung branchenspezifischer Energiekennzahlen*, Diplomarbeit, Abteilung für Industrielehre und Innovationsforschung, TU-Graz.

LAND, C. E. (1971). *Confidence intervals for linear functions of the normal mean and variance*, Ann. Math. Statist., 42, 1187-1205.

LAND, C. E. (1972). *An evaluation of approximate confidence intervals for lognormal means*, Technometrics, 14, 145-158.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*, John Wiley & Sons, Inc., New York.

PEARSON, E. S. AND HARTLEY, H. O. (1972). *Biometrika Tables for Statisticans*, Vol. II, 58-65, 242-248, Cambridge University Press, Cambridge.

Rappitsch, G. (1991). *Resampling-Methoden zur Konstruktion von Fehlerschätzern und Konfidenzintervallen*, Diplomarbeit, Institut für Statistik, TU-Graz.

Shimizu, K., Iwase, K., (1981). *Uniformly minimum variance unbiased estimation in lognormal and related distributions*, Commun. Statist.-Theor. Meth., A10, 1127-1147.

Zhou, X.-H., and Gao, S. (1997). *Confidence intervals for the log-normal mean*, Statistics in Medicine, 16, 783-790.