

# Statistische Analyse von Konzentrationsmessungen in Abwasseranlagen

Klaus Felsenstein

Institut für Statistik, Wahrscheinlichkeitsth. & Versicherungsmath.  
Technische Universität Wien

**Zusammenfassung:** Für die Überwachung von Abwasserkläranlagen ist die Messung und Kontrolle der Konzentration verschiedener chemischer Substanzen erforderlich. Wir untersuchen die statistischen Eigenschaften der Meßwerte und die dadurch gegebenen Auswirkungen auf international verwendete Überwachungskriterien. Die Ergebnisse der Verteilungsanalyse ermöglichen es, Vorschläge für eine effizientere Analyse der Meßdaten vorzubringen. Wir geben optimale Methoden für die Berechnung mittlerer Konzentrationswerte an, woraus auch die optimale Anzahl von Messungen für das Erreichen vorgegebener Genauigkeitsanforderungen bestimmt werden kann.

**Abstract:** Monitoring of purification plants requires the measurement of several chemical parameters. We analyze the statistical properties of those measured values and consider the impact on official regulations for monitoring. The results of the analysis of the distribution yield to suggestions for an efficient application of monitoring regulations. Using optimal methods for estimation of mean values we calculate the necessary sample sizes in order to reach a given level of precision and compare it to necessary sample sizes derived from standard methods.

**Schlüsselwörter:** Verteilungsanalyse, Meßverfahren, Probenumfang, Log-Normalverteilung.

## 1 Meßverfahren

Die wichtigsten Gruppen von Stoffen in Abwässern sind Kohlenstoff-, Stickstoff- und Phosphorverbindungen. Die Beschaffenheit des Abwassers beschreiben einerseits Summenparameter ( $CSB_n$  - chemischer Sauerstoffbedarf -,  $BSB_n$  - biochemischer Sauerstoffbedarf -,  $TOC$  - gesamter organisch gebundener Kohlenstoff,  $n$  bezeichnet die Anzahl der Tage), die in erster Linie organische Kohlenstoffverbindungen erfassen. Die gesamte Stickstoffbelastung ( $TKN$ ) setzt sich aus reduzierten Stickstoffverbindungen,  $NH_4$ -N, org. N und  $NO_2$ -N (Nitrit) und  $NO_3$ -N (Nitrat) zusammen. Für den gesamten Phosphor-Gehalt,  $P$ -gesamt, ist ortho-Phosphat,  $PO_4$ -P, der wesentlichste Einzelstoff.

In den Kläranlagen werden nach vorgegebenen Prüfplänen möglichst homogene Abwasserproben aus dem Zulauf und dem Ablauf entnommen. Die Ziehung und Analyse der Proben erfolgt in unterschiedlichen Intervallen, von mehrfach täglich bis zu Wochenintervallen. Die Analytik zur Bestimmung obiger Parameter erfolgt nach verschiedenen biochemischen Verfahren, beispielsweise Verdünnungsverfahren, Kolorimetrie, Photometrie, Ionenchromatographie, und unterliegt Normen. In Moser und Thonhauser (1992) findet man eine detaillierte Beschreibung solcher Verfahren und damit verbundener Probleme.

Hier sind in der Folge die statistischen Eigenschaften der Parameterwerte von Interesse. Als Beobachtungswerte liegen für eine der drei Gruppen repräsentative Meßwerte (etwa *CSB*, *N*-gesamt, *P*-gesamt) vor.

## 2 Verteilung der Konzentrationswerte

Die für die verschiedenen Abwasseranlagen vorgeschriebenen Kriterien und Grenzwerte sind oft durch zwei grundlegende Annahmen die Daten betreffend motiviert. Die eine ist die Hypothese einer Normalverteilung für die Beobachtungen, die zweite Annahme betrifft die - zumindest weitgehende - Unabhängigkeit der Meßwerte. Werden diese Voraussetzungen akzeptiert, so vereinfacht sich die statistische Interpretation der Kriterien beträchtlich. Eine umfangreiche Analyse von Meßwerten verschiedener Kläranlagen im Rahmen eines interdisziplinären Forschungsprojekts an der Technischen Universität ergab allerdings, daß diese beiden Annahmen in der Praxis nicht haltbar sind. Die beiden Untersuchungen (Verteilungs- und Zeitreihenanalyse) sind als getrennte Fragestellungen zu verstehen. Erstere war vor allem für die Interpretation und Wirksamkeit der Kontrolle wichtig, zweite war für die Auswahl der Zeitpunkte für die Ziehung einer Probe für die Betreiber der Kläranlagen von Interesse.

Die Unabhängigkeitshypothese, die bei Betrachtung der chemischen Vorgänge von vornherein in Zweifel zu ziehen ist, bleibt dabei überraschender Weise zumindest für manche Modelle in modifizierter Form vertretbar oder hat nur relativ geringe Auswirkungen auf den statistischen Fehler. In den durchgeführten Analysen stellte sich aber die Normalverteilung als durchwegs ungeeignet für die statistische Modellierung der chemischen Meßwerte (z.B. Phosphor, Stickstoff, Ammonium, Ammoniak etc.) heraus. Die für verschiedene Verteilungstests notwendige Unabhängigkeit der Meßwerte wurde durch die Auswahl weit auseinanderliegender Werte erreicht und mit parameterfreien Tests abgesichert. Dabei erwies sich als hilfreich, daß in den Zeitreihen keinerlei Perioden bzw. lineare Trends nachweisbar waren und außerdem Meßwerte über mehrere Jahre vorhanden waren.

Die in den Daten vorhandene Asymmetrie legt die Verwendung einer Verteilung mit entsprechender schiefer Dichte nahe. Fast alle Meßreihen paßten sich an eine der beiden folgenden Verteilungstypen gut an.

Der größte Teil der Datensätze entspricht einer logarithmischen Normalverteilung  $X \sim \text{Log}N(\mu, \sigma^2)$ , wobei aber Ausreißer bzw. Fehlmessungen die Meßreihen kontaminieren können. Die Entfernung oder gesonderte Behandlung der Ausreißer ermöglichen adaptierte Ausreißertests, wie beispielsweise der David-Hartley-, Grubbs- oder der Dixon-Test (siehe Madansky, 1988). In den vorliegenden Daten waren bis zu 4% der Werte der logarithmischen Normalverteilung fremd und waren mit statistischen Methoden relativ leicht zu erkennen, da sie nur im oberen Spektrum der Spannweite auftraten.

Abbildung 1 zeigt ein typisches Histogramm von Phosphor-Werten im Ablauf nach eventueller Bereinigung und die angepaßte Dichte der logarithmischen Normalverteilung. Die Modellierung des Anteils der Daten, die nicht logarithmisch normalverteilt sind, in Form einer stetigen Gleichverteilung erwies sich in einigen Fällen als erfolgreich. Verteilungstests, wie beispielsweise der Kolmogoroff-Smirnoff-Test, führten auf p-Werte von

mehr als 50% für eine Mischverteilung einer logarithmischen Normalverteilung und einer Gleichverteilung. Das Mischungsgewicht der Gleichverteilung lag unter 2%, die logarithmische Normalverteilung bildete die tragende Modellverteilung.

Bei den Stickstoffwerten ergab sich das gleiche Bild, die Annahme der logarithmischen Normalverteilung wurde für Ablauf- und Zulaufwerte bestätigt. In wenigen Fällen eignete sich auch die (der logarithmischen Normalverteilung ähnliche) Gammaverteilung als Modellverteilung. Mischverteilungen treten bei Stickstoffkonzentrationen wie auch bei anderen Substanzen seltener auf.

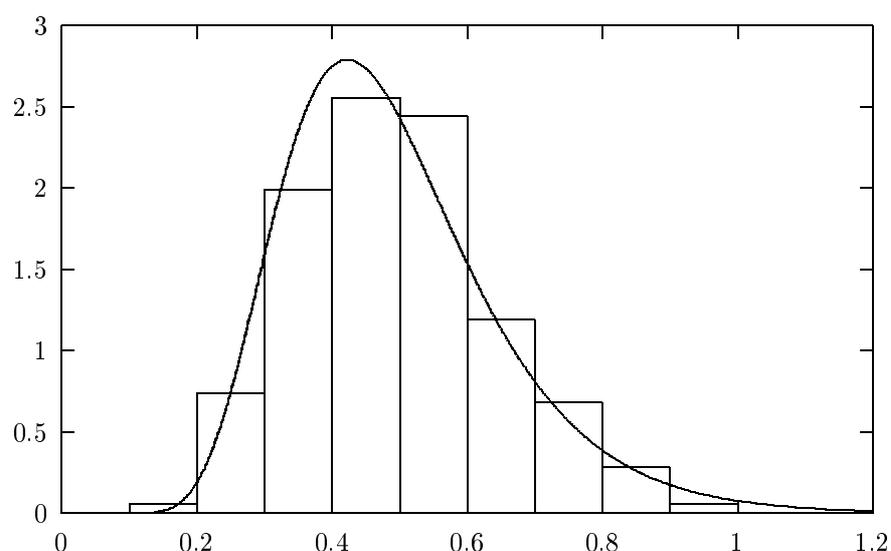


Abbildung 1: Phosphor Ablaufwerte [mg/l]

Die zweite Gruppe von Meßreihen bildete nicht unimodale Histogramme. Wird das Mischungsgewicht größer, kann man abweichende Messungen nicht mehr als Ausreißer behandeln, sondern kann den Prozeß aus zwei verschiedenen Prozessen bestehend ansehen (Abbildung 2). Durch geänderte Umweltbedingungen oder Aktivitäten der Anlagenbetreiber treten über einzelne Zeitspannen Prozesse auf verschiedenen Niveaus auf. Jeder der Prozesse entspricht einem Einzelprozeß einer logarithmisch normalverteilten Zeitreihe.

Als Gesamtverteilung empfiehlt sich daher eine Mischung von zwei (oder mehreren) logarithmischen Normalverteilungen für den gesamten Prozeß. Die Ergebnisse üblicher Anpassungstests bestätigen, daß sich solche Mischverteilungen gut an Meßreihen mit bimodalem Histogramm anpassen. Für die Werte aus Abbildung 2 zeigt Abbildung 3 das Histogramm mit entsprechender Dichte der Mischverteilung.

Die Anteile der beiden Verteilungen schwanken in den untersuchten Datensätzen vom bimodalen Typ zwischen 40% und 60%. Die Trennung der beiden logarithmisch normalverteilten Meßreihen sollte in der Praxis mit relativ einfachen Methoden der Clusteranalyse gelingen, zumindest wenn die verschiedenen chemischen Werte einzeln behandelt werden. Als Distanz für die Trennung der beiden Datensätze eignet sich ein eventuell gewichteter Euklidischer Abstand. Prinzipiell ist auch eine (Maximum Likelihood) Schätzung

des Anteils sowie der Parameter der logarithmischen Normalverteilungen durchführbar, wobei aber normalerweise numerische Probleme auftreten.

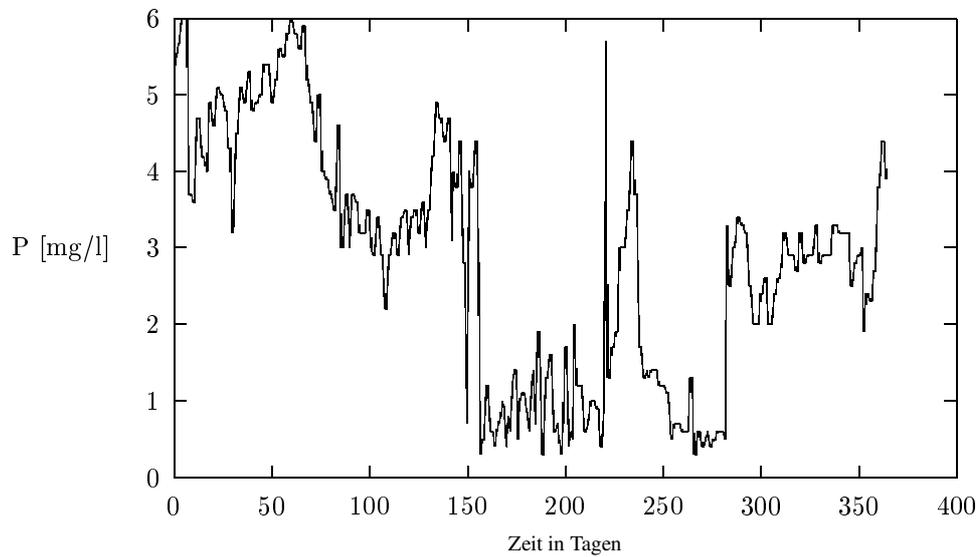


Abbildung 2: Zeitreihe Phosphor-Ablaufwerte

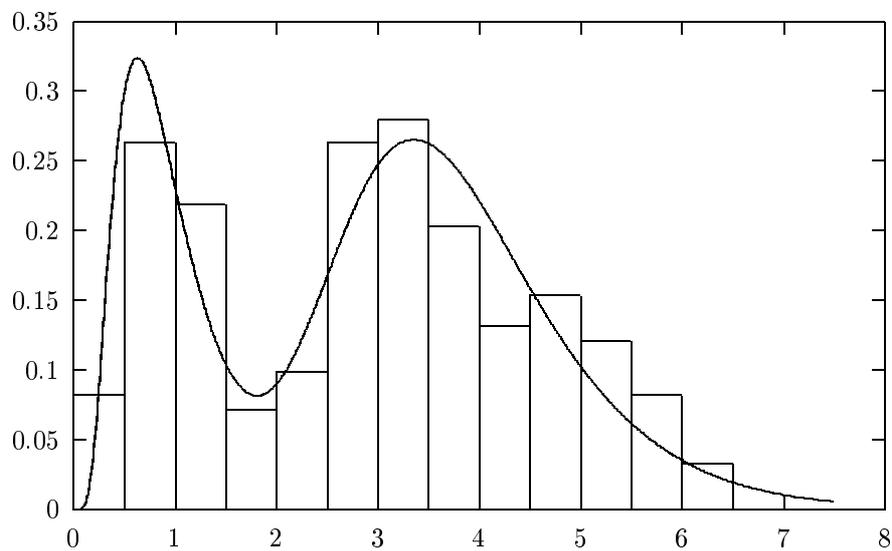


Abbildung 3: Histogramm und Dichte der Mischverteilung

Die Analyse von konkreten Ablaufdaten ergab, daß die Hypothese einer logarithmischen Normalverteilung eine brauchbare Grundlage für die Modellierung der chemischen Meßwerte in Abwasseranlagen ist, vorausgesetzt die Daten können von Ausreißern befreit und bei Mischverteilungen auch zerlegt werden.

### 3 Korrelationen

Für Daten aus Abwasseranlagen ist zu erwarten, daß aufeinanderfolgende Werte stark korreliert sind. Dies wird durch Unabhängigkeitstests, wie beispielsweise den von-Neumann-Test (siehe Madansky, 1988), für alle chemischen Parameter sowohl im Zulauf wie auch im Ablauf nachgewiesen.

Dem vorigen Abschnitt folgend ist die Untersuchung der logarithmierten Datenreihe statistisch einfacher durchführbar. Bei Messungen im Tagesabstand bleibt die Abhängigkeit normalerweise 3 bis 9 Tage signifikant, bei manchen Kläranlagen treten größere Korrelationen auch noch nach 14 Tagen auf. Abbildung 4 zeigt eine Autokorrelationsfunktion, die sich als Muster für viele Datensätze eignet.

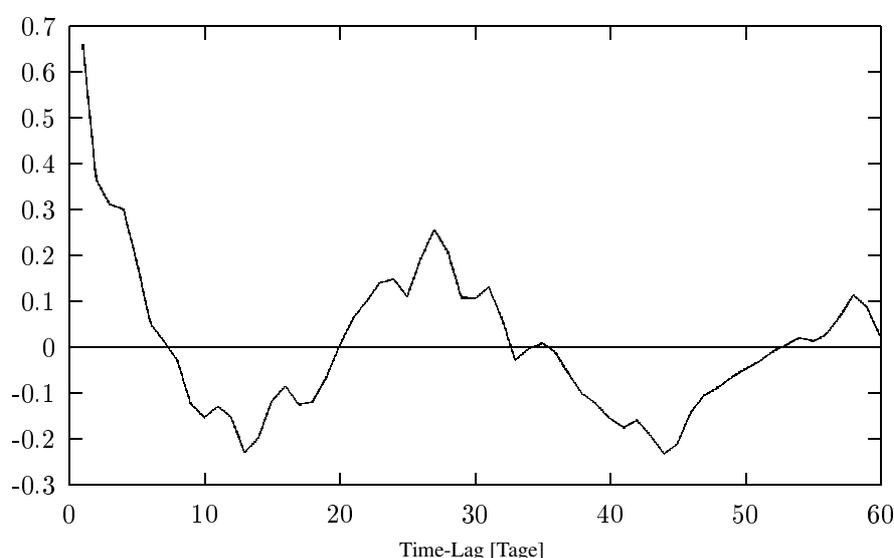


Abbildung 4: Autokorrelationsfunktion

Die meist auftretenden negativen Korrelationen für größere Zeitabstände lassen sich mit gesteigerter Aktivität der Kläranlagenbetreiber begründen, wenn hohe Schadstoffmengen im Zulauf registriert werden. Ein Ansatz als autoregressive Zeitreihe - AR bzw. ARIMA - führte auf eine zufriedenstellende Erklärung der Abhängigkeiten. Das Vorhersagefehler-Kriterium von Akaike (vgl. Akaike, 1971) empfiehlt bei den Zeitreihen häufig einen Time-Lag von 3 Tagen. Ein autoregressives Modell der Form

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \epsilon_t$$

bietet in der Regel eine brauchbare Grundlage für die Analyse und Vorhersage der (logarithmierten) Messungen.

Insbesondere bei der Kontrolle von Stickstoffwerten ist das Verhältnis der Ablaufwerte zu den Zulaufwerten die charakteristische Größe. Hier zeigt sich, daß die Abhängigkeit von Zulauf und Ablauf überraschend gering ist. Am selben Tag liegen so gut wie nie Abhängigkeiten zwischen den beiden Werten vor und falls sie überhaupt auftreten, dann erst nach ungefähr 3 Tagen.

Abbildung 5 zeigt eine Korrelationsfunktion für Zu- und Ablaufwerte des gesamten Stickstoffs einer Kläranlage.

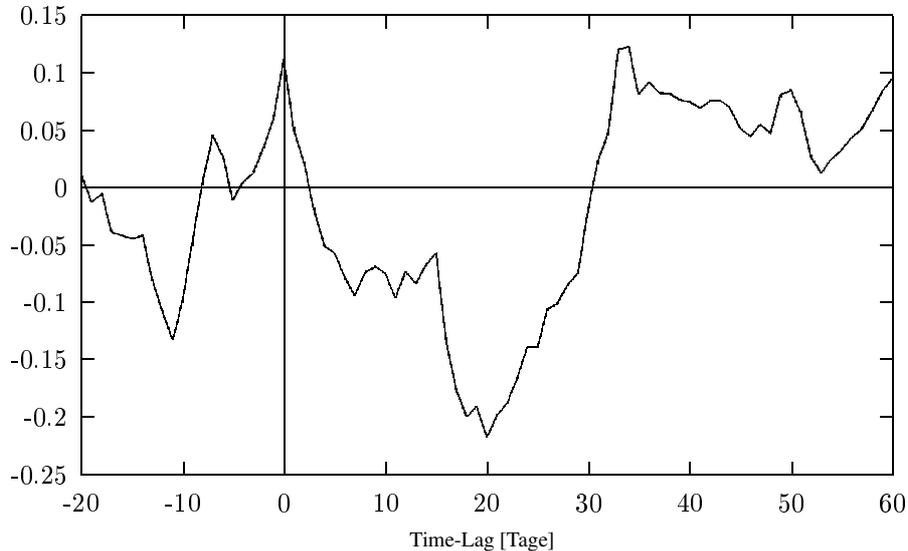


Abbildung 5: Korrelation von Zu- und Ablauf Stickstoff

## 4 Kontrollkriterien

Die Kontrollverfahren basieren auf der Vorgabe von Schwellenwerten für einzelne Meßwerte, Mittelwerte mehrerer Messungen über einen Zeitabschnitt oder von Wirkungsgraden. Unter der Voraussetzung, daß die Verteilung eine logarithmische Normalverteilung  $X \sim \text{LogN}(\mu, \sigma^2)$  ist und  $g$  den Grenzwert bezeichnet, gibt

$$p_g := \Phi\left(\frac{\mu - \log(g)}{\sigma}\right)$$

die Wahrscheinlichkeit für eine Überschreitung des Grenzwertes an. Dabei bezeichnet  $\Phi(\cdot)$  die Verteilungsfunktion der Standardnormalverteilung und  $\tilde{\mu} = \exp(\mu + \sigma^2/2)$  das Mittel und  $\tilde{\sigma}^2 = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$  die Varianz der logarithmischen Normalverteilung.

Die Kontrolle des Stickstoffgehalts erfolgt über den Wirkungsgrad, also die Effizienz der Anlage für die Elimination der Substanz. Wenn  $X_1$  den Zulaufmeßwert und  $X_2$  den Ablaufmeßwert bezeichnet, dann ist der Wirkungsgrad durch

$$\mathcal{W} = \frac{X_1 - X_2}{X_1}$$

definiert. Gilt  $X_1 \sim \text{LogN}(\mu_1, \sigma_1^2)$  und  $X_2 \sim \text{LogN}(\mu_2, \sigma_2^2)$ , so ist die Dichte von  $\mathcal{W}$  durch

$$f(w) = \frac{1}{1-w} \frac{1}{\sigma_{\mathcal{W}}} \phi\left(\frac{\log(1-w) - \mu_1 + \mu_2}{\sigma_{\mathcal{W}}}\right)$$

für  $w \leq 1$  gegeben. Dabei bezeichnet  $\phi(\cdot)$  die Dichte der Standardnormalverteilung,  $\sigma_{\mathcal{W}}^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$  und  $\rho$  den Korrelationskoeffizienten von  $(\log(X_1), \log(X_2))$ . Der Erwartungswert des Wirkungsgrads ist

$$\begin{aligned}\mathbb{E}\mathcal{W} &= 1 - e^{\mu_1 - \mu_2 + \sigma_{\mathcal{W}}/2} \\ &= 1 - \frac{\mathbb{E}X_1}{\mathbb{E}X_2} e^{\sigma_2^2 - \rho\sigma_1\sigma_2}.\end{aligned}$$

Die Anlage erfüllt das Kriterium, wenn der Wirkungsgrad im Mittel über einem Grenzwert (etwa 70 oder 80%) liegt.

In jedem Fall liegen Werte  $X_1, \dots, X_n$  (absolute Messungen oder Wirkungsgrade) und ein gegebener Grenzwert  $g$  vor. Dabei ist zu bemerken, daß Werte  $X_i$  oft selbst Mittelwerte sind. So werden oft Proben gleichzeitig an verschiedenen Stellen gezogen oder ein und dieselbe Probe nach unterschiedlichen biochemischen Verfahren analysiert. Beschränkungen für die Häufigkeit von Grenzwertüberschreitungen der  $X_i$  ergeben Überwachungskriterien für den erlaubten Gehalt von Substanzen, wie etwa Phosphor, im Abwasser.

Als Prüfkriterien sind hauptsächlich drei Methoden in Verwendung. Bei der ersten wird eine maximale Zahl von Überschreitungen vorgegeben, die zweite ist die *4von5-Regel* und bei der dritten wird der mittlere Wirkungsgrad über einen Zeitraum kontrolliert.

In der Europäischen Gemeinschaft ist für jede Anzahl  $n$  von Messungen des Phosphorgehalts eine maximale Anzahl an tolerierten Überschreitungen  $k_n$  vorgeschrieben. Dabei erfüllen die Zahlen  $k_n$  (möglichst genau) die Gleichung

$$\sum_{k=0}^{k_n} \binom{n}{k} p_g^k (1 - p_g)^{n-k} = p_g$$

für  $p_g = 0.05$ . Wird der Grenzwert mit 5% Wahrscheinlichkeit bei einer einzelnen Messung überschritten, dann ist auch die Wahrscheinlichkeit für die Verletzung des Kriteriums bei jeder Anzahl von Messungen 5%.

In Österreich wird die sogenannte *4von5-Regel* für die Überwachung des Phosphorgehalts angewandt. Von 5 aufeinanderfolgenden Meßwerten müssen mindestens 4 im Toleranzbereich sein und der zweifache Grenzwert darf keinesfalls überschritten werden. Betrachtet man die Ordnungsstatistik  $(X_{(1)}, \dots, X_{(5)})$  von 5 Messungen  $(X_1, \dots, X_5)$ , so muß für die Einhaltung der *4von5-Regel*  $X_{(4)} \leq g; X_{(5)} \leq 2g$  gelten. Die Wahrscheinlichkeit für dieses Ereignis ist

$$\mathbf{P}[X_{(4)} \leq g; X_{(5)} \leq 2g] = \sum_{i=1}^5 F_i(2g) - 4F_1(g),$$

wobei  $F_i(t)$  der Wert der gemeinsamen Verteilungsfunktion von  $(X_1, \dots, X_5)$  mit  $t$  als Argument an der  $i$ -ten Stelle und  $g$  als Argument an den übrigen Stellen bezeichnet. Für die *4von5-Regel* ist die Annahme, daß die 5 Messungen unabhängig sind, nicht vertretbar, während bei der EG Regelung von unabhängigen Werten ausgegangen werden kann, vorausgesetzt es liegen mindestens 3 bis 7 Tage zwischen zwei Messungen. Ausgehend von einer autoregressiven Zeitreihe  $\{X_t\}$  tritt eine Verletzung der *4von5-Regel* bei mehr als

8 Beobachtungen häufiger als die Überschreitung der EG-Vorschrift auf. Im praktischen Betrieb erweist sich die EU-Regel als ausgewogener und gegenüber außergewöhnlichen Einzelereignissen unempfindlicher.

## 5 Genauigkeit des Mittelwertes

Sowohl für die Einhaltung der Vorschriften als auch für die Überwachung der Anlage ist eine möglichst genaue Bestimmung des Mittelwertes jeder einzelnen chemischen Substanz äußerst wertvoll. Liegt zwischen zwei Messungen ein Zeitraum von mindestens 3 Tagen, dann kann man im allgemeinen die Werte als (annähernd) unabhängig annehmen. Die Schätzung des Erwartungswertes erfolgt meist mit dem Stichprobenmittel  $\bar{X}_n$ . Dies ist aber nur unter der Annahme, daß die Daten normalverteilt sind, die beste Schätzung des Erwartungswertes. Sind die Meßwerte aber, wie sich in der Praxis herausstellt, logarithmisch normalverteilt, basiert eine effiziente Mittelwertschätzung auf dem geometrischen Datenmittel

$$\bar{X}_G = \left( \prod_{i=1}^n X_i \right)^{1/n}.$$

Die erwartungstreue optimale (varianzminimale) Schätzung  $\tilde{T}_n$  des Mittels einer logarithmischen Normalverteilung hat die Form

$$\tilde{T}_n = 2^{(n-3)/2} \Gamma\left(\frac{n-1}{2}\right) \bar{X}_G \mathcal{I}_{(n-3)/2}\left(\tilde{S} \frac{n-1}{\sqrt{n}}\right) \left(\tilde{S} \frac{n-1}{\sqrt{n}}\right)^{(3-n)/2},$$

wobei  $\tilde{S}^2$  die Stichprobenvarianz der logarithmierten Werte und  $\mathcal{I}_\nu$  die modifizierte Besselfunktion erster Art ist,

$$\mathcal{I}_\nu(z) = \left(\frac{z}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(z^2/4)^k}{k! \Gamma(\nu + k + 1)}.$$

(Siehe Abschnitt 6). Mit wachsendem  $n$  entstehen numerische Schwierigkeiten bei der Berechnung der Besselfunktion. Eine Vereinfachung des Schätzers  $\tilde{T}_n$  erreicht man durch den Ersatz des Varianzschätzers durch den Parameter  $\sigma^2$ . Bei bekanntem Parameter  $\sigma^2$  ist

$$T_n := \bar{X}_G e^{(n-1)\sigma^2/2n} \quad (1)$$

der effiziente Schätzer (siehe Abschnitt 6). Die Annahme, daß  $\sigma^2$  bekannt ist, erweist sich für die untersuchten Daten als meist akzeptabel, dieser Parameter variiert für die meisten Meßreihen kaum. Setzt man den Schätzer  $\tilde{S}^2$  anstelle von  $\sigma^2$  in (1) ein, dann überschätzt  $T_n$  den Mittelwert. Diese Verzerrung war bei den gegenständlichen Meßreihen aber kaum feststellbar.

Die Verteilung der Schätzfunktion (1) ist wieder eine logarithmischen Normalverteilung  $T_n \sim \text{LogN}(\mu + (n-1)\sigma^2/2n; \sigma^2/n)$ . Der Effizienzverlust des Stichprobenmittels  $\bar{X}_n$  gegenüber  $T_n$ , dargestellt als Verhältnis der Varianzen, beträgt

$$\frac{e^{\sigma^2/n} - 1}{(e^{\sigma^2} - 1)/n}$$

und konvergiert für  $n \rightarrow \infty$  gegen  $\sigma^2/(e^{\sigma^2} - 1)$ . Die Brauchbarkeit von  $\bar{X}_n$  nimmt also mit wachsendem  $\sigma^2$  ab.

Die verbesserte Schätzmethode des Mittels durch  $T_n$  zeigt sich auch in der Verkleinerung des relativen Fehlers  $(T_n - \tilde{\mu})/\tilde{\mu}$ . Die Anzahl der für das Erreichen einer bestimmten Genauigkeit notwendigen Messungen verringert sich bei der Anwendung von  $T_n$ . Die (approximierten) Abweichungswahrscheinlichkeiten für den relativen Fehler sind für  $\bar{X}_n$

$$\mathbf{P} \left[ \left| \frac{\bar{X}_n - \tilde{\mu}}{\tilde{\mu}} \right| \leq t \right] \cong 2\Phi \left( \frac{t\sqrt{n}}{\sqrt{e^{\sigma^2} - 1}} \right) - 1,$$

und für  $T_n$  gilt bei  $t < 1$

$$\mathbf{P} \left[ \left| \frac{T_n - \tilde{\mu}}{\tilde{\mu}} \right| \leq t \right] = \Phi \left( \frac{\log(1+t)\sqrt{n}}{\sigma} + \frac{\sigma}{2\sqrt{n}} \right) - \Phi \left( \frac{\log(1-t)\sqrt{n}}{\sigma} + \frac{\sigma}{2\sqrt{n}} \right),$$

und für  $t \geq 1$

$$\mathbf{P} \left[ \left| \frac{T_n - \tilde{\mu}}{\tilde{\mu}} \right| \leq t \right] = \Phi \left( \frac{\log(1+t)\sqrt{n}}{\sigma} + \frac{\sigma}{2\sqrt{n}} \right).$$

Der erwartete relative Fehler bei beiden Schätzverfahren beträgt

$$\mathbf{E} \left| \frac{\bar{X}_n - \tilde{\mu}}{\tilde{\mu}} \right| \cong \sqrt{\frac{2(\exp(\sigma^2) - 1)}{\pi n}}$$

bzw.

$$\mathbf{E} \left| \frac{T_n - \tilde{\mu}}{\tilde{\mu}} \right| = 4\Phi \left( \frac{\sigma}{2\sqrt{n}} \right) - 2.$$

Bei den Analysen der Abwasserparameter wurden öfters Werte für  $\sigma^2$  in der Größenordnung  $\sigma^2 = 0.12$ ,  $\sigma^2 = 0.35$  und  $\sigma^2 = 1.5$  errechnet. In den folgenden Tabellen werden die Abweichungswahrscheinlichkeiten für diese Werte von  $\sigma^2$  und eine Grenze für den relativen Fehler von 5% ( $t = 0.05$ ) verglichen.

Tabelle 1: Abweichungswahrscheinlichkeit für  $\sigma^2 = 0.12$

n	$\mathbf{P} \left[ \left  \frac{\bar{X}_n - \tilde{\mu}}{\tilde{\mu}} \right  \leq t \right]$	$\mathbf{P} \left[ \left  \frac{T_n - \tilde{\mu}}{\tilde{\mu}} \right  \leq t \right]$
10	.3421	.3519
15	.4124	.4240
30	.5569	.5712
50	.6779	.6930
100	.8386	.8514

Tabelle 2: Abweichungswahrscheinlichkeit für  $\sigma^2 = 0.35$ 

n	$\mathbf{P} \left[ \left  \frac{\bar{X}_n - \hat{\mu}}{\hat{\mu}} \right  \leq t \right]$	$\mathbf{P} \left[ \left  \frac{T_n - \hat{\mu}}{\hat{\mu}} \right  \leq t \right]$
10	.1930	.2101
15	.2352	.2562
30	.3277	.3565
50	.4150	.4501
100	.5601	.6024

Bei größerem  $\sigma^2$  sollte auch die Grenze für den relativer Fehler höher angesetzt werden, im nächsten Beispiel ist diese Grenze 10%.

Tabelle 3: Abweichungswahrscheinlichkeit für  $\sigma^2 = 1.5$ 

n	$\mathbf{P} \left[ \left  \frac{\bar{X}_n - \hat{\mu}}{\hat{\mu}} \right  \leq t \right]$	$\mathbf{P} \left[ \left  \frac{T_n - \hat{\mu}}{\hat{\mu}} \right  \leq t \right]$
10	.1346	.2012
15	.1644	.2466
30	.2309	.3451
50	.2953	.4370
100	.4080	.5873

Die Vorgabe einer Sicherheit  $\beta$  für den relativen Fehler ermöglicht die Bestimmung einer optimalen Anzahl von Messungen für das Erreichen dieser Sicherheit. Für das Stichprobenmittel ist die Anzahl

$$n_{\bar{X}} = \Phi^{-1} \left( \frac{1 + \beta}{2} \right) (e^{\sigma^2} - 1) t^{-2}$$

und die optimale Anzahl  $n_T$  für  $T_n$  löst (für  $t < 1$ ) die Gleichung

$$\Phi \left( \frac{\log(1+t)\sqrt{n}}{\sigma} + \frac{\sigma}{2\sqrt{n}} \right) - \Phi \left( \frac{\log(1-t)\sqrt{n}}{\sigma} + \frac{\sigma}{2\sqrt{n}} \right) = \beta.$$

Zu den obigen Beispielen erhält man bei  $\sigma^2 = 0.12$ ,  $t = 0.05$  und  $\beta = 0.9$  einen Stichprobenumfang von  $n_{\bar{X}} = 138$  bzw.  $n_T = 130$ . Bei größerem  $\sigma$  ist die Einsparung an Beobachtungen bei  $T_n$  deutlicher, für  $\sigma^2 = 1.5$ ,  $t = 0.15$  und  $\beta = 0.85$  ergibt sich  $n_{\bar{X}} = 321$  bzw.  $n_T = 137$ .

Die effiziente Schätzung des mittleren Wirkungsgrads bei festem  $\varrho$ ,  $\sigma_1$ ,  $\sigma_2$  und dem Quotienten  $Z = X_2/X_1$  aus Ablauf- und Zulaufgröße ist demnach

$$\mathcal{W}^* = 1 - Z_G e^{(n-1)\sigma_W^2/2n}.$$

Den Grad der Effizienzsteigerung dieses Schätzers gegenüber dem Stichprobenmittel bestimmt wieder der Parameter  $\sigma_W^2$ .

## 6 Anhang

Dieser Abschnitt enthält die theoretische Basis für die in den vorigen Abschnitten beschriebene Vorgangsweise.

**RESULTAT 1** Die Stichprobe  $X_1, \dots, X_n$  folge einer logarithmischen Normalverteilung  $\text{Log}N(\mu, \sigma^2)$ . Dann ist die effiziente Schätzung des Erwartungswertes durch

$$\tilde{T}_n = 2^{(n-3)/2} \Gamma\left(\frac{n-1}{2}\right) \bar{X}_G \mathcal{I}_{(n-3)/2}\left(\tilde{S} \frac{n-1}{\sqrt{n}}\right) \left(\tilde{S} \frac{n-1}{\sqrt{n}}\right)^{(3-n)/2}$$

gegeben, wobei  $\tilde{S}^2$  die Stichprobenvarianz für die logarithmierten Beobachtungen,  $\bar{X}_G$  das geometrische Mittel der Beobachtungen und  $\mathcal{I}_\nu$  die modifizierte Besselfunktion erster Art ist,

$$\mathcal{I}_\nu(z) = \left(\frac{z}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(z^2/4)^k}{k! \Gamma(\nu + k + 1)}.$$

**BEWEIS.** Aus der Likelihoodfunktion

$$l(\mu, \sigma^2) = [(2\pi)^{n/2} \sigma^n \prod_i x_i]^{-1} \exp\left(-\frac{1}{2\sigma^2} \sum_i (\log(x_i) - \mu)^2\right)$$

ergibt sich die Suffizienz von  $\tilde{S}^2$  und  $\tilde{X}$ , dem Stichprobenmittel der logarithmierten Werte. Die Vollständigkeit von  $\tilde{X}$ ,  $\tilde{S}^2$  ergibt sich aus der Eigenschaft der Exponentialfamilie, zu der die logarithmische Normalverteilung gehört. Betrachtet man für  $j \in \{1, \dots, n\}$

$$\tilde{X}_j := \sum_{i \neq j} \log(X_i)$$

und

$$\tilde{S}_j := \sum_{i \neq j} \log(X_i)^2 - \frac{\tilde{X}_j^2}{n-1},$$

dann ist die Dichte von  $X_j, \tilde{X}_j, \tilde{S}_j$  durch

$$f(x, y, z) = \frac{1}{\sigma^4 \sqrt{n-1} x} \phi\left(\frac{\log(x) - \mu}{\sigma}\right) \phi\left(\frac{y - (n-1)\mu}{\sigma \sqrt{n-1}}\right) g\left(\frac{z}{\sigma^2}\right)$$

gegeben, wobei  $g$  die Dichte einer  $\chi^2$ -Verteilung mit  $n-2$  Freiheitsgraden ist. Durch den Zusammenhang

$$\tilde{X} = \tilde{X}_j + \log(X_j)$$

und

$$(n-1)\tilde{S}^2 = \tilde{S}_j + \frac{n-1}{n} \log(X_j)^2 + \tilde{X}_j^2 \frac{1}{n(n-1)} - \frac{2}{n} \tilde{X}_j \log(X_j)$$

ergibt sich für die Verteilung von  $X_j, \tilde{X}, (n-1)\tilde{S}^2$  die Dichte

$$f(x, \tilde{x}, \tilde{s}) = \frac{1}{\sigma^4 \sqrt{n-1} x} \phi\left(\frac{\log(x) - \mu}{\sigma}\right) \phi\left(\frac{\tilde{x} - \log(x) - (n-1)\mu}{\sigma \sqrt{n-1}}\right) g\left(\frac{\tilde{s}}{\sigma^2}\right)$$

mit  $z = \tilde{s} - n(\log x - \tilde{x}/n)^2/(n-1)$ . Die bedingte Dichte von  $X_j|\tilde{X}, (n-1)\tilde{S}^2$  ist

$$f(x|\tilde{x}, \tilde{s}) = \frac{\sqrt{n}}{\sqrt{(n-1)\tilde{s}}} \frac{1}{B(1/2, n/2-1)} \frac{1}{x} \left(1 - (\log(x) - \frac{\tilde{x}}{n})^2 \frac{n}{(n-1)\tilde{s}}\right)^{n/2-2}$$

für  $\exp(\tilde{x}/n - \sqrt{\tilde{s}(n-1)/n}) \leq x \leq \exp(\tilde{x}/n + \sqrt{\tilde{s}(n-1)/n})$ .

Die bedingte Verteilung von

$$\left(\log(X_j) - \frac{\tilde{X}}{n}\right)^2 \frac{n}{(n-1)\tilde{S}}$$

ist somit eine Beta-Verteilung  $Be(1/2, n/2-1)$ . Nach dem Satz von Lehmann/Scheffé ist

$$\tilde{T}_n = \mathbb{E}(\bar{X}_n|\tilde{X}, \tilde{S}^2),$$

und dieser bedingte Erwartungswert ist

$$\mathbb{E}(\bar{X}_n|\tilde{X}, \tilde{S}^2) = \exp\left(\frac{\tilde{X}}{n}\right) \frac{1}{B(1/2, n/2-1)} \int_{-1}^1 (1-t^2)^{n/2-2} \exp\left(\frac{t\sqrt{(n-1)\tilde{S}}}{\sqrt{n}}\right) dt.$$

Die Integraldarstellung der modifizierten Besselfunktion erster Art bringt dann das Ergebnis.  $\square$

**RESULTAT 2** Die Stichprobe  $X_1, \dots, X_n$  folge einer logarithmischen Normalverteilung  $LogN(\mu, \sigma^2)$  mit bekanntem Parameter  $\sigma^2$ . Dann ist die effiziente Schätzung des Erwartungswertes durch

$$T_n := \bar{X}_G e^{(n-1)\sigma^2/2n}$$

gegeben.

**BEWEIS.** Die Statistik  $\tilde{X}$  des vorigen Beweises erfüllt jetzt die selben Voraussetzungen. Die Transformation  $\tilde{X} = \tilde{X}_j + \log(X_j)$  führt auf die bedingte Dichte

$$f(x|\tilde{x}) = \frac{\sqrt{n}}{\sigma\sqrt{n-1}x} \phi\left(\frac{\log(x) - \mu}{\sigma}\right) \phi\left(\frac{\tilde{x} - \log(x) - (n-1)\mu}{\sigma\sqrt{n-1}}\right) \phi\left(\frac{\tilde{x} - n\mu}{\sqrt{n}\sigma}\right)^{-1}$$

von  $X_j|\tilde{X}$ . Die optimale Schätzung erfolgt wieder durch den bedingten Erwartungswert

$$\mathbb{E}(\bar{X}_n|\tilde{X}) = \exp\left(\frac{\tilde{X}}{n} + \frac{(n-1)\sigma^2}{2n}\right),$$

dies ist eine Darstellung von  $T_n$ .  $\square$

## Literatur

H. Akaike. Autoregressive Model Fitting for Control. *Ann. Inst. Statist. Math.*, 23, pp. 163-180, 1971.

- G. Box and G. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, 1976.
- E. Crow and K. Shimizu (eds.). *Lognormal Distributions*. Marcel Dekker, New York, 1988.
- A. Lawson and L. Waller. A Review of Point Pattern Methods for Spatial Modelling of Events around Sources of Pollution. *Environmetrics*, Vol. 7, pp. 471 - 489, 1996.
- E. Lehmann. *Theory of Point Estimation*. Wiley, New York, 1983.
- A. Madansky. *Prescriptions for Working Statisticians*. Springer-Verlag, New York, 1988.
- G. Mance. Comparison of the Waste Water Controls and River Water Quality in EU. Forschungsbericht, Bristol, 1994.
- D. Moser und C. Thonhauser. Probleme bei den Summenparametern CSB, BSB5, TOC. *Wiener Mitteilungen - Wasser-Abwasser-Gewässer*, Bd. 108, F1-F46, 1992.
- V. Neitzel und R. Klopp. Statistischer Vergleich der Anforderungen der EG-Richtlinie über die Behandlung kommunaler Abwässer und des Anhanges 1 zur Rahmenabwasser VwV für Stickstoff. *Korrespondenz Abwasser*, 6/93, 948-957, 1993.
- L. Sachs. *Angewandte Statistik*. 7. Auflage. Springer-Verlag, Berlin, 1992.
- Standard Methods for the Examination of Water and Wastewater*. 17<sup>th</sup> Ed., American Public Health Association, 1989.
- W. Timischl. Statistische Betrachtungen zur Überprüfung der Einhaltung von Grenzwerten. *Wiener Mitteilungen - Wasser-Abwasser-Gewässer*, Bd. 108, D1-D13, 1992.

Adresse des Autors:

ao. Univ.-Prof. Dr. Klaus Felsenstein  
Institut für Statistik, Wahrscheinlichkeitstheorie und Versicherungsmathematik  
Technische Universität Wien  
Wiedner Hauptstr. 8-10  
A-1040 Wien  
Tel. +43 1 58801 / 10722  
Elec. Mail: Klaus.Felsenstein@tuwien.ac.at