# Verfahren zur statistischen Auswertung von Daten mit als „< G“ dokumentierten Werten

DOKUM/STATAUSW

Bearbeiter:
G. Kanisch

ISSN 1865-8725 Version November 1998

# Verfahren zur statistischen Auswertung von als „< G" dokumentierten Werten

## 1 Vorbemerkungen

Ein Datensatz ist im Folgenden eine Stichprobe, die signifikante Daten, welche oberhalb der ihnen zuzuordnenden Erkennungsgrenzen *G** liegen und mit ihrem Wert dokumentiert sind, sowie nicht-signifikante Daten, die unterhalb der ihnen zuzuordnenden Erkennungsgrenzen *G** liegen und als „< G"-Wert dokumentiert sind, enthält. G bezeichnet hier die Nachweisgrenze.

Es wird empfohlen, ein auf dem Wahrscheinlichkeits-Plot beruhendes Verfahren (7, 10) zu verwenden. Mit Hilfe linearer Regression der im Wahrscheinlichkeitsnetz aufgetragenen signifikanten Daten werden für die nicht-signifikanten Werte (dokumentiert als „< G") Ersatzwerte ermittelt, so dass die Stichprobe anschließend frei von „< G"-Werten ist. Dann wird in gewohnter Weise die Berechnung der üblichen Kennwerte wie Mittelwert, Median, Standardabweichung u. a. m. durchgeführt. Dieses Verfahren kann bei größeren Datensätzen problemlos angewendet werden, erfordert aber den Einsatz eines Rechenprogrammes.

## 2 Verfahren mit Wahrscheinlichkeitscheinlichkeits-Plot

### 2.1 Im Falle einer einzigen Nachweisgrenze *G*

Dieses Verfahren stellt eine Vorstufe zum Verfahren nach Helsel und Cohn (1) dar.

Es wird eine Stichprobe betrachtet, in der „< G"-Werte mit nur einer einzigen numerisch verschiedenen Nachweisgrenze *G* auftreten. Unterhalb dieser Nachweisgrenze dürfen keine weiteren, als nachgewiesen betrachteten signifikanten Werte auftreten.

Liegen insgesamt n Werte $y_i$ (i = 1, 2, ..., n) vor, wobei k Werte (k < n) unter der einen Nachweisgrenze liegen, wird dazu wie nachfolgend beschrieben vorgegangen.

Die Messwerte werden zunächst der Größe nach aufsteigend sortiert. Für den Wahrscheinlichkeits-Plot wird jedem Wert $y_i$, auch den „< G"-Werten, die empirische Plot-Wahrscheinlichkeiten $p_i = i/(n+1)$ zugeordnet, zu denen dann die Quantile der Normalverteilung $z_i$ bestimmt werden. Mit Hilfe von Tabellen aus statistischer Standardliteratur oder eines Rechenprogrammes werden letztere über das Inverse der (integralen) Verteilungsfunktion der Normalverteilung als $z_i = \Phi^{-1}(p_i)$ berechnet. Allein mit den signifikanten Werten (i > k) werden dann mit den logarithmierten $y_i$ und ihren Quantilen $z_i$ die Parameter a und b einer Regressionsgeraden ermittelt:

$$\ln y_i = a + b \cdot z_i \qquad (i > k) \tag{1}$$

Die – hier nicht erforderliche – grafische Darstellung der Quantile $z_i$ als "x-Werte" und $\ln y_i$ als "y-Werte" (oder vice versa) bezeichnet man hierbei als Wahrscheinlichkeits-Plot (in diesem Falle ein logarithmischer Plot). Nach Vorschlag von Helsel (2) erfolgt die Logarithmierung der Daten in Gleichung (1) auf jeden Fall, unabhängig

davon, ob eine normale, lognormale oder eine andere Verteilung vorliegt. Es sei darauf hingewiesen, dass von der Nachweisgrenze G nur die Anzahl ihres Auftretens im Datensatz direkt in die Rechnung eingeht, nicht aber ihr Wert!

Die Ersatzwerte $q_i$ für die „< G"-Werte $y_i$ ($i \leq k$) können nun unter Verwendung ihrer zugeordneten Quantile $z_i$ und der Regressionsparameter a und b berechnet werden:

$$q_i = e^{a+b \cdot z_i} \quad \text{mit } (i \leq k) \tag{2}$$

Die $q_i$ ($i \leq k$) und $y_i$ ($i > k$) zusammengenommen stellen einen vollständigen Datensatz dar, mit dem in üblicher Weise verschiedene Schätzer für Mittelwert, Standardabweichung und Medianwert bestimmt werden können.

Nach der Berechnung der Ersatzwerte $q_i$ wird die Regressionsgerade Gleichung (1) nicht weiter verwendet. Es werden, zusammen mit den $q_i$, wieder die originalen signifikanten $y_i$-Werte für die Parameterschätzung benutzt. Damit wird der Charakter der zugrunde liegenden Verteilung, die in der Praxis meist weder normal- noch lognormalverteilt sein wird, weitgehend wieder hergestellt, soweit der Prozentsatz der „< G"-Werte nicht zu hoch ist. Dieses Verfahren wird als ein sogenanntes **robustes** Verfahren bezeichnet, das relativ unempfindlich gegenüber Abweichungen von Verteilungen wie Normal- oder Lognormalverteilung ist. Alternative Verfahren der Anpassung an bekannte Verteilungen, z. B. „klassische" Maximum-Likelihood-Verfahren (3), haben exakte Verteilungsannahmen zur Voraussetzung und reagieren daher viel empfindlicher auf solche Verteilungsabweichungen. Ein einfaches Beispiel eines weniger robusten Verfahrens wäre die Verwendung der Fitparameter a und b aus Gleichung (1) (Anpassung an eine Lognormalverteilung!) zur Bestimmung des Medians und der Standardabweichung.

Für die Durchführung des geschilderten Rechenganges **ist Voraussetzung, dass mindestens zwei signifikante Werte vorliegen müssen**. Diese müssen überdies verschieden voneinander sein. Hieraus ergibt sich eine Einschränkung der Anwendungsmöglichkeit bei Datensätzen mit kleiner Anzahl von Werten.

*Beispiel:*
Dokumentierte Messwerte: 2,13; 1,55; 1,40; < 1,30; 1,80; < 1,30.
Es ist n = 6, k = 2.

Die sich für dieses Beispiel ergebenden obigen Rechengrößen sind in der nachfolgenden Tabelle aufgeführt.

**Tab. 1:**

| sortierte $y_i$ | ln $y_i$ | $p_i = i/(n+1)$ | $z_i = \Phi^{-1}(p_i)$ | $q_i = e^{a+b \cdot z_i}$ [1)] | "vollständige" $y_i$ [2)] |
|---|---|---|---|---|---|
| < 1,30 | | 0,14 | -1,07 | 1,03 | 1,03 |
| < 1,30 | | 0,28 | -0,56 | 1,22 | 1,22 |
| 1,40 | 0,33 | 0,42 | -0,18 | | 1,4 |
| 1,55 | 0,43 | 0,57 | 0,18 | | 1,55 |
| 1,80 | 0,58 | 0,71 | 0,56 | | 1,8 |
| 2,13 | 0,75 | 0,85 | 1,07 | | 2,13 |

[1)] Parameter aus linearer Regression: a = 0,39020 und b = 0,34143

[2)] letzte Spalte ergibt: Mittelwert = 1,52 und Standardabweichung = 0,400

Die nachfolgende Abbildung zeigt den dazugehörigen Wahrscheinlichkeitsplot.

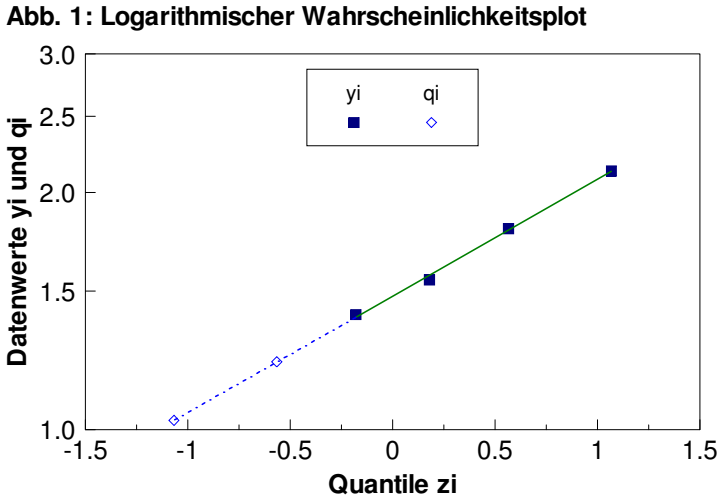


**Abb. 1:** Logarithmischer Wahrscheinlichkeitsplot

## 2.2 Im Falle mehrerer Nachweisgrenzen *G* (Verfahren nach Helsel und Cohn)

Dieser Fall ist im Wesentlichen dadurch gekennzeichnet, dass nach einer (speziellen) Sortierung der n Messwerte (einschließlich der verschieden großen Nachweisgrenzenwerte) signifikante und nicht-signifikante Werte sich in beliebiger Weise in der Reihenfolge abwechseln können. So können mehrere „< G"-Werte auch oberhalb des größten signifikanten Wertes, aber auch signifikante Daten unterhalb der kleinsten Nachweisgrenze liegen. Für die Behandlung dieses Falles wird eine auf Arbeiten von Helsel und Cohn (1) sowie von Hirsch und Stedinger (4) basierende Erweiterung des unter Abschnitt 2.1 beschriebenen Verfahrens des Wahrscheinlichkeits-Plots für den Fall einer einzigen Nachweisgrenze vorgeschlagen.

Zunächst wird eine auf einem Ansatz aus der Wahrscheinlichkeitsrechnung basierende Methode zur Bestimmung der Plot-Wahrscheinlichkeiten dargestellt.

Es sei ein Datensatz mit insgesamt n Elementen $y_i$ und darin enthaltenen nicht-signifikanten Werten („< G"-Werte) gegeben. Die Daten werden zunächst aufsteigend sortiert. Sind als Beispiel (1) folgende n Werte gegeben:

< 1, 7, 9, < 1, < 1, < 1, 15, < 1, 12, < 1, < 10, < 10, < 10, 3, 33, 27, 20 und 50,

so entsteht nach Sortierung die Reihenfolge

< 1, < 1, < 1, < 1, < 1, < 1, 3, 7, 9, < 10, < 10, < 10, 12, 15, 20, 27, 33 und 50.

Die insgesamt m numerisch voneinander verschiedenen nicht-signifikanten Werte („< "G-Werte), seien $X_1$, $X_2$, ..., $X_m$. Dann gilt: $X_1 < X_2 < ... < X_m$. Es wird $X_{m+1} = \infty$ definiert.

Für jede der (numerisch verschiedenen) Nachweisgrenzen $X_j$ (j = 1, 2, ... , m) werden dann durch Abzählen in der sortierten Folge Anzahlen $A_j$, $B_j$ und $C_j$ wie folgt definiert:

$A_j$ = Anzahl der signifikanten Werte y, die oberhalb der j-ten und unterhalb der nächsthöheren Nachweisgrenze liegen: $X_j \le y < X_{j+1}$;

$B_j$ = Anzahl aller Werte (signifikante und nicht-signifikante), die unterhalb der j-ten Nachweisgrenze liegen (bis herunter zum kleinsten Wert des gesamten Datensatzes): $y \le X_j$;

$C_j$ = Multiplizität der Nachweisgrenze $X_j$ (wenn $X_i$ mehrfach vorkommt).

Für das obige Beispiel gilt:

$A_1 = 3$ und $A_2 = 6$;

$B_1 = 6$ und $B_2 = 12$;

$C_1 = 6$ und $C_2 = 3$.

Nun wird nach der empirischen Wahrscheinlichkeit $pe_j$ für die Überschreitung der Nachweisgrenze $X_j$ gefragt. Diese wird definiert als (P bezeichnet hier allgemein eine Wahrscheinlichkeit)

$$pe_j = P[y \ge X_j].$$

Zu ihrer Bestimmung wird ein Ansatz aus der Wahrscheinlichkeitsrechnung verwendet (4). Ist die empirische Wahrscheinlichkeit $pe_{j+1}$ zur Überschreitung der nächsthöheren Nachweisgrenze $X_{j+1}$ gegeben (definitionsgemäß soll $pe_{m+1} = 0$ gelten), lässt sich ein rekursiver Zusammenhang wie folgt herstellen:

$$pe_j = P[y \ge X_{j+1}] + P[X_j \le y < X_{j+1} \mid y < X_{j+1}] \cdot P[y < X_{j+1}] \qquad (3)$$

Zum Verständnis dieser Gleichung lässt sie sich (textlich) zunächst so darstellen:

$[y \ge X_j]$, wenn gilt :

$[y \ge X_j]$ **oder** $[(y < X_{j+1})$ **und gleichzeitig** (y liegt zwischen $X_j$ und $X_{j+1})]$.

Die zweite eckige Klammer charakterisiert die Wahrscheinlichkeit des gleichzeitigen Eintreffens zweier Ereignisse A und B: P[AB]. Nach dem Multiplikationssatz für Wahrscheinlichkeiten (5) gilt dafür unter Benutzung des Begriffes der bedingten Wahrscheinlichkeit:

$$P[AB] = P[B \mid A] \cdot P[A]$$

Damit erhält man aber gerade den rechten Term in Gleichung (3)! Nach Gleichung (3) erhält man weiter:

$$pe_j = pe_{j+1} + P[X_j \le y < X_{j+1} \mid y < X_{j+1}] \cdot (1 - pe_{j+1}).$$

Die bedingte Wahrscheinlichkeit $P[X_j \le y < X_{j+1} \mid y < X_{j+1}]$ lässt sich mit Hilfe der Anzahlen $A_j$ und $B_j$ wie folgt angeben:

$$P\left[X_j \le y < X_{j+1} \,\middle|\, y < X_{j+1}\right] = \frac{A_j}{A_j + B_j}$$

Für die empirische Wahrscheinlichkeit $pe_j$ für die Überschreitung der Nachweisgrenze $X_j$ erhalten wir also die Rekursionsformel ($pe_{m+1}=0$ angenommen):

$$pe_j = pe_{j+1} + \left(1 - pe_{j+1}\right) \cdot \frac{A_j}{A_j + B_j} \qquad (4)$$

Nach Gleichung (4) ergeben sich für das obige Beispiel folgende Überschreitungswahrscheinlichkeiten $pe_j$, indem bei der höchsten Nachweisgrenze angefangen wird:

$$pe_2 = 0 + (1 - 0) \cdot \frac{6}{6 + 12} = 0{,}3333$$

$$pe_1 = 0{,}3333 + (1 - 0{,}3333) \cdot \frac{3}{3 + 6} = 0{,}5556$$

Einem zwischen $X_j$ und $X_{j+1}$ liegenden Wert y, der den Rang r innerhalb von insgesamt $A_j$ solchen Werten einnimmt, wird dann die folgende zwischen $pe_j$ und $pe_{j+1}$ liegende Wahrscheinlichkeit zugeordnet (es ist $pe_{j+1} < pe_j$ !):

$$pe_j - \left(pe_j - pe_{j+1}\right) \cdot \frac{r}{A_j + 1}$$

Durch die Bildung der komplementären Wahrscheinlichkeit ergeben sich hieraus die empirischen Plot-Wahrscheinlichkeiten $p_i$ für die $A_j$ signifikanten Werte oberhalb der j-ten Nachweisgrenze und unterhalb der (j+1)-ten Nachweisgrenze:

$$p_i = \left(1 - pe_j\right) + \left(pe_j - pe_{j+1}\right) \cdot \frac{r}{A_j + 1} \qquad (5)$$

Hierin ist r der Rang des i-ten signifikanten Wertes aus den $A_j$ Werten zwischen $X_j$ und $X_{j+1}$

Die empirischen Plot-Wahrscheinlichkeiten $pc_i$ für die j-te Nachweisgrenze ergeben sich in ähnlicher Weise zu:

$$pc_i = \left(1 - pe_j\right) \cdot \frac{r}{C_j + 1} \qquad (6)$$

Hierin ist r der Rang der i-ten Nachweisgrenze mit demselben Wert, sofern die j-te Nachweisgrenze mehrfach mit der Multiplizität $C_j$ mit demselben Wert vorkommt. Für eine nur einfach im Datensatz vorkommende j-te Nachweisgrenze ($C_j = 1$) ist $r = i = 1$, d. h. $pc_i = (1 - pe_j)/2$.

Wie unter Abschnitt 2.1 beschrieben, werden nun allen Plot-Wahrscheinlichkeiten $p_i$ bzw. $pc_i$ Quantile der Normalverteilung $z_i = \Phi^{-1}(p_i)$ bzw. $z_i = \Phi^{-1}(pc_i)$ zugeordnet. Für die signifikanten Messwerte werden mit $z_i = \Phi^{-1}(p_i)$ nach Gleichung (1) die Parameter a und b der Regressionsgeraden ermittelt.

Mit $z_i = \Phi^{-1}(pc_i)$ werden dann nach Gleichung (2) die Ersatzwerte qi für die „< G"-Werte bestimmt. Danach liegt mit den signifikanten Werte $y_i$ einerseits und den Ersatzwerten $q_i$ für die „< G"-Werte andererseits wieder ein vollständiger Datensatz vor, mit dem, wie unter Abschnitt 2.1 angedeutet, zur Bestimmung von Mittelwert und Standardabweichung, aber auch vom Medianwert verfahren werden kann.

Die für das obige Beispiel nach den Gleichungen (5) und (6) berechneten Werte $p_i$ bzw. $pc_i$ für die Plot-Wahrscheinlichkeiten der signifikanten Werte bzw. der nicht-signifikanten Werte und die ihnen jeweils zugeordneten Quantile $z_i$ sind in der folgenden Tabelle dargestellt. In Spalte 6 der Tabelle sind die nach Gleichung (1) berechneten Ersatzwerte $q_i$ für die „< G"-Werte angegeben. Spalte 7 enthält den aus den $q_i$ und signifikanten $y_i$ vereinigten "vollständigen" Datensatz. Aus den Daten in Spalte 7 ergeben sich folgende statistische Kennwerte:

**Tab. 2:**

| | | | |
|---|---|---|---|
| Mittelwert: | 10,789 | zu dokumentieren als: | 11 |
| Standardabweichung: | 13,681 | zu dokumentieren als: | 14 |
| Medianwert: | 3,926 | zu dokumentieren als: | 3,9 |

| Sp. 1 | Sp. 2 | Sp. 3 | Sp. 4 | Sp. 5 | Sp. 6 | Sp. 7 |
|---|---|---|---|---|---|---|
| sortierte $y_i$ | ln $y_i$ | $p_i$ (Gl. (5)) | $pc_i$ (Gl. (6)) | $z_i=\Phi^{-1}(p_i)$ bzw. $z_i=\Phi^{-1}(pc_i)$ | $q_i = e^{a+b\cdot z_i}$ [1)] | "vollständige Werte" $y_i$ |
| < 1,0 | – | – | 0,06 | –1,53 | 0,47 | 0,47 |
| < 1,0 | – | – | 0,12 | –1,14 | 0,85 | 0,85 |
| < 1,0 | – | – | 0,19 | –0,87 | 1,28 | 1,28 |
| < 1,0 | – | – | 0,25 | –0,66 | 1,77 | 1,77 |
| < 1,0 | – | – | 0,31 | –0,47 | 2,35 | 2,35 |
| < 1,0 | – | – | 0,38 | –0,3 | 3,04 | 3,04 |
| 3 | 1,1 | 0,5 | – | 0 | – | 3 |
| 7 | 1,95 | 0,55 | – | +0,1394 | – | 7 |
| 9 | 2,2 | 0,61 | – | +0,2818 | – | 9 |
| < 10 | – | – | 0,16 | –0,96 | 1,11 | 1,11 |
| < 10 | – | – | 0,33 | –0,43 | 2,51 | 2,51 |
| < 10 | – | – | 0,5 | 0 | 4,81 | 4,81 |
| 12 | 2,48 | 0,71 | – | +0,5656 | – | 12 |
| 15 | 2,71 | 0,76 | – | +0,7122 | – | 15 |
| 20 | 3 | 0,81 | – | +0,8760 | – | 20 |
| 27 | 3,3 | 0,85 | – | +1,0676 | – | 27 |
| 33 | 3,5 | 0,9 | – | +1,3094 | – | 33 |
| 50 | 3,91 | 0,95 | – | +1,6688 | – | 50 |

[1)] Parameter aus linearer Regression (Sp. 2 und 5) nach Gleichung (1): a=1,5702, b=1,5108

Die nachfolgende Abbildung zeigt den dazugehörigen Wahrscheinlichkeitsplot.

**Abb. 2 : Logarithmischer Wahrscheinlichkeitsplot**

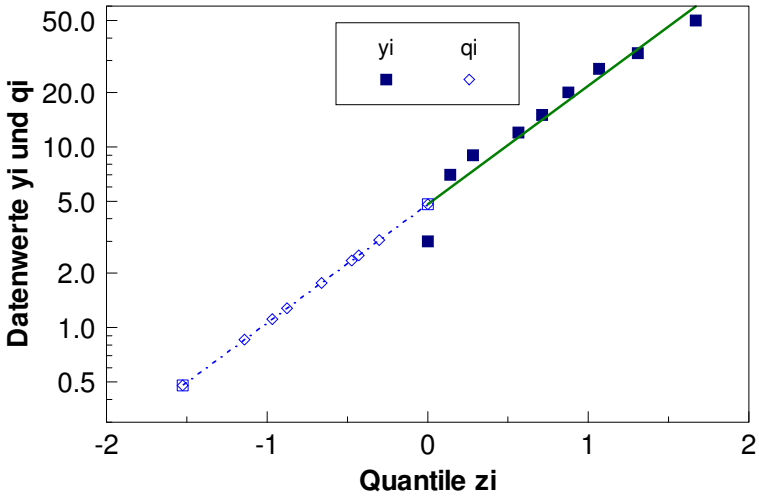


**Abb. 2:** Logarithmischer Wahrscheinlichkeitsplot

# 3 Ergänzungen zum Verfahren nach Helsel und Cohn

Zur weiteren Verbesserung des Verfahrens wurden die folgenden Ergänzungen hinzugefügt.

— Treten nach der am Beginn durchzuführenden Sortierung der originalen Daten j verschieden große „< G"-Werte hintereinander auf, ohne dass ein signifikanter Wert dazwischen liegt, werden für die folgenden Berechnungen alle diese „< G"-Werte durch ihren größten „< $G_{max}$"-Wert ersetzt, so dass j „< $G_{max}$"-Werte vorliegen.

— In Gleichung (1) stellt der Regressionsparameter b die empirische Standardabweichung der (n-k) Werte ln $y_i$ dar. Diese ist, im Gegensatz zur Varianz, für kleinere Werte von (n-k) mit einem systematischen Fehler (Bias, b wird unterschätzt) behaftet. Für die Standardabweichung b werden daher multiplikative, von (n-k) abhängige Korrektionsfaktoren aus der Literatur verwendet (6).

— Aus Spalte 7 der Tabelle 2 ist ersichtlich, dass einige der Ersatzwerte für die < 1,0-Werte tatsächlich größer als die ihnen zugeordnete Nachweisgrenze sind, was aber nicht sein sollte. Abhilfe kann hier dadurch erreicht werden, dass zunächst die Steigung der Regressionsgeraden (vgl. Abbildung 2) geändert wird. Da sich die Gerade hierbei zu weit von den signifikanten Messwerten entfernen kann, wird sie danach durch Parallelverschiebung wieder auf die Kurve der signifikanten Messwerte gebracht, bis die Ersatzwerte aus der verlängerten Geraden auch unter ihren zugeordneten Nachweisgrenzen liegen. Dieses (Optimierungs-)Verfahren lässt sich nur noch mit Hilfe eines Rechenprogrammes durchführen. Nach Anwendung dieser Korrektur erhält man folgende Kennwerte für die Daten aus Tabelle 2:

| | |
|---|---|
| Mittelwert: | 10,082; |
| Standardabweichung: | 14,149; |
| Medianwert: | 2,588. |

Diese Korrektur ist in der Regel nur bei „extremen" Datensätzen erforderlich.

**Literatur**

(1) Helsel, D. R., Cohn, T. A.: Estimation of Descriptive Statistics for Multiply Censored Water Quality Data. Water Resources Research, 1988, Vol. 24, S. 1997-2004

(2) Helsel, D. R.: Less than obvious. Environ. Sci. Technol., 1990, Vol. 24, S. 1766-1774

(3) Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit (Hrsg.): Richtlinie zur Emissions- und Immissionsüberwachung kerntechnischer Anlagen. GMBl. 44, Nr. 29 vom 19. August 1993

(4) Hirsch, R. M., Stedinger, J. R.: Plotting Positions for Historical Floods and Their Precision. Water Resources Research 23, 1987, S. 715-727

(5) Kreyszig, E.: Statistische Methoden und ihre Anwendungen. Vandenhoeck & Ruprecht, Göttingen, 1975

(6) Johnson, N., Kotz, S.: Continuous univariate distributions. Part 1. Verlag Houghton and Mifflin, Boston,1970