

**ANPASSUNGSTESTS
FÜR SURVIVALVERTEILUNGEN**

Karl AUINGER

Forschungsbericht/
Research Memorandum Nr. 253

Januar 1989

All contributions are to be regarded as preliminary and should not be quoted without consent of the respective author. All contributions are personal and any opinions expressed should never be regarded as opinion of the Institute for Advanced Studies.

Abstract. In this paper we propose a general method for the construction of goodness of fit tests for lifetime distributions. The method is the following: first find an identity which holds for the survival function or the cumulative hazard function of the null distribution. Then replace the function by a consistent estimate. The resulting statistic is asymptotically normal. Estimating its asymptotic variance then gives a teststatistic which is under H_0 asymptotically χ^2 . The method can be used for randomly right censored and single type I censored data. We apply this method to the following distributions: Weibull, Log-logistic, Log-normal, Rayleigh, Gompertz, Pareto.

Zusammenfassung. In dieser Arbeit wird eine allgemeine Methode für die Konstruktion von Anpassungstests für Überlebensverteilungen vorgeschlagen. Die Methode beruht auf einer zu findenden Identität, der die Überlebensfunktion oder die kumulierte Hazardfunktion der Null-Verteilung genügt. Wird in der Identität die jeweilige Funktion durch eine konsistente Schätzung ersetzt, dann ist die daraus resultierende Statistik asymptotisch normalverteilt, woraus durch Schätzung der asymptotischen Varianz eine unter der Nullhypothese asymptotisch χ^2 -verteilte Teststatistik gewonnen wird. Diese Methode ist für zufällig- und single-typ-I- rechtszensierte Daten geeignet. Auf die folgenden Verteilungen wird diese Methode angewandt: Weibull, Log-logistisch, Log-normal, Rayleigh, Gompertz, Pareto.

Inhaltsverzeichnis

I. Einleitung: Grundbegriffe der Survivalverteilungen.....	1
II. Anpassungstests für Survivalverteilungen.....	6
1. Zufällig zensierte Daten.....	6
2. Schätzungen von Funktionsverläufen, die die zugrundeliegende Verteilung eindeutig bestimmen.....	7
a. Der Produkt-Limit-Schätzer.....	7
b. Die empirische kummulierte Hazardfunktion.....	8
3. Grenzverteilungen.....	9
a. Die asymptotische Verteilung der Prozesse.....	9
b. Die endlichdimensionalen Verteilungen.....	10
c. Die Schätzung der Kovarianzen.....	12
4. Typ-I-Zensierung.....	13
a. Typ-I-Zensierung als Grenzfall der zufälligen Zensierung....	13
b. Modifikation der Statistiken.....	15
5. Konstruktion von Anpassungstests.....	17
III. Anwendungen.....	19
1. Die Weibull Verteilung.....	19
2. Die Log-logistische Verteilung.....	23
3. Die Log-normal Verteilung.....	26
4. Die Rayleigh Verteilung.....	29
5. Die Gompertz Verteilung.....	31
6. Die Pareto Verteilung.....	32
7. Zusammenfassung.....	36
Literatur.....	37

I. Einleitung

Es sollen hier kurz die mathematisch-statistischen Grundlagen der Survivalanalyse skizziert werden.

Survivalanalyse untersucht - ganz allgemein gesprochen - den Übergang von Individuen von einem Stadium (Zustand) in ein anderes, die Dauer des Verweilens in den einzelnen Stadien und die Wahrscheinlichkeit eines Überganges. Wir wollen uns auf den Spezialfall des sogenannten Zwei-Zustände-Modells mit absorbierendem Zielzustand beschränken. Hier befinden sich die interessierenden Objekte zu Beginn der Untersuchung in einem Ausgangszustand, um im Lauf der Zeit in einen anderen Zustand überzugehen, der vom Standpunkt der jeweiligen Untersuchung aus als endgültiger ("Tod") aufgefaßt wird.

Der Zeitpunkt T des Eintritts des Ereignisses - auch Ankunftszeit oder Überlebenszeit genannt - wird als nichtnegative zufällige Veränderliche aufgefaßt. Ganz allgemein besteht die Aufgabe der Survivalanalyse darin, aus dem gegebenen Datenmaterial Aufschlüsse über die Wahrscheinlichkeitsverteilung der zufälligen Variablen T zu gewinnen.

Angenommen, diese Wahrscheinlichkeitsverteilung besitzt eine Dichtefunktion f , dann ist

$$F(t) = P(T \leq t) = \int_0^t f(s) ds$$

die (kummulierte) Verteilungsfunktion der Überlebenszeiten. $F(t)$ gibt für ein Individuum die Wahrscheinlichkeit dafür an, daß sein "Tod" innerhalb des Zeitintervalls $[0, t]$ eintritt. Umgekehrt gibt

die "Überlebensfunktion" S :

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(s) ds$$

die Wahrscheinlichkeit an, mit der ein Individuum den Zeitpunkt t "überlebt", d.h. das besagte Ereignis bis zum Zeitpunkt t noch nicht eingetreten ist.

Die Hazardfunktion $h(t)$ ist definiert als

$$h(t) := \lim_{h \rightarrow 0+} q(t, t+h)/h = \lim_{h \rightarrow 0+} P(t < T \leq t+h | T > t)/h$$

wobei $q(t, t+h)$ die Übergangswahrscheinlichkeit: die Wahrscheinlichkeit des Eintritts des Ereignisses im Intervall $(t, t+h]$, unter der Bedingung, daß es bis zum Zeitpunkt t noch nicht eingetreten ist, ausdrückt. $h(t)$ heißt dann die Hazardrate oder Übergangsrate zum Zeitpunkt t . (Bei Mehrzustandsmodellen sind die Begriffe Hazardrate und Übergangsrate nicht identisch.) Wir wollen voraussetzen, daß die Hazardrate zu jedem Zeitpunkt $t > 0$ definiert ist. Die Hazardrate ist keine Wahrscheinlichkeit (es können Werte > 1 vorkommen). Wohl aber sind für "sehr kleine" Zeitintervalle der Länge h die Hazardrate und die Übergangswahrscheinlichkeit ungefähr proportional:

$$q(t, t+h) \approx h(t) \cdot h$$

Anschaulich heißt dies: je höher die Hazardrate zu einem Zeitpunkt t ist, desto größer ist für ein Individuum das Risiko, zu diesem Zeitpunkt t zu "sterben"; damit ist in der Hazardrate quasi die Intensität des Übergangsprozesses ausgedrückt. In der Hazard- oder Ratenfunktion $h(t)$ ist festgehalten, wie sich diese "momentane Todesrate" während der Zeit ändert.

Es gelten folgende Beziehungen:

$$h(t) = f(t)/S(t) = -(dS(t)/dt)/S(t) = -d(\log S(t))/dt$$

$$S(t) = \exp\left(-\int_0^t h(s)ds\right)$$

$$f(t) = h(t) \cdot \exp\left(-\int_0^t h(s)ds\right)$$

Die Funktion $H(t) = \int_0^t h(s)ds = -\log S(t)$ wird auch als die kumulierte Hazardfunktion bezeichnet. Durch jede der drei Funktionen $h(t)$, $H(t)$ und $S(t)$ ist die Wahrscheinlichkeitsverteilung der Ankunftszeit T eindeutig bestimmt.

In dieser Arbeit wollen wir Anpassungstests für die folgenden Survival Verteilungen konstruieren:

Weibull, Log-logistisch, Log-normal, Rayleigh, Gompertz und Pareto.

In allen Fällen wird nach dem folgenden Prinzip vorgegangen: man finde eine funktionale Beziehung $g(x_1, \dots, x_n)$ sodaß unter der Nullhypothese jeweils $g(S(t_1), \dots, S(t_n)) = 0$ resp.

$g(H(t_1), \dots, H(t_n)) = 0$ für (gewisse oder beliebige) Werte $t_1, \dots, t_n$

gilt. Werden nun die Werte $S(t_i)$ resp. $H(t_i)$ durch konsistente Schätzungen ersetzt, dann sollte unter der Nullhypothese

$g(\hat{S}(t_1), \dots, \hat{S}(t_n))^2$ resp. $g(\bar{H}(t_1), \dots, \bar{H}(t_n))^2$ nicht zu

groß werden. Als Maß für die zulässige Abweichung wird die

asymptotische Verteilung von $\sqrt{Ng}(\hat{S}(t_1), \dots, \hat{S}(t_n))$ resp.

$\sqrt{Ng}(\bar{H}(t_1), \dots, \bar{H}(t_n))$ betrachtet. Diese Statistiken sind in den

untersuchten Fällen asymptotisch normalverteilt, sodaß mittels

Bestimmung einer konsistenten Schätzung der asymptotischen Ko-

varianzmatrix Teststatistiken angegeben werden können, die unter

der Nullhypothese asymptotisch χ^2 -verteilt sind.

Außerdem wird diskutiert, gegen welche Klassen von Verteilungs-

alternativen diese Tests konsistent sind, und wie diese Klassen vergrößert werden können.

Geeignet sind die Tests für zufällig zensierte (unter gewissen Annahmen) und *Typ-I* zensierte Daten. Für den letzteren Fall ergibt sich meist eine Vereinfachung der Teststatistiken.

Anwendungen dieser Methode auf die Exponentialverteilung wurden ausführlich vom Autor (1988) studiert.

Abschließend wollen wir noch zwei Theoreme aus der mathematischen Statistik zitieren, die grundlegend für unsere Untersuchung sind:

Satz 1. (Zentraler Grenzwertsatz)

Sei (Y_i) eine Folge von unabhängigen, identischverteilten (n -dimensionalen) zufälligen Vektoren mit Erwartungswert $EY_i = m$ und Kovarianzmatrix $E(Y_i - m)(Y_i - m)' = T$. Dann konvergiert die Variable

$$\sqrt{N(N-1)} \sum_{i=1}^N (Y_i - m)$$

in Verteilung gegen eine multivariate Normalverteilung $N(0, T)$ mit Mittelwert 0 und Kovarianzmatrix T .

Beweis: Z.B.: Anderson (1958), p.74.

Satz 2. (δ -Methode)

Seien $T_{1N}, \dots, T_{kN}$ Statistiken derart, daß die Statistik

$$\sqrt{N}(T_{1N} - m_1, \dots, T_{kN} - m_k)$$

für $N \rightarrow \infty$ in Verteilung gegen eine multivariate Normalverteilung $N(0, T)$ konvergiert. Sei $g: \mathbb{R}^k \rightarrow \mathbb{R}^n$ eine Abbildung, deren partielle Ableitungen an der Stelle $m = (m_1, \dots, m_k)$ existieren. Dann konvergiert

$$\sqrt{N}(g(T_{1N}, \dots, T_{kN}) - g(m_1, \dots, m_k))$$

in Verteilung gegen eine multivariate Normalverteilung $N(0, V)$, wobei $V = D'TD$ und D die Matrix der partiellen Ableitungen von g an der Stelle m bezeichnet.

Beweis: Z.B.: Rao (1965), Ch.6.

II. Anpassungstests für Survivalverteilungen

1. Zufällig zensierte Daten

Von zufällig zensierten Daten spricht man, wenn einige der Realisierungen t_i der zu untersuchenden Variablen T nicht exakt bekannt, sondern nur obere oder untere Schranken derselben gegeben sind, die selbst vom Zufall abhängen. Da es sich bei den t_i in den meisten Anwendungen um Zeiten handelt, ist nur der Fall interessant, daß sogenannte Rechtszensierung vorliegt, d.h. daß anstatt der exakten Realisierungen untere Schranken derselben bekannt sind. Diese können als *Beobachtungsdauer* des entsprechenden Individuums interpretiert werden.

Formalisiert läßt sich das Problem der *zufällig zensierten Daten* folgendermaßen darstellen: Seien (T_i, C_i) $i = 1, \dots, n$, n unabhängige, identisch verteilte Paare von Zufallsvariablen, deren Verteilungsfunktionen mit $F_1(t) = P(T_i \leq t)$ und $F_2(t) = P(C_i \leq t)$ bezeichnet werden. Das Problem besteht darin, diese Verteilungsfunktionen zu schätzen (zumindest die Verteilungsfunktion der Variablen T_i - das Problem ist aber symmetrisch in T_i und C_i), obwohl nicht Realisierungen (t_i, c_i) die Paare (T_i, C_i) beobachtet werden, sondern nur Realisierungen der Paare (X_i, k_i) , wobei diese Variablen definiert sind durch:

$$X_i := \min(T_i, C_i)$$

und

$$k_i := I\{T_i \leq C_i\}$$

Dabei bedeutet $I\{x\}$ die Indikatorfunktion der Menge $\{x\}$. Obwohl das Problem symmetrisch in T_i und C_i ist, bezieht man sich (vom Ausgangspunkt her) auf T_i als die interessierende Variable und auf C_i als die Zensierungsvariable. Die Indikatorvariable k_i zeigt an,

ob es sich bei der gemachten Beobachtung um eine tatsächliche oder eine zensierte handelt.

Im folgenden soll $S(t)$ die Überlebensfunktion der Variablen T_i , $G(t)$ die Überlebensfunktion der Zensierungsvariablen C_i bezeichnen. Außerdem wollen wir für die ganze Arbeit über den Zensierungsmechanismus die folgenden einschränkenden Annahmen treffen:

- (i) die interessierende Variable $T_i \geq 0$ und die Zensierungsvariable $C_i \geq 0$ haben stetige Verteilungen
- (ii) T_i und C_i seien unabhängig.

2. Schätzungen von Funktionsverläufen, die die zugrundeliegende Verteilung eindeutig bestimmen

2.a. Der Produkt-Limit-Schätzer

Der Produkt-Limit-Schätzer oder Kaplan-Meier-Schätzer (PLS) für die Überlebensfunktion $S(t) = 1 - F_1(t)$ ist folgendermaßen definiert: Bei gegebenem Sample $\{(X_1, k_1), \dots, (X_N, k_N)\}$ seien $t_{(1)} < t_{(2)} < \dots < t_{(k)}$ die "tatsächlich beobachteten" Überlebenszeiten, d.h. jene X_i , für die $k_i = 1$ gilt, der Größe nach geordnet. Es kann angenommen werden, daß zu einem beliebigen Zeitpunkt höchstens eine exakte oder zensierte Beobachtung gemacht wurde, da für $i \neq j$ $P(X_i = X_j) = 0$ gilt. Weiters sei n_i ($1 \leq i \leq k$) die Mächtigkeit der Risikomenge zum Zeitpunkt $t_{(i)}$, d.h. die Anzahl der Individuen, die unmittelbar vor dem Zeitpunkt $t_{(i)}$ am Leben und unzensiert sind, oder in der obigen Terminologie:

$$n_i := |\{j | X_j \geq t_{(i)}\}|$$

Der Produkt-Limit-Schätzer ist dann definiert durch

$$\hat{S}(t) = \prod_{i: t_{(i)} < t} \frac{n_i - 1}{n_i}$$

Außerdem ist $\hat{S}(t)$ nicht definiert, wenn $t > X_{(N)}$ und $k_{(N)} = 0$.

Anschaulich kann man den PLS für $S(t)$ folgendermaßen interpretieren: die Wahrscheinlichkeit, den Zeitpunkt t zu überleben, ist das Produkt der bedingten Wahrscheinlichkeiten, $t_{(i)}$ zu überleben, unter der Bedingung $t_{(i-1)}$ überlebt zu haben:

$$S(t) = P(T_i > t) = P(T_i > t | T_i > t_{(u)}) P(T_i > t_{(u)} | T_i > t_{(u-1)}) \dots P(T_i > t_{(1)}),$$

wobei $t_{(u+1)} > t \geq t_{(u)}$. Die Ausdrücke $(n_i - 1)/n_i$ sind dann gerade Schätzwerte für diese bedingten Wahrscheinlichkeiten.

Bemerkung. Üblicherweise wird der PLS etwas allgemeiner definiert.

Im obigen Produkt wird der Ausdruck $n_i - 1$ ersetzt durch $n_i - d_i$, wobei d_i die Anzahl der Individuen, die im Zeitpunkt $t_{(i)}$ sterben, bezeichnet. Oder formal ausgedrückt: $d_i = |\{j | X_j = t_{(i)}, k_i = 1\}|$. Da die Variable T eine stetige Verteilung besitzt, ist $P(d_i > 1) = 0$. In diesem Fall reduziert sich die allgemeine Definition also auf die gegebene. Der Terminus "Risikomenge" wird oft für die Menge der Individuen ohne die zum jeweiligen Zeitpunkt Verstorbenen verwendet; deren Mächtigkeit ist dann $n_i - d_i$ statt n_i .

2.b. Die empirische kummulierte Hazardfunktion

Will man die kummulierte Hazardfunktion schätzen, bietet sich zunächst die Beziehung $H(t) = -\log S(t)$ an. Setzt man für $S(t)$ den PLS ein, dann erhält man folgende Schätzung für die kummulierte Hazardfunktion:

$$\hat{H}(t) = -\log \hat{S}(t).$$

Eine andere Möglichkeit bietet die sogenannte *empirische kummulierte Hazardfunktion*:

$$\tilde{H}(t) := \sum_{t_{(j)} < t} \frac{1}{n_j} = \sum_{X_{(i)} < t} \frac{k_{(i)}}{N+1-i}$$

Es gilt $\hat{H}(t) = - \sum \log(1 - 1/n_j) = \sum (1/n_j + 1/2n_j^2 + \dots)$. Damit ist $\tilde{H}(t)$ eine Approximation erster Ordnung von $\hat{H}(t)$, die recht brauchbar ist, wenn t nicht allzu groß ist. Asymptotisch sind die beiden Schätzungen äquivalent. Es gilt der

Satz 3. (Breslow & Crowley, 1974)

$$\text{plim } (\hat{H}(t) - \tilde{H}(t)) = 0.$$

Für einige Anwendungen erscheint die Schätzung $\tilde{H}(t)$ aufgrund der rechnerischen Einfachheit empfehlenswerter, obwohl theoretisch alle Tests in Abschnitt III mit der Schätzung $\hat{H}(t)$ durchgeführt werden könnten.

Bemerkung. Ähnlich wie beim PLS wird auch die empirische kumulierte Hazardfunktion üblicherweise etwas allgemeiner definiert: die 1-en in der ersten Summe werden durch d_i ersetzt. Unter der Voraussetzungen der Stetigkeit der Verteilung von T sind die beiden Definitionen fast sicher identisch.

3. Grenzverteilungen

a. Die Verteilung des Grenzprozesses

Das asymptotische Verhalten der Prozesse $\hat{S}(t)$ und $\tilde{H}(t)$ wurde von Breslow und Crowley (1974) unter den genannten Voraussetzungen für den Zensierungsmechanismus untersucht. Wir brauchen folgende Notation: $F(t) = P(X_i \leq t, k_i = 1)$ bezeichne die Subverteilungsfunktion der nicht zensierten Beobachtungen. Weiters sei $R(t)$ die Überlebensfunktion der Variablen $X_i = \min(T_i, C_i)$, d.h. $R(t) = S(t)G(t) = P(\min(T_i, C_i) > t)$. Unter den eingangs gemachten Voraussetzungen über den Zensierungsprozeß gilt folgender

Satz 4. Sei $K > 0$ eine Konstante mit $R(K) > 0$. Dann konvergiert für $0 < t < K$ der Prozeß $\sqrt{N}(\hat{S}(t) - S(t))$ schwach gegen einen Gauß'schen Prozeß $Z(t)$ mit Mittelwert 0 und Kovarianzfunktion

$$\text{Cov}(Z(v), Z(t)) = S(v)S(t) \int_0^v \frac{-dS(u)}{S(u)R(u)} = S(v)S(t) \int_0^v \frac{dF(u)}{R^2(u)} \quad (v \leq t)$$

A fortiori ist dann $\hat{S}(t)$ eine konsistente Schätzung von $S(t)$.

Bemerkung. Die Einschränkung auf das Intervall $[0, K]$ besagt anschaulich nur, daß ab einem Zeitpunkt, ab dem mit Wahrscheinlichkeit 1 keine Beobachtungen (weder zensierte noch unzensierte) mehr gemacht werden, keine "vernünftige" Schätzung (und daher auch Testung) für den Verlauf der Überlebensfunktion gemacht werden kann: ist $R(K) = S(K)G(K) = 0$, dann ist $\hat{S}(t) = 0$ oder $\hat{S}(t)$ ist nicht definiert für $t \geq K$.

Für die empirische kummulierte Hazardfunktion $\tilde{H}(t)$ gilt ein ähnliches Resultat:

Satz 5. Sei $K > 0$ mit $R(K) > 0$. Dann konvergiert für $0 < t < K$ der Prozeß $\sqrt{N}(\tilde{H}(t) - H(t))$ schwach gegen einen Gauß'schen Prozeß $Z(t)$ mit Mittelwert 0 und Kovarianzfunktion

$$\text{Cov}(Z(s), Z(t)) = \int_0^s \frac{-dS(u)}{S(u)R(u)} = \int_0^s \frac{dF(u)}{R^2(u)} \quad (s \leq t)$$

b. Endlichdimensionale Verteilungen

Für die Konstruktion von Anpassungstests werden wir vor allem zwei Resultate über die endlichdimensionalen Verteilungen des Prozesses

brauchen, die nach einem bekannten Theorem aus der Theorie der stochastischen Prozesse (siehe Billingsley (1968) oder Hall und Heyde (1980)) aus den beiden zitierten Sätzen abgeleitet werden können (wieder wird angenommen, daß die Voraussetzungen über den Zensierungsprozeß erfüllt sind):

Satz 6. Sei $K = \inf\{t | S(t)G(t) = 0\}$ und $0 < t_1 < t_2 < \dots < t_n < K$. Dann ist die Statistik $\sqrt{N}(\hat{S}(t_1) - S(t_1), \dots, \hat{S}(t_n) - S(t_n))'$ asymptotisch normalverteilt mit Mittelwert 0 und Kovarianzmatrix $T = (t_{ij})$, wobei t_{ij} gegeben ist durch:

$$t_{ij} = S(t_i)S(t_j) \int_0^{t_i} \frac{-dS(u)}{S(u)R(u)} = S(t_i)S(t_j) \int_0^{t_i} \frac{dF(u)}{R^2(u)} \quad (i \leq j)$$

Für die endlichdimensionalen Verteilungen der empirischen kumulierten Hazardfunktion gilt das analoge Resultat:

Satz 7. Sei $K = \inf\{t | S(t)G(t) = 0\}$ und $0 < t_1 < t_2 \dots t_n < K$. Dann ist die Statistik $\sqrt{N}(\bar{H}(t_1) - H(t_1), \dots, \bar{H}(t_n) - H(t_n))'$ asymptotisch normalverteilt mit Mittelwert 0 und Kovarianzmatrix $U = (u_{ij})$, wobei u_{ij} gegeben ist durch

$$u_{ij} = \int_0^{t_i} \frac{-dS(u)}{S(u)R(u)} = \int_0^{t_i} \frac{dF(u)}{R^2(u)} \quad (i \leq j)$$

Für die Invertierbarkeit dieser beiden Kovarianzmatrizen wurde vom Autor (1988) folgendes gezeigt:

Satz 8. Die Kovarianzmatrix T (resp. V) ist genau dann invertierbar, wenn $0 < S(t_n) < S(t_{n-1}) < \dots < S(t_1)$ gilt.

c. Schätzung der Kovarianzen

Die Eintragungen der Kovarianzmatrizen T (resp. U):

$t_{ij} = S(t_i)S(t_j)u_{ij}$ sind i.a. unbekannt. Für viele Anwendungen werden konsistente Schätzungen benötigt. Aus obiger Beziehung folgt, daß es genügt, konsistente Schätzungen für U zu finden. Eine Schätzung für T ist dann gegeben durch

$\hat{t}_{ij} = \hat{S}(t_i)\hat{S}(t_j)\hat{u}_{ij}$. Um

$$u_{ij} = \int_0^{t_i} \frac{-dS(u)}{S(u)R(u)} = \int_0^{t_i} \frac{dF(u)}{R^2(u)} \quad (i \leq j)$$

konsistent zu schätzen, ist es naheliegend, die unter dem Integralzeichen vorkommenden Funktionen S, R und F durch ihre empirischen Analoga $\hat{S}, \hat{R}$ und $\hat{F}$ zu ersetzen. Für $\hat{S}$ wählt man den Produkt-Limit-Schätzer. $R(t)$ bezeichnet die Wahrscheinlichkeit, nach dem Zeitpunkt t eine (zensierte oder unzensierte) Beobachtung zu machen, d.h. $R(t) = P(X_i = \min(T_i, C_i) > t)$. Eine natürliche Schätzung dafür ist gegeben durch $\hat{R}(t) = n_t/N$, wobei n_t die Größe der Risikomenge zum Zeitpunkt t bezeichnet. Formal ausgedrückt: $n_t = |\{j | X_j \geq t\}|$. Für $F(t)$, die Subverteilungsfunktion der nicht zensierten Beobachtungen: $F(t) = P(X_i \leq t, k_i = 1)$ wählt man deren empirische Verteilungsfunktion:

$$\hat{F}(t) = N^{-1} \sum_{X(j) < t} k(j) = N^{-1} \sum_{t(j) < t} 1 = |\{j | t(j) < t\}|/N$$

Bildet man nun die Ausdrücke $\int -d\hat{S}(t)/(\hat{S}(t)\hat{R}(t))$ respektive $\int d\hat{F}(t)/\hat{R}^2(t)$, dann sind die jeweiligen Integrale als

Riemannintegrale nicht definiert, da Integrand und Integrator an den gleichen Stellen, nämlich $t(j)$, Sprungstellen besitzen.

Nichtsdestoweniger können sie aber als Integrale bezüglich der diskreten Wahrscheinlichkeitsmaße $-d\hat{S}(t)$ respektive $d\hat{F}(t)$

interpretiert werden. In diesem Fall ist der Wert dieser Integrale

davon abhängig, ob als Werte der Integranden in den Sprungstellen jeweils die linksseitigen oder rechtsseitigen Grenzwerte gewählt werden (d.h. ob in den Definitionen von $\hat{S}(t)$, $\hat{R}(t)$ und $\hat{F}(t)$ "<" oder "<=" gesetzt wird). Variiert man über alle diese Möglichkeiten, dann erhält man folgende Schätzungen für u_{ij} ($i \leq j$):

$$\hat{u}_{ij}(1) = N \sum_{t_{(1)} < t_i} 1/n_1^2$$

$$\hat{u}_{ij}(2) = N \sum_{t_{(1)} < t_i} 1/n_1(n_1-1)$$

$$\hat{u}_{ij}(3) = N \sum_{t_{(1)} < t_i} 1/(n_1-1)^2$$

Offensichtlich gilt $\hat{U}(1) < \hat{U}(2) < \hat{U}(3)$. Unter den angeführten Annahmen über den Zensierungsprozeß wurde vom Autor (1988) gezeigt:

Satz 9. Mit wachsender Stichprobengröße N konvergiert $\hat{U}(i)$ fast sicher gegen U für $i = 1, 2, 3$.

Im Buch von Lawless (1982) wird die "in der Mitte liegende" Schätzung $\hat{U}(2)$ empfohlen.

4. Typ-I-Zensierung

a. Typ-I-Zensierung als Grenzfall der zufälligen Zensierung

Anschaulich bedeutet *Single*-Typ-I-Zensierung folgendes: es wird im voraus die Dauer der Beobachtung festgelegt, die Anzahl der tatsächlichen Beobachtungen ist dann eine zufällige Variable. Formal ausgedrückt: sind $T_1, \dots, T_N$ die Realisierungen der interessierenden Variablen und ist C eine festgelegte Konstante, dann werden

Formal sieht ein derartiger Zensierungsmechanismus wie der unter II.1 erwähnte aus. Es können die Ergebnisse, die bisher für zufällige Zensierungsmuster abgeleitet wurden, jedoch nicht unmittelbar auf den Fall der Typ-I-Zensierung übertragen werden, da die Annahme, daß C eine stetige Verteilung besitzt, nun verletzt ist. Da C eine Konstante ist, gilt

$$G(t) = P(C > t) = \begin{cases} 0 & \text{für } t \geq C \\ 1 & \text{für } t < C \end{cases}$$

Sei $\delta > 0$ beliebig ($\delta < C$). Wir betrachten nun folgende Variable:

$$X_i' = \min(T_i, C_i') \text{ und } k_i' = I\{T_i \leq C_i'\},$$

wobei die C_i' i.i.d. gleichverteilt auf dem Intervall $(C-\delta, C)$ und C_i' und T_i voneinander unabhängig sind. Die Zensierungsvariablen C_i' erfüllen nun alle gemachten Voraussetzungen, und die Sätze über die Grenzverteilungen von $\hat{S}(t)$ und $\tilde{H}(t)$ und deren endlichdimensionale Verteilungen können angewandt werden. Die Verteilungen von (X_i, k_i) und (X_i', k_i') sind im Intervall $(0, C-\delta)$ identisch, daher gelten die Sätze über die Grenzverteilungen von $\hat{S}(t)$ und $\tilde{H}(t)$ und deren endlichdimensionale Verteilungen auch für Single-Typ-I-Zensierung, sofern nur $K < C$ gewählt wird. Für die Grenzverteilung von $\sqrt{N}(\hat{S}(t_1) - S(t_1), \dots, \hat{S}(t_n) - S(t_n))'$ kann dies direkt aus dem zentralen Grenzwertsatz abgeleitet werden (siehe auch den nächsten Abschnitt). Die asymptotische Normalverteilung von $\sqrt{N}(\tilde{H}(t_1) - H(t_1), \dots, \tilde{H}(t_n) - H(t_n))'$ wird direkt im nächsten Abschnitt gezeigt werden.

Insgesamt gelten also alle abgeleiteten Aussagen, insbesondere die in III. konstruierten Tests für zufällig und Typ-I-zensierte Daten (und daher auch für nicht zensierte Daten).

b. Modifikationen

Werden im Intervall $(0, K)$ nur nicht zensierte Beobachtungen gemacht, dann lassen sich die Schätzformeln folgendermaßen modifizieren: $n_i = |\{j | X_j \geq t_{(i)}\}| = |\{i, \dots, N\}| = N - (i-1) = N - i + 1$.

Daher ist:

$\hat{S}(t) = \pi(n_i - 1) / n_i = (N-1) / N * (N-2) / (N-1) * \dots * (N-K) / (N-K+1) =$
 $= (N-K) / N = 1 - (\text{\#Beobachtungen bis zum Zeitpunkt } t) / N =$
 $= 1 - \hat{F}_1(t)$, und dies ist genau die empirische Überlebensfunktion. Die Formel für die Grenzverteilung von

$\sqrt{N}(\hat{S}(t_1) - S(t_1), \dots, \hat{S}(t_n) - S(t_n))'$ läßt sich dann unmittelbar aus dem zentralen Grenzwertsatz ableiten (siehe Auinger (1988)).

Nun zur empirischen kummulierten Hazardfunktion:

$\bar{H}(t) = \sum 1/n_i = 1/N + 1/(N-1) + \dots + 1/(N-K+1)$, wobei K wieder die Anzahl der Beobachtungen bis zum Zeitpunkt t bedeutet. Wir zeigen nun direkt:

Satz 10. Seien $0 < t_1 < t_2 < \dots < t_n < C$ und $S(t_n) > 0$, dann ist $\sqrt{N}(\bar{H}(t_1) - H(t_1), \dots, \bar{H}(t_n) - H(t_n))'$ asymptotisch normalverteilt mit Mittelwert 0 und Kovarianzmatrix $U = (u_{ij})$ mit

$$u_{ij} = \frac{1 - S(t_i)}{S(t_i)} = \frac{t_{ij}}{S(t_i)S(t_j)} \quad (i \leq j)$$

Dabei ist $(t_{ij}) = (S(t_j)(1 - S(t_i)))$ die Kovarianzmatrix der Grenzverteilung von $\sqrt{N}(\hat{S}(t_1) - S(t_1), \dots, \hat{S}(t_n) - S(t_n))'$.

Bemerkung. Letzteres kann in Auinger (1988) nachgelesen werden. Die Behauptung des Satzes ist in Übereinstimmung mit der vorher

berechneten Formel für die (Ko)varianz $u_{ij} = \int_0^{t_i} \frac{-dS(u)}{S(u)R(u)}$. Da $R(u) =$

$$= S(u)G(u) = S(u), \text{ ist } u_{ij} = \int_0^{t_i} \frac{-dS(u)}{S^2(u)} = \frac{1}{S(u)} \Big|_0^{t_i} = \frac{1}{S(t_i)} - 1 = \frac{1-S(t_i)}{S(t_i)}.$$

Beweis. Sei K_i die Anzahl der Beobachtungen bis zum Zeitpunkt t_i .

Dann ist $\bar{H}(t_i) = 1/N + 1/(N-1) + \dots + 1/(N-K_i+1)$. Es gilt

$0 < \log(N+1) - \log(N-K_i+1) < \bar{H}(t_i) < \log(N) - \log(N-K_i)$; daher

ist $0 < \log(N/(N-K_i)) - \bar{H}(t_i) < \log[N(N-K_i+1)/\{(N-K_i)(N+1)\}]$.

Sei nun $g: \mathbb{R}^n \rightarrow \mathbb{R}^n$ gegeben durch $g = (g_1, \dots, g_n)'$ und

$g_i(x_1, \dots, x_n) = -\log(x_i)$. Dann ist $g_i(S(t_1), \dots, S(t_n)) =$

$-\log S(t_i) = \log 1/S(t_i) = H(t_i)$ und daher

$g(S(t_1), \dots, S(t_n)) = (H(t_1), \dots, H(t_n))'$,

$g_i(\hat{S}(t_1), \dots, \hat{S}(t_n)) = -\log \hat{S}(t_i) = \log N/(N-K_i)$ und

$g(\hat{S}(t_1), \dots, \hat{S}(t_n)) = (\log N/(N-K_1), \dots, \log N/(N-K_n))'$.

Daher ist

$$\begin{aligned} & \sqrt{N}(\bar{H}(t_1) - H(t_1), \dots, \bar{H}(t_n) - H(t_n))' = \\ & \sqrt{N}[(\bar{H}(t_1) - H(t_1), \dots, \bar{H}(t_n) - H(t_n))' + \\ & + g(\hat{S}(t_1), \dots, \hat{S}(t_n)) - g(\hat{S}(t_1), \dots, \hat{S}(t_n))] = \\ & \sqrt{N}[(\bar{H}(t_1), \dots, \bar{H}(t_n))' - g(\hat{S}(t_1), \dots, \hat{S}(t_n))] + \\ & + \sqrt{N}[g(\hat{S}(t_1), \dots, \hat{S}(t_n)) - g(S(t_1), \dots, S(t_n))]. \end{aligned}$$

Der zweite Teil konvergiert nach der δ -Methode in Wahrscheinlichkeit gegen eine Normalverteilung mit Mittelwert 0 und Kovarianzmatrix U mit $u_{ij} = t_{ij}/(S(t_i)S(t_j))$, wobei $T = (t_{ij})$ die asymptotische Kovarianzmatrix von $\sqrt{N}(\hat{S}(t_1) - S(t_1), \dots, \hat{S}(t_n) - S(t_n))$ bezeichnet. Es ist nämlich $U = D'TD$ und $D = -(\delta_{ij}/S(t_i))$. Es bleibt noch zu zeigen, daß der erste Ausdruck in Wahrscheinlichkeit gegen 0 strebt, was komponentenweise gezeigt wird:

$$\begin{aligned} 0 < \log N/(N-K_i) - \bar{H}(t_i) < \log[N(N-K_i+1)/\{(N-K_i)(N+1)\}] &\leq \\ \log[(N-K_i+1)/(N-K_i)] = \log(1+1/(N-K_i)) &\leq 1/(N-K_i) + 1/[2(N-K_i)^2] + \\ \dots = (N-K_i)^{-1}(1 + 1/[2(N-K_i)] + 1/[3(N-K_i)^2] + \dots) & \quad (*) \end{aligned}$$

Nun ist $1/(N-K_i) = (1/N)(1/[1-K_i/N])$ und $\text{plim}(1-K_i/N) = S(t_i) > 0$.

Daher ist $\text{plim}(\sqrt{N}/(N-K_1)) = 0$. Die Reihe $1 + x/2 + x^2/3 + \dots$ konvergiert für $|x| < 1$ absolut und für $|x| \leq 1-\delta$ gleichmäßig. Daher konvergiert in (*) der in Klammer stehende Ausdruck in Wahrscheinlichkeit gegen 1. Insbesondere erhalten wir, daß $\sqrt{N}(\log N/(N-K_1) - \bar{H}(t_1))$ in Wahrscheinlichkeit gegen 0 konvergiert.

Als Schätzungen für die Kovarianzmatrizen T und U erhalten wir:

$$\hat{t}_{ij}(\text{Typ-I}) = K_i(N-K_j)/N^2 \quad i \leq j$$

$$\hat{u}_{ij}(\text{Typ-I}) = K_i/(N-K_i) \quad i \leq j$$

Betrachtet man die analogen Schätzungen für den Fall der zufälligen Zensierung, dann erkennt man sofort den Zusammenhang:

$$\hat{u}_{ij} = N \sum_{t_{(k)} < t_i} 1/n_k(n_k-1) - N \sum_{N-K_i}^N 1/k^2 - N \int_{N-K_i}^N x^{-2} dx = N(-x^{-1} \Big|_{N-K_i}^N) =$$

$= N(1/(N-K_i) - 1/N) = K_i/(N-K_i) = \hat{u}_{ij}(\text{Typ-I})$. Dabei bedeutet - asymptotische Äquivalenz, deren exakter Beweis wie vorher für den Ausdruck $\bar{H}(t_1) - \log N/(N-K_1)$ geführt werden müßte. Die Formeln für den Fall zufällig zensierter Variable liefern also auch für diesen Fall richtige (= konsistente = asymptotisch äquivalente) Schätzungen, die aber komplizierter sind als die "natürlich" abgeleiteten.

Bemerkung. Die ganze Arbeit hindurch wollen wir die Bezeichnung "Typ-I-Zensierung" für "Single-Typ-I-Zensierung" verwenden.

5. Konstruktion von Anpassungstests

Mittels der Grenzverteilungen der endlichdimensionalen Verteilungen der Prozesse $\sqrt{N}(\hat{S}(t) - S(t))$ und $\sqrt{N}(\bar{H}(t) - H(t))$ und der δ -Methode sollen für verschiedene Survivalverteilungen Anpassungstests nach dem folgenden Muster konstruiert werden: bei gegebener Verteilung (mit unbekannten Parametern) suche man eine funktionale

Beziehung $g(x_1, \dots, x_n)$, sodaß $g(S(t_1), \dots, S(t_n)) = 0$ oder $g(H(t_1), \dots, H(t_n)) = 0$ gilt, für (gewisse oder beliebige) Werte $t_1, \dots, t_n$, unabhängig von den wahren Werten der unbekannten Parameter. Ersetzt man $S(t)$ durch die Schätzung $\hat{S}(t)$, $H(t)$ durch $\bar{H}(t)$, dann dürfen unter der Nullhypothese $g(\hat{S}(t_1), \dots, \hat{S}(t_n))$ resp. $g(\bar{H}(t_1), \dots, \bar{H}(t_n))$ nicht zu sehr von Null differieren. Als Maß für eine zulässige Abweichung wird die (asymptotische) Verteilung von $\sqrt{Ng}(\hat{S}(t_1), \dots, \hat{S}(t_n))$ resp. $\sqrt{Ng}(\bar{H}(t_1), \dots, \bar{H}(t_n))$ betrachtet. Diese Statistiken sind (unter sehr allgemeinen Bedingungen) asymptotisch normalverteilt, die Kovarianzmatrizen können konsistent geschätzt werden, sodaß Teststatistiken erstellt werden können, die unter der Nullhypothese asymptotisch χ^2 -verteilt sind. Konsistent sind derartige Tests gegen alle Verteilungen, für die für die gewählten $t_1, \dots, t_n$ die Beziehung $g(S(t_1), \dots, S(t_n)) \neq 0$ resp. $g(H(t_1), \dots, H(t_n)) \neq 0$ gilt. Durch Erhöhung der Anzahl der Punkte $t_1, \dots, t_n$ kann diese Menge erhöht werden. Die Erhöhung der Freiheitsgrade verringert aber gleichzeitig die Geschwindigkeit der Konvergenz gegen die Grenzverteilung. Für jede der betrachteten Verteilungen sind $n+1$ Punkte notwendig und hinreichend um eine Beziehung g zu finden, y die unabhängig von den Werten der unbekannten Parameter gilt (dabei ist n die Anzahl der unbekannten Parameter).

III. Anwendungen

1. Die Weibullverteilung

Eine zufällige Variable $T \geq 0$ heißt Weibull verteilt mit den Parametern $p, \alpha > 0$, wenn die Ratenfunktion $h(t)$ gegeben ist durch

$$h(t) = \alpha p (\alpha t)^{p-1} \quad (t \geq 0).$$

Für die kummulierte Hazardfunktion $H(t)$ erhalten wir daher

$$H(t) = (\alpha t)^p,$$

und für die Überlebensfunktion $S(t)$ ergibt sich

$$S(t) = \exp(-(\alpha t)^p).$$

Eine Funktion, wie in II.5 beschrieben, erhalten wir, indem wir Manipulationen durchführen, die die unbekannten Parameter zum Verschwinden bringen.

$$\log H(t) = p \log(\alpha t) = p \log \alpha + p \log t, \text{ daher ist}$$

$$\log H(t) - \log H(u) = p(\log t - \log u) \text{ und}$$

$$\frac{\log H(t) - \log H(u)}{\log H(u) - \log H(v)} = \frac{\log(t/u)}{\log(u/v)}$$

Durch Ausmultiplizieren erhalten wir:

$$\log(u/v)[\log H(t) - \log H(u)] = \log(t/u)[\log H(u) - \log H(v)].$$

Alles auf eine Seite gebracht ergibt:

$$\log(u/v)\log H(t) + \log(v/t)\log H(u) + \log(t/u)\log H(v) = 0.$$

Diese Beziehung gilt auch, wenn $H(u) = H(v)$ (i.e. $u = v$).

Wie in Satz 7. beschrieben, wähle man nun ein $K > 0$ mit $S(K)G(K) > 0$ und t, u, v mit $0 < t < u < v < K$ (als Richtgröße für K kann die größte Beobachtung dienen).

Sei $g: \mathbb{R}^3 \rightarrow \mathbb{R}$ definiert durch:

$$g(x, y, z) = \log(u/v)\log x + \log(v/t)\log y + \log(t/u)\log z.$$

Dann ist $g(H(t), H(u), H(v)) = 0$ und $\sqrt{Ng(H^*(t), H^*(u), H^*(v))}$ ist asymptotisch normalverteilt, wobei $H^* = \hat{H} = -\log \hat{S}$ oder

$H^* = \bar{H}$. Wir beschränken uns zunächst auf $H^*(t) = \bar{H}(t)$, d.h. H

wird durch die empirische kummulierte Hazardfunktion geschätzt:

$$\tilde{H}(x) = \sum_{t_{(i)} < x} 1/n_i \quad x \in \{t, u, v\}$$

Die asymptotische Kovarianzmatrix $V = (\text{cov}(x, y))$ ($x, y \in \{t, u, v\}$) von $\sqrt{N}(\tilde{H}(t) - H(t), \tilde{H}(u) - H(u), \tilde{H}(v) - H(v))$ wird konsistent geschätzt durch:

$$\text{cov}(x, y)^{-} = N\tilde{v}_{xy} = N \sum_{t_{(i)} < \min(x, y)} 1/n_i(n_i - 1)$$

Die asymptotische Varianz von $\sqrt{N}g(\tilde{H}(t), \tilde{H}(u), \tilde{H}(v))$ ist dann gegeben durch $G'VG$, wobei G der Vektor der partiellen Ableitungen von g an der Stelle $(H(t), H(u), H(v))$ ist. Wir erhalten für den Vektor G :

$$G = (\log(u/v)/H(t), \log(v/t)/H(u), \log(t/u)/H(v))'.$$

Eine konsistente Schätzung für die asymptotische Varianz σ^2 von $\sqrt{N}g(\tilde{H}(t), \tilde{H}(u), \tilde{H}(v))$ erhält man, indem man die unbekannten Ausdrücke in V und G durch konsistente Schätzungen ersetzt. Wir setzen $t' = u$, $u' = v$ und $v' = t$, dann ist eine konsistente Schätzung $\tilde{\sigma}^2$ von σ^2 gegeben durch:

$$\tilde{\sigma}^2 = N \sum_{x, y \in \{t, u, v\}} \tilde{v}_{xy} \tilde{g}_x \tilde{g}_y.$$

Dabei ist

$$\tilde{v}_{xy} = \sum_{t_{(i)} < \min(x, y)} 1/n_i(n_i - 1)$$

und

$$\tilde{g}_x = \log(x'/x'') / \sum_{t_{(i)} < x} 1/n_i$$

Als Teststatistik erhalten wir:

$$H = \frac{\left(\sum_{x \in \{t, u, v\}} \log(x'/x'') \tilde{H}(x) \right)^2}{\sum_{x, y \in \{t, u, v\}} \tilde{v}_{xy} \tilde{g}_x \tilde{g}_y},$$

die unter der Nullhypothese "T ist Weibull verteilt" asymptotisch

Chi²-verteilt ist mit einem Freiheitsgrad. Dies gilt für zufällige und Single-Typ-I-Zensierung. Für das letztere Zensierungsmuster kann noch folgende alternative Teststatistik angegeben werden: bei Typ-I Zensierung ist $S(t)$ einfacher zu schätzen als $H(t)$, wir betrachten daher $\hat{H}(t) = -\log \hat{S}(t)$. Sei $K > 0$ eine Konstante mit $K \leq C$, wobei C die Zensierungszeit bezeichnet. Weiters seien wieder t, u, v gewählt mit $0 < t < u < v < K$. Sei A_x die Anzahl der Beobachtungen bis zum Zeitpunkt x ($x \in \{t, u, v\}$) und $K_x = N - A_x$. Dann ist $\hat{S}(x) = K_x/N$. Wir betrachten nun die Funktion $g: \mathbb{R}^3 \rightarrow \mathbb{R}$, definiert durch:

$$g(x, y, z) = \log(u/v) \log(-\log x) + \log(v/t) \log(-\log y) + \\ + \log(t/u) \log(-\log z).$$

Dann ist $g(S(t), S(u), S(v)) = 0$ und $\sqrt{N}g(\hat{S}(t), \hat{S}(u), \hat{S}(v))$ ist asymptotisch normalverteilt mit Mittelwert 0. Die asymptotische Varianz berechnet sich nach der δ -Methode: $\sigma^2 = G' T G$, wobei $T = (\text{cov}(x, y))$ mit $\text{cov}(x, y) = S(\max(x, y))(1 - S(\min(x, y)))$. G ist der Vektor der partiellen Ableitungen von g , ausgewertet an der Stelle $(S(t), S(u), S(v))$. Für den Vektor G erhalten wir daher:

$$G' = (\log(u/v)/(S(t)\log S(t)), \log(v/t)/(S(u)\log S(u)), \\ \log(t/u)/(S(v)\log S(v))).$$

Eine konsistente Schätzung der asymptotischen Varianz σ^2 von $\sqrt{N}g(\hat{S}(t), \hat{S}(u), \hat{S}(v))$ ist gegeben durch den folgenden Ausdruck

$$\hat{\sigma}^2 = \sum_{x, y \in \{t, u, v\}} \hat{v}_{xy} \hat{g}_x \hat{g}_y,$$

wobei

$$\hat{v}_{xy} = K_y(N - K_x)/N^2 \quad (x \leq y)$$

und

$$\hat{g}_x = N \log(x'/x'') / (K_x \log(K_x/N))$$

Insgesamt erhalten wir damit:

$$\begin{aligned}\hat{v}_{xy}\hat{g}_x\hat{g}_y &= \frac{K_{\max}(x,y)(N-K_{\min}(x,y))\log(x'/x'')\log(y'/y'')}{K_x K_y \log(K_x/N)\log(K_y/N)} = \\ &= \frac{(N-K_{\min}(x,y))\log(x'/x'')\log(y'/y'')}{K_{\min}(x,y)\log(K_x/N)\log(K_y/N)}\end{aligned}$$

Als Teststatistik erhalten wir damit explizit:

$$H = \frac{N(\sum_x \log(x'/x'')\log(-\log(K_x/N)))^2}{\sum_{x,y} \frac{(N-K_{\min}(x,y))\log(x'/x'')\log(y'/y'')}{K_{\min}(x,y)\log(K_x/N)\log(K_y/N)}}$$

die unter der Nullhypothese wieder asymptotisch χ^2 -verteilt ist mit einem Freiheitsgrad.

Für die folgenden Verteilungen werden wir für die Teststatistiken Formeln von ähnlicher Struktur erhalten. Die Summation wird dann im Zähler jeweils über den Index x , im Nenner über die Indizes x, y erfolgen, ohne daß dies immer eigens erwähnt wird.

Konsistent ist der so konstruierte Test gegen alle Alternativen, für die $g(H(t), H(u), H(v)) \neq 0$ gilt. Wie vom Autor (1988) für die Exponentialverteilung durchgeführt, kann die Menge der konsistenten Alternativen vergrößert werden, wenn die Beziehung $g(x, y, z)$ nicht nur an den 3 Stellen t, u, v , sondern an mehreren $x_1, \dots, x_{n+2}$ in $(0, K)$ ausgewertet wird: anstatt der reellwertigen Funktion g wird die vektorwertige Funktion $(g(x_1, x_i, x_{n+2}))$ ($i = 2, \dots, n+1$) betrachtet. Ersetzt man x_i durch $\hat{S}(x_i)$, erhält man eine asymptotisch normalverteilte Variable $\sqrt{N}\hat{g}$ und daraus eine Teststatistik, die unter der Nullhypothese asymptotisch χ^2 -verteilt ist mit n Freiheitsgraden. Konsistent ist ein derartiger Test gegen alle Alternativen, für die $g(S(x_1), S(x_i), S(x_{n+2})) \neq 0$ für mindestens ein $i \in \{2, \dots, n+1\}$ gilt. Auf Details soll hier nicht

näher eingegangen werden. Für die Exponentialverteilung wurde dieses Prinzip ausführlich vom Autor (1988) studiert.

Wichtige Alternativen zur Weibullverteilung sind Verteilungen, deren Ratenverläufe nicht monoton sind. Insbesondere solche mit einmaligem Wechsel des Ratenverlaufs. Seien $0 < t < v < K$ derart, daß der Wechsel im Ratenverlauf zwischen t und v stattfindet. Wie müßte bei gegebener Alternative ein u zwischen t und v gewählt werden, damit vom Test das Nichtzutreffen der Nullhypothese möglichst gut erkannt wird?

Wir betrachten die Funktion:

$$f(u) = \log(u/v)\log H(t) + \log(v/t)\log H(u) + \log(t/u)\log H(v),$$

wobei u im Intervall $[t, v]$ variieren soll. Dann gilt $f(t) = f(v) = 0$ und $f'(u) = \log(v/t)h(u)/H(u) - (\log H(v) - \log H(t))/u$. Der Test beruht darauf, daß unter H_0 $f(u) = 0$. Wir wählen daher bei gegebener Alternative u derart, daß $|f(u)|$ möglichst groß wird. Derartige Punkte u sind als Lösung der Gleichung $f'(u) = 0$ zu finden.

Es ist $f'(u) = 0 \Leftrightarrow uh(u)/H(u) = \log(H(v)/H(t))/\log(v/t)$. Sei nun

$x(u) < u$ derart, daß $\int_0^u h(s)ds = h(x(u))u$. Ein Wert u , sodaß bei

gegebener Alternative der Wert $|f(u)|$ möglichst groß wird, muß also der Gleichung

$$\frac{h(u)}{h(x(u))} = \frac{\log(H(v)/H(t))}{\log(v/t)} \quad \text{genügen.}$$

2. Die Log-logistische Verteilung

Eine zufällige Variable $T \geq 0$ heißt log-logistisch verteilt, wenn $\log T$ logistisch verteilt ist.

Die Ratenfunktion $h(t)$ ist gegeben durch

$$h(t) = \frac{\alpha p(\alpha t)^{p-1}}{1+(\alpha t)^p}$$

Für die kummulierte Hazardfunktion $H(t)$ erhalten wir:

$$H(t) = \log(1+(\alpha t)^p)$$

und für die Überlebensfunktion $S(t)$ erhalten wir damit

$$S(t) = 1/(1+(\alpha t)^p).$$

Wir gehen ähnlich wie in 1. vor, um die gesuchte Funktion zu erhalten:

$$1/S(t) = 1 + (\alpha t)^p \Rightarrow$$

$$1/S(t) - 1 = (\alpha t)^p \Rightarrow$$

$$\log(1/S(t)-1) = p \log \alpha + p \log t.$$

Daraus erhalten wir ähnlich wie 1.:

$$\begin{aligned} \log(u/v) \log(1/S(t)-1) + \log(v/t) \log(1/S(u)-1) + \\ + \log(t/u) \log(1/S(v)-1) = 0 \end{aligned}$$

Wieder wähle man ein $K > 0$ mit $S(K)G(K) > 0$ und t, u, v mit $0 < t < u < v < K$. Sei $g: \mathbb{R}^3 \rightarrow \mathbb{R}$ definiert durch:

$$g(x, y, z) = \log(u/v) \log(1/x-1) + \log(v/t) \log(1/y-1) + \log(t/v) \log(1/z-1).$$

Dann ist $g(S(t), S(u), S(v)) = 0$, und $\sqrt{N}g(\hat{S}(t), \hat{S}(u), \hat{S}(v))$ ist asymptotisch normalverteilt mit Mittelwert 0. Die asymptotische Varianz wird wieder nach der Formel der δ -Methode berechnet. Der Vektor G der partiellen Ableitungen von g an der Stelle $(S(t), S(u), S(v))$ ist gegeben durch:

$$G = \left(\frac{1}{S(t)(S(t)-1)} \log \frac{u}{v}, \frac{1}{S(u)(S(u)-1)} \log \frac{v}{t}, \frac{1}{S(v)(S(v)-1)} \log \frac{t}{u} \right)'$$

Eine konsistente Schätzung für die asymptotische Varianz σ^2 von $\sqrt{N}g(\hat{S}(t), \hat{S}(u), \hat{S}(v))$ kann dann folgendermaßen angegeben werden:

$$\hat{\sigma}^2 = N \Sigma \hat{v}_{xy} \hat{g}_x \hat{g}_y,$$

wobei:

$$\begin{aligned}\hat{v}_{xy} &= \hat{S}(x)\hat{S}(y) \sum_{t(j) < \min(x,y)}^{\Sigma} 1/n_j(n_j-1) = \\ &= \hat{S}(x)\hat{S}(y)\hat{d}_{\min(x,y)},\end{aligned}$$

und

$$\hat{g}_x = \log(x'/x'') / [\hat{S}(x)(\hat{S}(x)-1)].$$

Bezeichnet man mit $F(t) = 1 - S(t)$ die Verteilungsfunktion und setzt man $\hat{F}(t) = 1 - \hat{S}(t)$, dann erhalten wir folgende Teststatistik:

$$H = \frac{(\sum \log(x'/x'') \log(\hat{F}(x)/\hat{S}(x)))^2}{\sum \log(x'/x'') \log(y'/y'') \hat{d}_{\min(x,y)} / [\hat{F}(x)\hat{F}(y)]},$$

die unter der Nullhypothese "T ist log-logistisch verteilt" asymptotisch χ^2 -verteilt ist mit einem Freiheitsgrad.

Wieder gilt die Teststatistik für zufällig und Typ-I-zensierte Daten. Alles in 1. über die Menge der konsistenten Alternativen gesagte gilt auch hier.

Wir wollen nun wieder die Modifikationen studieren, die sich ergeben, wenn nur Typ-I-Zensierung betrachtet wird:

K_x habe wieder die Bedeutung wie in 1. Dann ist $\hat{S}(x) = K_x/N$. Die asymptotische Kovarianzmatrix V von

$\sqrt{N}(\hat{S}(t)-S(t), \hat{S}(u)-S(u), \hat{S}(v)-S(v))$ ist gegeben durch:

$$\text{cov}(x,y) = v_{xy} = S(y)(1-S(x)) \quad (x \leq y)$$

Berechnet man mit der δ -Methode die asymptotische Varianz von $\sqrt{Ng}(\hat{S}(t), \hat{S}(u), \hat{S}(v))$ dann erhalten wir:

$$\sigma^2 = \sum v_{xy} g_x g_y.$$

Für $v_{xy} g_x g_y$ ergibt sich folgender Ausdruck:

$$v_{xy} g_x g_y = S(\max(x,y)) F(\min(x,y)) \frac{\log(x'/x'') \log(y'/y'')}{S(x) F(x) S(y) F(y)} =$$

$$= \frac{\log(x'/x'')\log(y'/y'')}{S(\min(x,y))F(\max(x,y))}.$$

Ersetzt man die unbekannten Werte durch Schätzungen, dann erhalten wir folgende Teststatistik:

$$H = \frac{(\sum \log(x'/x'')\log[(N-K_x)/K_x])^2}{N \sum \log(x'/x'')\log(y'/y'')/[K_{\min(x,y)}(N-K_{\max(x,y)})]},$$

die unter der Nullhypothese "T ist log-logistisch verteilt" asymptotisch χ^2 -verteilt ist mit einem Freiheitsgrad.

3. Die Log-normal Verteilung

Ein Gesetz von ähnlicher Form erhalten wir auch für diese Verteilung. Eine Variable $T \geq 0$ heißt log-normal verteilt, wenn $\log T$ normalverteilt ist. Man kann daher auf $\{\log T_i\}$ die bekannten Tests für die Normalverteilung anwenden. Diese Methoden versagen aber, wenn, wie in unserem Fall, zensierte Daten vorliegen.

Sei also T log-normalverteilt, also $\log T$ normalverteilt mit den Parametern μ und σ , d.h.

$$P((\log T - \mu)/\sigma > t) = \Phi(t), \text{ wobei}$$

$$\Phi(t) = (2\pi)^{-1/2} \int_t^{\infty} \exp(-v^2/2) dv$$

Bemerkung. Üblicherweise wird das Komplement dieser Funktion, die dazugehörige Verteilungsfunktion mit dem Buchstaben Φ bezeichnet. Wir erhalten:

$$P(\log T > \sigma t + \mu) = \Phi(t) = P(T > \exp(\sigma t + \mu)),$$

und daher ist die Überlebensfunktion der Log-normal-Verteilung gegeben durch:

$$S(t) = P(T > t) = \Phi((\log t - \mu)/\sigma).$$

Die Hazardfunktion und die kummulierte Hazardfunktion bestimmen sich dann nach den in der Einleitung gegebenen Regeln. Ist θ die eindeutig bestimmte Umkehrfunktion von Φ , dann gilt:

$$\theta S(t) = (\log t - \mu)/\sigma, \text{ daher ist}$$

$$\theta S(t) - \theta S(u) = (\log t - \log u)/\sigma = \log(t/u)/\sigma.$$

Somit erhalten wir:

$$(\theta S(t) - \theta S(u))\log(u/v) = (\theta S(u) - \theta S(v))\log(t/u).$$

Daraus ergibt sich eine Funktion ähnlicher Form wie in 1. und 2.:

$$\log(u/v)\theta S(t) + \log(v/t)\theta S(u) + \log(t/u)\theta S(v) = 0$$

Wir wählen wieder ein $K > 0$ mit $S(K)G(K) > 0$ und t, u, v mit $0 < t < u < v < K$ und betrachten die Funktion $g: \mathbb{R}^3 \rightarrow \mathbb{R}$, definiert durch:

$$g(x, y, z) = \log(u/v)\theta(t) + \log(v/t)\theta(u) + \log(t/u)\theta(v).$$

Dann ist $\sqrt{N}g(\hat{S}(t), \hat{S}(u), \hat{S}(v))$ asymptotisch normalverteilt mit Mittelwert 0. Die asymptotische Varianz wird wieder nach der Formel der δ -Methode berechnet.

Zunächst ist $\Phi'(t) = -\exp(-t^2/2)/\sqrt{2\pi}$. Nach der Regel für die Differentiation der Umkehrfunktion gilt:

$$\theta'(t) = 1/\Phi'(\theta(t)). \text{ Daher ist}$$

$$\theta'(t) = -\sqrt{2\pi}\exp(\theta(t)^2/2).$$

Für den Vektor der partiellen Ableitungen von g , ausgewertet an der Stelle $(S(t), S(u), S(v))$, ergibt sich somit:

$$G = -\sqrt{2\pi}(\log(u/v)\exp(\theta S(t)^2/2), \log(v/t)\exp(\theta S(u)^2/2), \log(t/u)\exp(\theta S(v)^2/2))'.$$

Für die asymptotische Varianz von $\sqrt{N}g(\hat{S}(t), \hat{S}(u), \hat{S}(v))$ gilt dann $\sigma^2 = G'VG$, wobei V die asymptotische Kovarianzmatrix von $\sqrt{N}(\hat{S}(t)-S(t), \hat{S}(u)-S(u), \hat{S}(v)-S(v))$ ist. Als konsistente Schätzung für σ^2 erhalten wir:

$$\hat{\sigma}^2 = N\hat{\Sigma}\hat{v}_{xy}\hat{g}_x\hat{g}_y, \text{ wobei}$$

$$\hat{v}_{xy} = \hat{S}(x)\hat{S}(y) \sum_{t_{(i)} < \min(x,y)} 1/n_i(n_i-1)$$

Setzt man $\hat{d}_x = \sum_{t(i) < x} 1/n_i(n_i-1)$,

dann erhält man:

$$\hat{\sigma}^2 = 2\pi N \sum_{x,y \in \{t,u,v\}} \log(x'/x'') \log(y'/y'') \hat{d}_{\min(x,y)} \hat{S}(x) \hat{S}(y) \cdot \exp([\theta \hat{S}(x)^2 + \theta \hat{S}(y)^2]/2)$$

Die Teststatistik wird hier also nicht besonders einfach, da sich nichts wegekürzt (außer einem N):

Es gilt:

$$H = \frac{N(\sum \log(x'/x'') \theta \hat{S}(x))^2}{\hat{\sigma}^2}$$

ist unter der Nullhypothese "T ist log-normal verteilt" asymptotisch χ^2 -verteilt mit einem Freiheitsgrad.

Wir betrachten nun wieder die Modifikationen, die sich bei Typ-I-Zensierung ergeben: K_x habe wieder die Bedeutung, wie in den vorhergehenden Abschnitten, dann ist $\hat{S}(x) = K_x/N$. Die asymptotische Varianz wird wieder nach der Formel $\sigma^2 = G'VG$ berechnet. Daher ist

$$\hat{\sigma}^2 = \sum \hat{v}_{xy} \hat{g}_x \hat{g}_y, \text{ wobei}$$

$$\hat{v}_{xy} = K_{\max(x,y)}(N - K_{\min(x,y)})/N^2, \text{ und}$$

$$\hat{g}_x = -\sqrt{2\pi} \log(x'/x'') \exp(\theta(K_x/N)^2/2).$$

Hier lassen sich offensichtlich keine Vereinfachungen durchführen, sodaß sich folgende Teststatistik ergibt:

$$H = \frac{N^3 (\sum \log(x'/x'') \theta (K_x/N))^2}{2\pi \sum \log(x'y'/x''y'') K_{\max(x,y)} \cdot (N - K_{\min(x,y)}) \exp([\theta (K_x/N)^2 + \theta (K_y/N)^2]/2)}$$

Unter der Nullhypothese "T ist log-normal verteilt" ist H asymptotisch χ^2 -verteilt mit einem Freiheitsgrad.

Ansonsten gilt wieder das schon in 1. und 2. Gesagte.

4. Die Rayleigh Verteilung

Eine Variable $T \geq 0$ heißt Rayleigh verteilt mit den Parametern $a, b > 0$, wenn die Ratenfunktion gegeben ist durch:

$$h(t) = a + 2bt.$$

Für die kummulierte Hazardfunktion $H(t)$ ergibt sich somit:

$$H(t) = at + bt^2,$$

weshalb sich für die Überlebensfunktion:

$$S(t) = \exp(-at-bt^2)$$

ergibt. Wir betrachten nur die (einfachere) Beziehung für die kummulierte Hazardfunktion.

Zunächst gilt:

$$H(x+y) = a(x+y) + b(x+y)^2 = H(x) + H(y) + 2bxy.$$

Der Parameter b wird eliminiert, indem von der rechten Seite $H(x)+H(y)$ subtrahiert wird und dann diese Zahl durch $H(s+t)-H(s)-H(t)$ dividiert wird. Um eine möglichst einfache Form zu erhalten, wählen wir $x = y = t$ und $s = 2x$ mit $3x < K$. Dann erhalten wir:

$$\frac{H(2x)-2H(x)}{H(3x)-H(2x)-H(x)} = \frac{1}{2}$$

und damit:

$$H(3x) - 3H(2x) + 3H(x) = 0.$$

Wir betrachten die Funktion $g: \mathbb{R}^3 \rightarrow \mathbb{R}$, definiert durch:

$$g(r,s,t) = t - 3s + 3r,$$

dann ist $\sqrt{N}g(\tilde{H}(x), \tilde{H}(2x), \tilde{H}(3x))$ asymptotisch normalverteilt mit Mittelwert 0. Um die Varianz σ^2 zu ermitteln, berechnen wir zuerst den Vektor der partiellen Ableitungen von g , ausgewertet an der Stelle $(H(x), H(2x), H(3x))$. Es ergibt sich: $G' = (3, -3, 1)$.

Damit kann man die asymptotische Varianz explizit ausrechnen:

$$\sigma^2 = -3\text{var}(\tilde{H}(x)) + 3\text{var}(\tilde{H}(2x)) + \text{var}(\tilde{H}(3x)).$$

Als konsistente Schätzung von σ^2 erhalten wir dann:

$$\tilde{\sigma}^2 = N \left(\sum_{t(j) < 3x} \frac{1}{n_j(n_j-1)} + 3 \sum_{x \leq t(j) < 2x} \frac{1}{n_j(n_j-1)} \right)$$

Somit erhalten wir als (relativ einfache) Teststatistik:

$$H = \frac{\left(\sum_{t(j) < 3x} \frac{1}{n_j} - 3 \sum_{x \leq t(j) < 2x} \frac{1}{n_j} \right)^2}{\sum_{t(j) < 3x} \frac{1}{n_j(n_j-1)} + 3 \sum_{x \leq t(j) < 2x} \frac{1}{n_j(n_j-1)}},$$

die unter der Nullhypothese "T ist Rayleigh verteilt" asymptotisch χ^2 -verteilt ist mit einem Freiheitsgrad.

Für den Fall Typ-I zensierter Daten betrachten wir die Schätzung

$\hat{H}(x) = -\log \hat{S}(x) = \log(N/K_x)$. Wir müssen beachten, daß nun die

Funktion g folgende Form hat:

$$g(r, s, t) = -\log t + 3\log s - 3\log r.$$

Für den Vektor der partiellen Ableitungen von g an der Stelle

$(S(x), S(2x), S(3x))$ erhalten wir daher:

$$G' = (-3/S(x), 3/S(2x), -1/S(3x)).$$

Rechnet man nach der Formel der δ -Methode die asymptotische Varianz σ^2 aus, dann erhält man:

$$\sigma^2 = -3F(x)/S(x) + 3F(2x)/S(2x) + F(3x)/S(3x),$$

wobei $F(t) = 1 - S(t)$ die Verteilungsfunktion bezeichnet.

Berechnet man $g(\hat{S}(x), \hat{S}(2x), \hat{S}(3x))$ explizit, dann erhält man:

$$g(\hat{S}(x), \hat{S}(2x), \hat{S}(3x)) = \log(N/K_{3x}) + 3\log(K_{2x}/K_x).$$

Daraus erhalten wir folgende Teststatistik:

$$H = \frac{N(\log(N/K_{3x}) + 3\log(K_{2x}/K_x))^2}{(N-K_{3x})/K_{3x} + 3(N-K_{2x})/K_{2x} - 3(N-K_x)/K_x},$$

die unter der Nullhypothese "T ist Rayleigh verteilt" asymptotisch

χ^2 -verteilt ist mit einem Freiheitsgrad.

5. Die Gompertz Verteilung

Eine Variable $T \geq 0$ heißt Gompertz verteilt mit den Parametern $b, c > 0$, wenn die Ratenfunktion $h(t)$ die folgende Gestalt besitzt:

$$h(t) = b \exp(ct).$$

Für die kummulierte Hazardfunktion ergibt sich daraus:

$$H(t) = b(\exp(ct) - 1)/c,$$

und für die Überlebensfunktion erhalten wir:

$$S(t) = \exp(b(1 - \exp(ct))/c).$$

Wir wählen wieder die kummulierte Hazardfunktion, um die gesuchte Beziehung g zu finden:

$$H(x) = b \exp(cx)/c - b/c, \text{ daher ist } H(x+y) = b \exp(c(x+y))/c - b/c.$$

Daraus ergibt sich:

$$H(x+y) - H(x) = b \exp(cx) \exp(cy)/c - b \exp(cx)/c = b \exp(cx) [\exp(cy) - 1]/c = b \exp(cx) H(y). \text{ Wir erhalten also:}$$

$$[H(x+y) - H(x)]/H(y) = b \exp(cx).$$

Die rechte Seite ist offenbar unabhängig von y , daher gilt:

$$[H(x+y) - H(x)]/H(y) = [H(x+z) - H(x)]/H(z).$$

Wir wählen die Variablen wieder so, daß der gesuchte Zusammenhang eine möglichst einfache Form hat: $x = y$, $z = 2x$. Dann erhalten wir:

$$[H(2x) - H(x)]/H(x) = [H(3x) - H(x)]/H(2x), \text{ oder, nach Multiplikation mit } H(x)H(2x):$$

$$H(3x)H(x) - H^2(2x) + H(2x)H(x) - H^2(x) = 0.$$

Setzen wir nun $g: \mathbb{R}^3 \rightarrow \mathbb{R}$:

$$g(r, s, t) = rt - s^2 + sr - r^2,$$

dann ist $g(H(x), H(2x), H(3x)) = 0$ und $\sqrt{N}g(\bar{H}(x), \bar{H}(2x), \bar{H}(3x))$ ist asymptotisch normalverteilt mit Mittelwert 0.

Der Vektor G der partiellen Ableitungen von g an der Stelle $(H(x), H(2x), H(3x))$ ist gegeben durch:

$$G = (H(3x) + H(2x) - 2H(x), -2H(2x) + H(x), H(3x))'.$$

Wieder ist die Teststatistik:

$$H = \frac{(\bar{H}(3x)\bar{H}(x) - \bar{H}^2(2x) + \bar{H}(2x)\bar{H}(x) - \bar{H}^2(x))^2}{\bar{G}'\bar{U}\bar{G}}$$

unter der Nullhypothese "T ist Gompertz verteilt" asymptotisch χ^2 -verteilt mit einem Freiheitsgrad.

Dabei bedeutet:

$$\bar{G} = (\bar{H}(3x) + \bar{H}(2x) - 2\bar{H}(x), -2\bar{H}(2x) + \bar{H}(x), \bar{H}(3x))' \text{ und}$$

$$\bar{U} = (\bar{u}_{ij}) \text{ wobei } \bar{u}_{ij} = \frac{\sum_{t_{(1)} < \min(i,j)x} 1}{n_1(n_1-1)}.$$

Auf ein explizites Ausrechnen der Teststatistik H für diesen Fall und der Modifikation bei Typ-I zensierten Daten wollen wir diesmal verzichten, weil es ziemlich kompliziert werden würde und keine Vereinfachungen abzusehen sind (nur ein N kürzt sich in Zähler und Nenner weg).

6. Die Pareto Verteilung

Eine zufällige Variable $T \geq 0$ heißt Pareto-verteilt mit den Parametern $a, b > 0$, wenn ihre Ratenfunktion $h(t)$ gegeben ist durch

$$h(t) = a/(1+bt).$$

Für die kummulierte Hazardfunktion $H(t)$ erhält man dann:

$$H(t) = a \log(1+bt)/b,$$

und die Überlebensfunktion $S(t)$ lautet:

$$S(t) = (1+bt)^{b/a}.$$

Wir wählen wieder die Funktion H , um die gesuchte funktionale Be-

ziehung g zu finden. Zunächst ist:

$$H(x)/H(y) = \log(1+bx)/\log(1+by).$$

Hier erscheint es nicht möglich, durch "herkömmliche" (algebraische) Operationen den Parameter b zu eliminieren. Wir verwenden folgendes:

Lemma. Seien $0 < y < x$ vorgegebene reelle Zahlen. Die Funktion

$$F_{x,y}: z \longmapsto \log(1+xz)/\log(1+yz) \quad (z > 0)$$

ist eine stetig differenzierbare Bijektion zwischen den Intervallen $(0, \infty)$ und $(1, x/y)$.

Beweis. (i) Nach der Regel von de l'Hospital gilt

$$\lim_{z \rightarrow 0} F(z) = \lim_{z \rightarrow 0} \frac{x(1+yz)}{(1+xz)y} = x/y \text{ und}$$

$$\lim_{z \rightarrow \infty} F(z) = \lim_{z \rightarrow \infty} \frac{x(1+yz)}{(1+xz)y} = \lim_{z \rightarrow \infty} \frac{yx}{xy} = 1.$$

Da F stetig ist, ist $(1, x/y)$ im Wertebereich von F enthalten.

(ii) Wir zeigen nun, daß $F'(z) < 0$ für alle $z > 0$. Dann ist F streng monoton fallend, daher injektiv und nimmt außerhalb von $(1, x/y)$ keine Werte an. Zunächst ist:

$$F'(z) = \frac{x \log(1+yz)/(1+xz) - y \log(1+xz)/(1+yz)}{\log^2(1+yz)}$$

Es ist $F'(z) < 0$ genau dann, wenn

$$(1+yz)x \log(1+yz)/y - (1+xz) \log(1+xz) < 0 \quad (*)$$

Wir behaupten, daß die Ungleichung $(*)$ für beliebige $x, y, z > 0$ mit $x > y$ gültig ist. Dazu betrachten wir für fixe $y, z > 0$ die Funktion

$$f(x) = x(1+yz) \log(1+yz)/y - (1+xz) \log(1+xz) \quad (x \geq y).$$

Dann ist $f(y) = 0$ und

$$f'(x) = (1+yz) \log(1+yz)/y - z \log(1+xz) - z \text{ und daher}$$

$$\begin{aligned}
yf'(x) &= (1+yz)\log(1+yz) - yz - yz\log(1+xz) = \\
&= \log(1+yz) + yz\log(1+yz) - yz - yz\log(1+xz) = \\
&= yz[\log(1+yz) - \log(1+xz)] + [\log(1+yz) - yz].
\end{aligned}$$

Der erste Summand ist negativ, da $y < x$ und der Logarithmus monoton wachsend ist, der zweite Summand ist negativ wegen der Ungleichung $\log(1+u) < u$. Die Funktion f ist im Intervall (y, ∞) streng monoton fallend. Wegen $f(y) = 0$ ist $f(x) < 0$ für $x > y$ und daher $F'(z) < 0$ für beliebige $x, y, z > 0$ mit $x > y$.

Damit existiert zu F die eindeutig bestimmte Umkehrabbildung F^{-1} : $(1, x/y) \longrightarrow (0, \infty)$ mit $F^{-1}(1) = \infty$ und $F^{-1}(x/y) = 0$. Seien nun $x > y > 0$ fix gewählt. Mit $G_{x,y}(t)$ wollen wir folgende Funktion bezeichnen:

$$G_{x,y}(t) = \begin{cases} \infty & \text{für } t \leq 1 \\ F_{x,y}^{-1}(t) & \text{für } 1 < t < x/y \\ 0 & \text{für } x/y \leq t \end{cases}$$

Dann ist $G_{x,y}$ stetig, im Intervall $(1, x/y)$ differenzierbar und es gilt $G_{x,y}(H(x)/H(y)) = b$.

Nun wählen wir wieder t, u, v mit $0 < t < u < v < K$ und eine Funktion $g: \mathbb{R}^3 \longrightarrow \mathbb{R}$ wie folgt:

$$g(x, y, z) = G_{v,u}(z/y) - G_{u,t}(y/x),$$

dann gilt $g(H(t), H(u), H(v)) = 0$ und $\sqrt{N}g(\tilde{H}(t), \tilde{H}(u), \tilde{H}(v))$ ist asymptotisch normalverteilt mit Mittelwert 0. Um die asymptotische Varianz zu berechnen, muß der Vektor der partiellen Ableitungen von $g(x, y, z)$ berechnet werden. Die Ableitung von $G_{u,t}(x)$ wird dann nach der Formel für die Differentiation der Umkehrfunktion berechnet $G'_{u,t}(x) = 1/F'_{u,t}(G_{u,t}(x))$. Auf eine explizite Berechnung des Vektors G und der asymptotischen Varianz $\sigma^2 = G'VG$ wollen wir auch hier wieder verzichten (da keine Vereinfachungen abzusehen sind).

Die Teststatistik

$$H = \frac{(G_{v,u}(\tilde{H}(v)/\tilde{H}(u)) - G_{u,t}(\tilde{H}(u)/\tilde{H}(t)))^2}{\tilde{G}'\tilde{U}\tilde{G}},$$

ist unter der Nullhypothese "T ist Pareto verteilt" asymptotisch χ^2 -verteilt mit einem Freiheitsgrad. Dabei ist $\tilde{G}$ der Vektor der partiellen Ableitungen von g , ausgewertet an der Stelle $(\tilde{H}(t), \tilde{H}(u), \tilde{H}(v))$, und $\tilde{u}_{ij} = \frac{\sum_{t(1) < \min(i,j)} 1}{n_1(n_1-1)}$, wobei $i, j \in \{t, u, v\}$.

Auf die Berechnung der Modifikationen bei Typ-I-Zensierung wollen wir hier ebenfalls wieder verzichten.

Bemerkungen.

(i) Der Wert $G_{v,u}(\tilde{H}(v)/\tilde{H}(u))$ kann ∞ werden. Dies ist aber nur der Fall, wenn zwischen v und u keine Beobachtung gemacht wurde. Dann sollte die Nullhypothese abgelehnt werden, weil der Funktionsverlauf (unter H_0) nicht der Datenlage entspricht. Sind überhaupt nur wenige Beobachtungen vorhanden, dann sollte der Test nicht angewandt werden.

(ii) $G_{v,u}$ ist im Punkt $x = v/u$ nicht differenzierbar. Sollte daher $\tilde{H}(v)/\tilde{H}(u)$ mit v/u zusammenfallen, dann sollte als Ableitung der linksseitige Limes $G'_{v,u}(\tilde{H}(v)/\tilde{H}(u) - 0)$ gewählt werden. Es ändert sich aber nichts an den asymptotischen Aussagen, da $H(v)/H(u)$ unter H_0 ein innerer Punkt von $(1, v/u)$ ist und daher die δ -Methode angewandt werden kann.

7. Zusammenfassung

Alle in 1. - 6. konstruierten Tests sind konsistent gegen alle Alternativen, für die für die gewählten Werte t, u, v (resp. $x, 2x, 3x$) $g(S(t), S(u), S(v)) \neq 0$ oder $g(H(t), H(u), H(v)) \neq 0$ gilt.

Bei einer Erhöhung der Anzahl der Freiheitsgrade kann die Menge dieser konsistenten Alternativen vergrößert werden, indem die Beziehung g für mehrere Tripel simultan getestet wird:

Man wählt Zahlen $0 < t < u_1 < \dots < u_n < v < K$ und betrachtet den Vektor:

$$\sqrt{N} \begin{bmatrix} g(\hat{S}(t), \hat{S}(u_1), \hat{S}(v)) \\ \vdots \\ g(\hat{S}(t), \hat{S}(u_i), \hat{S}(v)) \\ \vdots \\ g(\hat{S}(t), \hat{S}(u_n), \hat{S}(v)) \end{bmatrix}$$

der asymptotisch normalverteilt ist mit Mittelwert 0. Die asymptotische Kovarianzmatrix wird berechnet durch Anwendung der δ -Methode auf $\sqrt{N}(\hat{S}(t) - S(t), \dots, \hat{S}(u_1) - S(u_1), \dots, \hat{S}(v) - S(v))'$.

(Analog, wenn statt der Überlebensfunktion die kummulierte Hazardfunktion geschätzt wird.) Für die Fälle 4. und 5. müssen die Maschenweiten $u_1 - t, u_{i+1} - u_i, v - u_n$ gleich groß gewählt werden. Auf diese Weise überprüft man, ausgehend von den beiden "Stützpunkten" t und v , schrittweise das Intervall dazwischen. Konsistent ist ein derartiger Test gegen alle Alternativen, für die $g(S(t), S(u_i), S(v)) \neq 0$ resp. $g(H(t), H(u_i), H(v)) \neq 0$ für mindestens ein i . Jede Alternative, für die $g(S(t), S(u), S(v)) \neq 0$ resp. $g(H(t), H(u), H(v)) \neq 0$ für feste "Randwerte" t, v und variablem u dazwischen, kann so mit Hilfe eines hinreichend feinen Netzes aufgespürt werden. Allerdings müssen genügend Beobachtungen vorhanden sein: in jedem Intervall sollte mindestens eine Beobachtung liegen (sonst wäre auch die geschätzte Kovarianzmatrix singulär).

LITERATUR

Anderson, T.W., An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1958

Auinger, K., Spezifikationstests vom Hausmantyp für Übergangsratenmodelle, Forschungsbericht/Research Memorandum No.243, März 1988, Institut für Höhere Studien, Wien.

Billingsley, P., Convergence of Probability Measures, Wiley, New York, 1968

Breslow, N. and J. Crowley, A large sample study of the life table and the product limit estimates under random censorship, Ann.Stat., 2, 3, 437-453, 1974

Hall, P. and C.C. Heyde, Martingale Limit Theory and Its Applications, Academic Press, New York, 1980

Lawless, J.F., Statistical Models and Methods for Lifetime Data, Wiley, New York, 1982

Rao, C.R., Linear Statistical Inference and Its Applications, Wiley, New York, 1965