Robuste und effiziente Konfidenzbereiche für nichtzentrale Perzentile

Cornelius Gutenbrunner
Siemens Healthcare Diagnostics
Products GmbH
Postfach 1149
35001 Marburg
Cornelius.Gutenbrunner@siemens.com

Zusammenfassung

Für die Punkt- und Intervall-Schätzung randständiger Perzentile (p=0.01,0.02, 0.05, 0.95, 0.98, .99 usw.) bei Stichprobenumfängen n im Bereich 100-400 Beobachtungen ist es nicht einfach, gute Methoden zu finden: Die meistgenutzte nichtparametrische Methode (Stichprobenperzentile, X_p0 < X_p<X_p1, p0 und p1 zum gewünschten Konfidenzniveau verteilungsfrei konstruiert) bietet je nach n und p keine oder eine sehr ineffiziente und Ausreißer-empfindliche Lösung. Die gängige parametrische Methode (Schätzer der Form MW +/- k*SD, k auf Basis der Normalverteilung berechnet) ist nicht robust und versagt schnell, wenn die wahre Verteilungsform nicht der Normalverteilung entspricht.
Im Vortrag wird ein sehr einfaches allgemeines Konstruktionsprinzip für Punkt- und Intervallschätzer gezeigt, das gute Kompromisse zwischen beiden oben genannten Verfahren erlaubt. Die Bestimmung der Konstanten des Verfahrens erfolgt über Simulation, die mit einem sehr einfachen SAS-Makro durchgeführt werden kann.

Schlüsselwörter: Perzentile, Quantile, Konfidenzbereiche, mediantreue Schätzung, Simulation, Robustheit, Effizienz

1 Einleitung und Aufgabenstellung

Die Schätzung der Perzentile einer Verteilung ist eine häufig auftretende Aufgabe in vielen Anwendungsbereichen statistischer Methoden (siehe z.B. [3],[4]). Zu einer Verteilungsfunktion $p=F(x)$ bezeichne $x=Q(p)=F^{-1}(p)$ die zugehörige Perzentil- oder Quantilfunktion. Die Punkt- und Intervallschätzung zentraler Perzentile Q(p), p nahe an 0.5 ($0.1 \le p \le 0.9$) ist i.allg. unproblematisch. Dies gilt jedoch nicht für randständige Perzentile Q(p). Je näher p an 0 oder 1 liegt, desto größer muss der Stichprobenumfang n sein, um eine brauchbare Punkt- und erst recht Intervallschätzung (Konfidenzintervall) zu bekommen.

Es gibt hier bei den beiden bekanntesten Verfahren: 1. Nichtparametrische Schätzung und 2. Schätzung unter Normalverteilungsannahme auf der Basis von Mittelwert und Standardabweichung, eine für die Praxis empfindliche Lücke. So braucht man zum Beispiel mindestens n=368 Beobachtungen, um das nichtparametrische 95%-Konfidenzintervall für Q(0.01) oder Q(0.99) zu berechnen. Selbst wenn man diese 368

(oder etwas mehr) Beobachtungen zusammenbekommt, enthält das resultierende Konfidenzintervall als eine Grenze die kleinste (bei p=0.01) oder größte Beobachung (bei p=0.99) der Stichprobe. Wenn aber eine Stichprobe überhaupt Ausreißer enthält, dann ist mit hoher Wahrscheinlichkeit die kleinste oder größte Beobachtung davon betroffen! Auf der anderen Seite ist das Intervall [MW+k_1(p)*SD, MW+k_2(p)*SD], das man auf Normalverteilungsbasis berechnen kann, sehr empfindlich gegen Abweichungen von dieser. Anders als etwa im Falle des t-Tests kann man sich hier also nicht mit einer gewissen Robustheit des Verfahrens gegenüber Abweichungen von der Normalverteilung beruhigen.

Die vorliegende Präsentation soll zeigen, wie leicht man durch statistische Simulation zu Verfahren kommt, die man als klugen Kompromiss zwischen den oben dargestellten Extremen auffassen kann. Die erforderlichen Simulationen lassen sich durch sehr einfache SAS-Programme realisieren.

Um Missverständnissen vorzubeugen: Die Präsentation ist kein Plädoyer für unverantwortlich kleine Stichprobenumfänge im Zusammenhang mit der Schätzung randständiger Perzentile. Im Gegenteil, da man immer die dazugehörigen Konfidenzintervalle mitberechnen kann, erkennt man schon an deren (u.U. großer) Breite, welche Unsicherheit jeweils mit der Punktschätzung verbunden ist.

Als kleiner Nebeneffekt ergibt sich aus den angegebenen Verfahren auch eine einfache Möglichkeit, aus den gängigen Definitionen für Stichprobenperzentile (etwa SAS-PCTLDEF 1-5), die meist nicht Erwartungs- oder Median-treue Schätzer liefern, Perzentilschätzer mit diesen Eigenschaften zu konstruieren.

2 Konstruktionsprinzip der Schätzer

Die betrachteten Punktschätzer und Konfidenzintervallgrenzen sind von der Form

(1) $\hat{Q}(p) = \hat{M} + c(p, F_0)\hat{D}$

Hierbei ist $\hat{M}$ ein sich gegenüber Lage- und Skalentransformationen der Stichprobe äquivariant verhaltendes Lokationsmaß, wie etwa der Median der Stichprobe.

$\hat{D}$ ist ein sich gegenüber Lagetransformationen invariant und gegenüber Skalentransformationen äquivariant verhaltendes Dispersionsmaß, wie etwa der MAD (Median der absoluten Abweichungen vom Median) oder eine einfache Perzentildifferenz, wie z.B.

(2) $\hat{D} = \hat{X}(p) - \hat{X}(0.5)$,

$\hat{X}(p)$ das Stichprobenperzentil.

Der Faktor $c(p,F_0)$ hängt neben p auch vom Verteilungstyp F_0 ab, daher ist das Verfahren parametrisch. Es ist aber leicht, in kurzer Zeit mehrere Verteilungstypen und auch verschiedene Typen von Dispersionsmaßen in die Simulationen einzubeziehen, um festzustellen, welche Wahl des Dispersionsmaßes zu einer möglichst geringen F_0-Abhängigkeit des Faktors führt. Die Wahl des Dispersionsmaßes hängt auch davon ab, ob es sinnvoll erscheint, Symmetrie von F_0 vorauszusetzen oder nicht. Symmetrie der Verteilungen ist eine starke Voraussetzung, die oft in der Praxis nicht erfüllt ist, andererseits aber, wenn sie als gegeben angenommen werden kann, zu einer ungefähren Verdoppelung der Effizienz der Schätzer führt.

Für die Konstruktion von Konfidenzgrenzen und mediantreuen Punktschätzern muss man Wahrscheinlichkeiten der Art

(4) $P(\hat{Q}(p) \le Q(p))$

kontrollieren. Unter (1) ist (4) gleich

(5) $P(Z \le c(p,F_0))$

mit dem unter Lage- und Skalentransformationen invarianten Ausdruck

(6) $Z = (Q(p) - \hat{M}) / \hat{D}$.

Die Invarianz von *Z* bewirkt, dass man für jeden Verteilungstyp nur einen Vertreter dieses Typs simulieren muss. Die Aufgabe der Bestimmung der korrekten Konstanten *c* in (1) reduziert sich also auf die Bestimmung der Perzentile der Verteilung von *Z* unter dem ausgewählten Vereilungstyp F_0 . Die Bestimmung der Perzentile der Verteilung von *Z* wiederum ist sehr einfach durch Simulation zu bewerkstelligen.

3 Beispiele

Beispiel 1: Gamma-Verteilung
Aufgabe: Bei n=100 95%-Konfidenzintervall für Q(0.025) und Q(0.975) angeben.

Anm.: Für das nichtparametrische Intervall wäre n=146 minimaler Stichprobenumfang.

Wir wählen $\hat{M} = \hat{X}(0.5)$ (Stichprobenmedian) und $\hat{D} = \hat{X}(0.5) - \hat{X}(0.025)$ (für Q(0.025)) bzw. $\hat{D} = \hat{X}(0.975) - \hat{X}(0.5)$ (für Q(0.975)). Es ist hier genau die gewählte Definition der Stichprobenperzentile zu berücksichtigen, bei SAS also z.B. die gewählte PCTLDEF (=1, 2, 3, 4 oder 5). Für PCTLDEF=5 (default) und eine angenommene Gamma-Verteilung mit Parameter *a=3* der Beobachtungen ergibt Simulation mit 100000 Replikationen (ca. 30 Sekunden Rechenzeit) folgende Perzentile der Verteilung von *Z*:

Für Q(0.025): $\hat{Z}(0.025)=0.89,\ \hat{Z}(0.500)=1.01,\ \hat{Z}(0.975)=1.18.$

Für Q(0.975): $\hat{Z}(0.025)=0.71,\ \hat{Z}(0.500)=1.02,\ \hat{Z}(0.975)=1.42.$

Die mediantreuen Punktschätzer sind also

$\hat{Q}(0.025)=\hat{X}(0.5)+1.01(\hat{X}(0.025)-\hat{X}(0.5))$, $\hat{Q}(0.975)=\hat{X}(0.5)+1.02(\hat{X}(0.975)-\hat{X}(0.5))$,

die 95%-Konfidenzintervalle sind

Für Q(0.025): $[\hat{X}(0.5)+1.18(\hat{X}(0.025)-\hat{X}(0.5)),\hat{X}(0.5)+0.89(\hat{X}(0.025)-\hat{X}(0.5))]$

Für Q(0.975): $[\hat{X}(0.5)+0.71(\hat{X}(0.975)-\hat{X}(0.5)),\hat{X}(0.5)+1.42(\hat{X}(0.975)-\hat{X}(0.5))]$

Bei PCTLDEF=5 gilt $\hat{X}(0.025)=X_{(3)}$ (dritte Ordnungsstatistik) und entsprechend $\hat{X}(0.975)=X_{(98)}$ (zu den PCTLDEF's vgl. auch [1]).

Hätte man PCTLDEF=4 gewählt, so wäre z.B. $\hat{X}(0.025)=0.475X_{(2)}+0.525X_{(3)}$ und es hätten sich für Q(0.025) die Werte $\hat{Z}(0.025)=0.87,\ \hat{Z}(0.500)=0.98,\ \hat{Z}(0.975)=1.14$ ergeben, somit als mediantreuer Schätzer $\hat{Q}(0.025)=\hat{X}(0.5)+0.98(\hat{X}(0.025)-\hat{X}(0.5))$ und als Konfidenzintervall $[\hat{X}(0.5)+1.14(\hat{X}(0.025)-\hat{X}(0.5)),\hat{X}(0.5)+0.87(\hat{X}(0.025)-\hat{X}(0.5))]$.

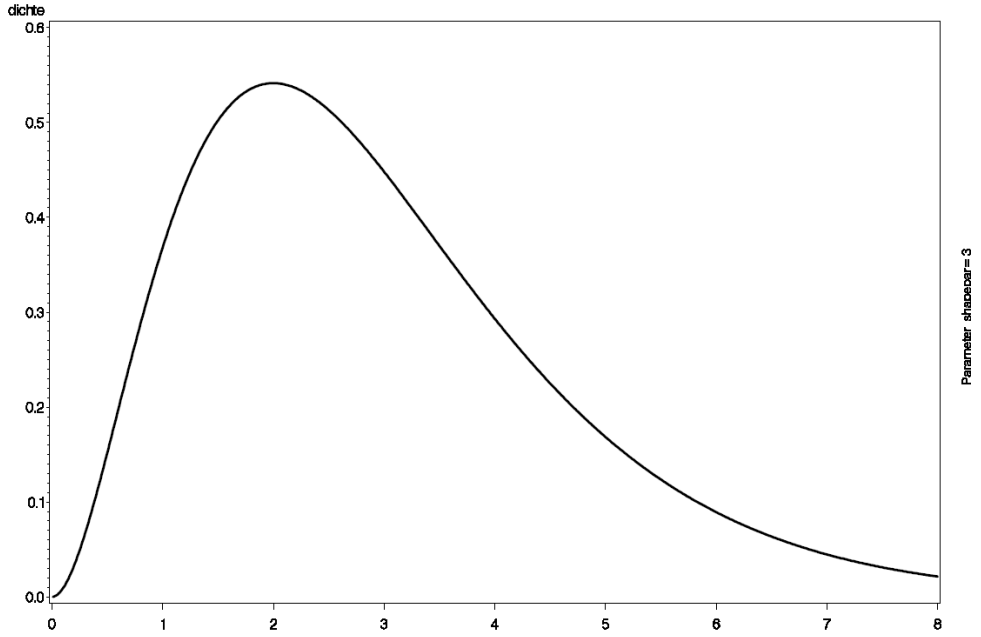

Abbildung 1: Dichte der Gamma-Verteilung mit Parameter a=3

Beispiel 2: t-Verteilung
Aufgabe analog zu Beispiel 1

Als Beispiel für eine symmetrische Verteilung mit starken Flanken wählen wir die t-Verteilung mit 2.345 Freiheitsgraden. Diese Wahl der Freiheitsgrade führt gerade zu einer Tailstärke 2.00 im Sinne von [2] (Tailstärke $t_{0.05,0.15}$=(Q(0.95)-Q(0.05))/(Q(0.85)-Q(0.15)); Gauß-Verteilung: Tailstärke 1.59; Cauchy-Verteilung: Tailstärke 3.22). Tatsächlich ist die t-Verteilung mit 2.345 Freiheitsgraden schon recht nahe an der Cauchy-Verteilung: Ihr zweites absolutes Moment $E|X|^2$ ist noch endlich, aber bereits das 2.5-te Moment $E|X|^{2.5}$ ist unendlich. Analog wie in Beispiel 1 erhalten wir für n=100 als Konfidenzintervall für Q(0.975):

$$[\,\hat{X}(0.5)+0.53(\hat{X}(0.975)-\hat{X}(0.5))\,,\,\hat{X}(0.5)+1.70(\hat{X}(0.975)-\hat{X}(0.5))\,],$$

als mediantreuen Punktschätzer

$$\hat{Q}(0.975)=\hat{X}(0.5)+1.03(\hat{X}(0.975)-\hat{X}(0.5))\,.$$

4 Programmierung in SAS

Code des SAS-Programms für obiges Beispiel:

```
%macro CI_for_Pctl_Gamma_Dist(
     p=0.025,
     n=100,
     shapepar=3,
     rep=10000,
     seed=38642159,
     out=tmp);

* evtl. schon vorh. Tabelle _d in work library löschen;
PROC DATASETS nolist LIB=work;DELETE  _d   /MEMTYPE=DATA;RUN;QUIT;

data _d;
array zz z1-z&rep;
array xx x1-x&n;

* number of replicates for simulation. 100000 is recommended;
rep=&rep;

*number of measurements per sample;
n=&n;

*Parameter of Gamma-distribution,-1 refers to Gaussian distribution;
shapepar=&shapepar;
*Percentage of Percentile;
p=&p;p100=100*p;
```

```
*true Percentile;
TruePctl=probit(&p);
if shapepar>0 then TruePctl=gaminv(&p,shapepar);

do i1=1 to rep;* Replications for simulation;

    do i3=1 to n;

        xx[i3]=rannor(&seed);**Gaussian random numbers;

        **transforming to gamma-distribution if shapepar>0;
        if shapepar>0 then do;
            xx[i3]=gaminv(probnorm(xx[i3]),shapepar);
        end;
     end;

    x_p=pctl(p100,of x1 - x&n);
    x_50=pctl(50 ,of x1 - x&n);

    zz[i1]=(TruePctl-x_50)/(x_50-x_p);

end;

z_025=pctl( 2.5,of z1 - z&rep);
z_500=pctl(50  ,of z1 - z&rep);
z_975=pctl(97.5,of z1 - z&rep);
run;

data &out;set _d;run;

proc print data=_d;
    var z_025 z_500 z_975 ;
    format z_025 z_500 z_975 8.2;
    label
        z_025='Factor_for_UL_of_95CI_Q(p)'
        z_500='Factor_for_Median_Unbiased_Q(p)'
        z_975='Factor_for_LL_of_95CI_Q(p)'
    ;
title"CI_for_Pctl_Gamma_Dist(p=&p,n=&n,shapepar=&shapepar,rep=&rep,
seed=&seed,out=&out)";
run;
%mend CI_for_Pctl_Gamma_Dist;

/*
Sample call of macro:
%CI_for_Pctl_Gamma_Dist(p=0.025,n=100,shapepar=3,rep=100000,
seed=38642159,out=tmp);
*/
```

Literatur

[1] H. Stürzl, C. Gutenbrunner: SAS Makro UNISTATS 2.0. 14.KSFE 2010 Berlin, U Rendtel, P Schirmbacher, O Kao, W.F. Lesener, R. Minkenberg (Hrsg.). Shaker Verlag, Aachen, 2010.

[2] W. Kössler, W. Lesener: Adaptive Lokationstests mit U-Statistiken. 14.KSFE 2010 Berlin, U Rendtel, P Schirmbacher, O Kao, W.F. Lesener, R. Minkenberg (Hrsg.). Shaker Verlag, Aachen, 2010.

[3] CLSI Guideline C28-A3: Clinical and Laboratory Standards Institute (CLSI). Defining, Establishing and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline—Third Edition. CLSI document C28-A3 (ISBN 1-56238-682-4). CLSI, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898 USA, 2008.

[4] CLSI Guideline EP17-A:Protocols for Determination of Limits of Detection and Limits of Quantitation; Approved Guideline. CLCI document EP17-A (ISBN 1-56238-551-8). CLSI, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898 USA, 2004.