Konfidenzintervalle für Raten und Korrelationskoeffizienten: zwei SAS-Macros

Friederike Rohlmann, Rainer Muche

Abteilung Biometrie und Medizinische Dokumentation, Universität Ulm
Telefon: 0731 / 50-26903
eMail: friederike.rohlmann@medizin.uni-ulm.de

Abstract

In dem Beitrag werden zwei SAS-Macros vorgestellt, in denen Konfidenzintervalle für die bei der statistischen Auswertung häufig benutzten Schätzer Rate und Korrelationskoeffizient berechnet werden. In den entsprechenden SAS-Prozeduren PROC FREQ (Rate) und PROC CORR (Korrelationskoeffizient) gibt es keine Option, diese Konfidenzintervalle direkt anzufordern. Da in der Biometrie und Epidemiologie die Angabe von Konfidenzintervallen immer häufiger gewünscht wird, werden die Berechnungen mittels SAS-Macros automatisiert.

Einleitung

Seit einigen Jahren gibt es in der Biometrie und insbesondere in der Epidemiologie Empfehlungen und Diskussionen, in statistischen Auswertungen eher die Konfidenzintervalle für die Schätzwerte interessierender Parameter als "nur" p-Werte der entsprechenden statistischen Tests anzugeben. [1-3].

Ein Konfidenzintervall gibt einen Bereich um den Schätzer eines Parameters an, in dem der wahre, unbekannte Parameter mit vorgegebener Wahrscheinlichkeit liegt. Man erhält somit zusätzlich zum Schätzer eine Bereichsangabe, die es ermöglicht, die Variabilität der Ergebnisse auf Grundlage der erhobenen Daten zu beurteilen. Da die meisten statistischen Tests direkt mit Konfidenzintervallen zusammenhängen und deren Ergebnis daraus abgelesen werden kann, verliert man nichts an Aussagekraft.

Wichtige und häufig benutzte Schätzer in der Biometrie und Epidemiologie sind Raten und Korrelationskoeffizienten. In den entsprechenden SAS-Prozeduren PROC FREQ bzw. PROC CORR gibt es keine Möglichkeit, Konfidenzintervalle berechnen zu lassen.

Dieser Beitrag soll u.a. ein Plädoyer dafür sein, dass die Konfidenzintervalle der häufig benutzten statistischen Kenngrößen in die entsprechenden SAS-Prozeduren integriert werden.

Konfidenzintervall für den Pearson-Korrelationskoeffizienten:

Im folgenden SAS-Macro wird ein Konfidenzintervall für den Pearson-Korrelations-koeffizienten berechnet. Mittels PROC CORR kann dies nicht ausgegeben werden, obwohl die Berechnung relativ einfach ist und in fast allen Statistiklehrbüchern dokumentiert ist. Wir haben hier die asymptotische Formel aus dem Buch von Zöfel [4] zugrunde gelegt.

Die Grenzen des Konfidenzintervalles werden folgendermaßen bestimmt:

$$CU = \mathbf{tanh}\left(\left(\frac{\mathbf{1}}{\mathbf{2}} \bullet \mathbf{ln}\frac{\mathbf{1}+r}{\mathbf{1}-r}\right) - \left(\frac{t_{\alpha,n-2}}{\sqrt{n-\mathbf{3}}}\right)\right)$$

$$CO = \mathbf{tanh}\left(\left(\frac{\mathbf{1}}{\mathbf{2}} \bullet \mathbf{ln}\frac{\mathbf{1}+r}{\mathbf{1}-r}\right) + \left(\frac{t_{\alpha,n-2}}{\sqrt{n-\mathbf{3}}}\right)\right)$$

Der Korrelationskoeffizient r wird dabei transformiert in eine asymptotisch normalverteilte Zufallsvariable z, für die ein asymptotisches, zweiseitiges Intervall bestimmt wird. Durch Rücktransformation mit dem tangens hyperbolicus (tanh) können die Grenzen für den Korrelationskoeffizienten angegeben werden.

Zur Berechnung der Konfidenzgrenzen müssen dem Macro CI_CORR.MAC folgende Parameter übergeben werden:

- OUT: SAS-Datei, in die die Ergebnisse (und Zwischenergebnisse) geschrieben werden
- R: Korrelationskoeffizient, um den das Konfidenzintervall berechnet werden soll
- CP: Konfidenzwahrscheinlichkeit in %
- N: Anzahl der Wertepaare, die zur Berechnung von R zur Verfügung stehen
- DEC: Anzahl Nachkommastellen im Output

SAS-Macro CI_CORR.MAC:

```
*******************************************************************************;
* BENOETIGTE FORMELN:                                                         *;
*                                                                             *;
* z = 1/2 * ln (1+r)/(1-r)                                                    *;
* c = t(alpha,df) / (n-3)**1/2   mit df(Freiheitsgrade) = n-2                 *;
* z0 - c < z < z0 + c                                                         *;
* r = (e**2z - 1) / (e**2z + 1) = tanh z   (Tangenz hyperbolicus von z)       *;
*                                                                             *;
*******************************************************************************;
%macro ci_corr(out= , r= , cp= , n= , dec=);

* out -> Dataset-Name fuer das CI                                  *;
* r   -> Korrelationskoeffizient                                   *;
* cp  -> Konfidenzwahrscheinlichkeit in % (confidence probability) *;
* n   -> Anzahl der Wertepaare                                     *;
* dec -> gewünschte Anzahl Nachkommastellen im Output              *;


data &out;
  retain r cp alpha n r_lb2 r_ub2 t2 c2 z02 z_lb2 z_ub2;

  r=&r;
  cp=&cp;
  alpha=(1-&cp/100);
  n=&n;

  t=tinv((1-alpha/2),n-2);    * t-Wert aus der t-Verteilung *;
  c=t/((n-3)**0.5);
  z0=0.5*log((1+r)/(1-r));    * log: natuerl. Log.*;
  z_lb=z0-c;                  * z lower bound *;
  z_ub=z0+c;                  * z upper bound *;
  r_lb=tanh(z_lb);            * r lower bound *;
  r_ub=tanh(z_ub);            * r upper bound *;

  t2=round(t,10**(-&dec));
  c2=round(c,10**(-&dec));
  z02=round(z0,10**(-&dec));
  z_lb2=round(z_lb,10**(-&dec));
  z_ub2=round(z_ub,10**(-&dec));
  r_lb2=round(r_lb,10**(-&dec));
  r_ub2=round(r_ub,10**(-&dec));

  label r='r'
        cp='Confidence Probability (%)'
        alpha='Alpha'
        n='N'
        t2='t(alpha;df)'
        c2='c'
        z02='z-Wert(r)'
        z_lb2='CI(z) lower bound'
        z_ub2='CI(z) upper bound'
        r_lb2='CI(r) lower bound'
        r_ub2='CI(r) upper bound'
        ;
  keep r cp alpha n t2 c2 z02 z_lb2 z_ub2 r_lb2 r_ub2;
run;

%mend ci_corr;
```

Nachfolgend wird ein beispielhafter Aufruf mit entsprechendem Output gezeigt. Dabei ist r der eingegebene Pearson-Korrelationskoeffizient, N die Anzahl der Wertepaare. CI (r) lower und upper bound geben die Grenzen des Konfidenzintervalles zur Confidence probability an. Die weiteren Werte sind Zwischenwerte aus den Formeln.

Beispiel:

```
** BSP-AUFRUF **;
*  %ci_corr(out=one, r=0.443, cp=95, n=20, dec=3);
*  options pageno=1 nodate nocenter ls=96 ps=64;
*  proc print data=one noobs label; run;
```

Output-Datei:

r	Confidence Probability (%)	Alpha	N	CI(r) lower bound	CI(r) upper bound	t(alpha;df)	c	z-Wert(r)	CI(z) lower bound	CI(z) upper bound
0.443	95	0.05	20	-0.034	0.755	2.101	0.51	0.476	-0.034	0.986

Konfidenzintervall für Raten:

Im nächsten SAS-Macro wird ein Konfidenzintervall für Raten berechnet. In PROC FREQ kann dies nicht angefordert werden. Wir haben hier die Formel für zweiseitige exakte Konfidenzgrenzen umgesetzt, die z.B. im Buch von Sachs [5] zu finden ist. Weitere Formeln und ein Vergleich finden sich in dem Artikel von Newcombe [6].

Die Grenzen des Konfidenzintervalles werden folgendermaßen bestimmt:

$$CU = \frac{x}{x + (n - x + 1) \cdot F_{1-\alpha/2,\,2(n-x+1),\,2x}}$$

$$CO = \frac{(x+1) \cdot F_{1-\alpha/2,\,2(x+1),\,2(n-x)}}{n - x + (x+1) \cdot F_{1-\alpha/2,\,2(x+1),\,2(n-x)}}$$

wobei . x : Anzahl Treffer
n : Anzahl Beobachtungen
F : Wert der F-Verteilung mit entsprechenden Freiheitsgraden

Entsprechend der Werte, die man in den Formeln braucht, sind auch die Parameter dem Macro zu übergeben:

OUT:	SAS-Datei, in die die Ergebnisse geschrieben werden
X:	Anzahl Treffer
N:	Anzahl Beobachtungen
CP:	Konfidenzwahrscheinlichkeit in %
PERCENT:	Angabe prozentuale Häufigkeiten gewünscht? 1=ja, 0=nein
DEC:	Anzahl Nachkommastellen im Output

Im Macro sind einige Besonderheiten abgefangen:

- Ist die Anzahl Treffer gleich der Anzahl Beobachtungen, kann der F-Wert für die obere Konfidenzgrenze nicht bestimmt werden (Anzahl Freiheitsgrade=0). Der oberen Grenze wird 1 bzw. 100% zugewiesen.

- Für den Fall, daß die Anzahl der Treffer gleich 0 und die Anzahl der Beobachtungen größer 0 ist, kann der F-Wert für die untere Konfidenzgrenze nicht bestimmt werden (Anzahl Freiheitsgrade =0). Die untere Grenze wird auf 0 bzw. 0% gesetzt.
- Für den Fall, daß die Anzahl der Treffer 0 und die Anzahl der Beobachtungen 0 ist, werden keine Werte berechnet.

SAS-Macro CI_FREQ.MAC:

```
*************************************************************************************;
* BENOETIGTE FORMELN:                                                               *;
* CI: p_lb <= p <= p_ub                                                             *;
* p_lb = x      / x+(n-x+1)F  mit F (df1=2(n-x+1), df2=2x)  F=F-Wert                *;
*                                                          df=degrees of freedom*;
* p_ub = (x+1)F / n-x+(x+1)F  mit F (df1=2(x+1), df2=2(n-x))                        *;
*************************************************************************************;
%macro ci_freq(out= , x= , n= , cp= , percent=, dec=);

* out     -> Dataset-Name fuer das CI                                    *;
* x       -> Anzahl Treffer                                             *;
* n       -> Gesamtzahl                                                 *;
* cp      -> Konfidenzwahrscheinlichkeit in % (confidence probability) *;
* percent -> Angabe prozentualer Häufigkeiten gewünscht? 1=ja 0=nein   *;
* dec     -> gewünschte Anzahl Nachkommastellen im Output              *;

data &out;
  retain x n p2 cp alpha p_lb2 p_ub2 f_l2 f_u2;

  x=&x;
  n=&n;
  if n ne 0 and &percent=1 then p=x/n*100;
    else if n ne 0 and &percent ne 1 then p=x/n;   * relative (prozentuale)
Häufigkeit *;
  cp=&cp;  alpha=(1-&cp/100);

  df_l1=2*(n-x+1);    * Freiheitsgrade *;
  df_l2=2*x;
  df_u1=2*(x+1);
  df_u2=2*(n-x);

  if x ne 0 then
  f_l=finv((1-alpha/2),df_l1,df_l2);    * F-Werte aus der F-Verteilung bei
Freiheitsgraden 1 und 2 *;
  if n ne x then
  f_u=finv((1-alpha/2),df_u1,df_u2);

  if &percent=1 then do;
    if x=0 and n ne 0 then p_lb=0;
      else p_lb=(x/(x+(n-x+1)*f_l))*100;             * lower bound *;
    if n=x and n ne 0 then p_ub=100;
      else p_ub=(((x+1)*f_u)/(n-x+(x+1)*f_u))*100;   * upper bound *;
    end;
    else do;
    if x=0 and n ne 0 then p_lb=0;
     else p_lb=x/(x+(n-x+1)*f_l);                    * lower bound *;
    if n=x and n ne 0 then p_ub=1;
      else p_ub=((x+1)*f_u)/(n-x+(x+1)*f_u);         * upper bound *;
    end;

  p_lb2=round(p_lb,10**(-&dec));
  p_ub2=round(p_ub,10**(-&dec));
  p2   =round(p,   10**(-&dec));
  f_l2 =round(f_l, 10**(-&dec));
  f_u2 =round(f_u, 10**(-&dec));
  label x='x (Counted hits)'
        cp='Confidence Probability (%)'
        alpha='Alpha'
        n='N'
        f_l2='F-Value lower bound'
        f_u2='F-Value upper bound'
        p_lb2='CI(p) lower bound'
        p_ub2='CI(p) upper bound'
        ;

keep x n p2 cp alpha f_l2 f_u2 p_lb2 p_ub2;

  %if &percent=1 %then label p2="Frequency (%)";
                 %else label p2="Relative frequency";;
run;
%mend ci_freq;
```

Im Anschluß wird für die oben beschriebenen Situationen beispielhaft jeweils das Macro benutzt und die resultierenden Ergebnisse dokumentiert.

Beispiele:

```
** BSP-AUFRUF **;
  %ci_freq(out=one,  x=7, n=20, cp=95, percent=0, dec=3);
  %ci_freq(out=two,  x=7, n=20, cp=95, percent=1, dec=1);

  %ci_freq(out=three,x=132, n=132, cp=95, percent=0, dec=3);
  %ci_freq(out=four, x=132, n=132, cp=95, percent=1, dec=1);

  %ci_freq(out=five, x=0, n=132, cp=95, percent=0, dec=3);
  %ci_freq(out=six,  x=0, n=132, cp=95, percent=1, dec=1);

  %ci_freq(out=seven,x=0, n=0, cp=95, percent=0, dec=3);

  options pageno=1 nodate nonumber;
  proc print data=one label noobs; run;
  proc print data=two label noobs; run;
  proc print data=three label noobs; run;
  proc print data=four label noobs; run;
  proc print data=five label noobs; run;
  proc print data=six label noobs; run;
  proc print data=seven label noobs; run;
```

Output-Dateien:

x (Counted hits)	N	Relative frequency	Confidence Probability (%)	Alpha	CI(p) lower bound	CI(p) upper bound	F-Value lower bound	F-Value upper bound
7	20	0.35	95	0.05	0.154	0.592	2.749	2.36

x (Counted hits)	N	Frequency (%)	Confidence Probability (%)	Alpha	CI(p) lower bound	CI(p) upper bound	F-Value lower bound	F-Value upper bound
7	20	35	95	0.05	15.4	59.2	2.7	2.4

x (Counted hits)	N	Relative frequency	Confidence Probability (%)	Alpha	CI(p) lower bound	CI(p) upper bound	F-Value lower bound	F-Value upper bound
132	132	1	95	0.05	0.972	1	3.741	.

x (Counted hits)	N	Frequency (%)	Confidence Probability (%)	Alpha	CI(p) lower bound	CI(p) upper bound	F-Value lower bound	F-Value upper bound
132	132	100	95	0.05	97.2	100	3.7	.

x (Counted hits)	N	Relative frequency	Confidence Probability (%)	Alpha	CI(p) lower bound	CI(p) upper bound	F-Value lower bound	F-Value upper bound
0	132	0	95	0.05	0	0.028	.	3.741

x (Counted hits)	N	Frequency (%)	Confidence Probability (%)	Alpha	CI(p) lower bound	CI(p) upper bound	F-Value lower bound	F-Value upper bound
0	132	0	95	0.05	0	2.8	.	3.7

x (Counted hits)	N	Relative frequency	Confidence Probability (%)	Alpha	CI(p) lower bound	CI(p) upper bound	F-Value lower bound	F-Value upper bound
0	0	.	95	0.05	.	.	.	.

Literatur

[1] D.S. Salsburg: The religion of statistics as practiced in medical journals
The American Statistician 39 (1985), S. 220-223

[2] M.J. Schervish: p-values: what they are and what they are not.
The American Statistician 50 (1996), S. 203-206

[3] D.S. Salsburg: The use of statistical methods in the analysis of clinical studies
J. Clin. Epidemiol. 46 (1993), S. 17-27

[4] P. Zöfel: Statistik in der Praxis
UTB 1293, Gustav Fischer Verlag, Stuttgart (1985), S. 214ff

[5] L. Sachs: Angewandte Statistik, 6. Auflage.
Springer Verlag, Heidelberg (1984), S. 258ff

[6] R.G. Newcombe: Two-sided confidence intervals for the single proportion:
comparison of seven methods. Stat. in Med. 17 (1998), S. 857-872

Die Macros können von den Autoren angefordert werden. Bitte dazu eine e-Mail (s. Überschrift) an die Autoren senden.