

SAS® Macros for Exact Confidence Limits

Richard L. Berg, Marshfield Medical Research Foundation, Marshfield, WI

Abstract

Confidence intervals have a long history and have become an indispensable tool in statistical applications. The computation of confidence intervals for a mean based on normal (large sample) theory is straightforward and readily available in the SAS system through such procedures as PROC ANOVA and PROC GLM. For the case of small samples and non-normal distributions, however, normal theory based confidence intervals may provide a poor approximation to exact results. This paper presents a system of SAS macros for exact results with binomial, hypergeometric, or Poisson data, as well as nonparametric confidence intervals for the percentiles of a continuous distribution.

Introduction

Confidence limits for the parameters of commonly encountered discrete distributions have long been available through tables, at least for select parameter values (Clopper and Pearson 1934, Chung and DeLury 1950, Pearson and Hartley 1954, Lieberman and Owen 1961). In practice, approximations based on large sample theory are often relied upon, but have well recognized limitations (Ghosh 1979, Blyth 1986, Samuels and Lu 1992). Similarly, confidence limits for continuous but non-normal data may be based upon a normal approximation, perhaps on a transformed scale.

Probability functions available within the SAS system allow direct calculation of certain exact confidence limits. In the following sections, we discuss the general confidence interval problem and describe a system of SAS macros which provides exact limits for the parameters of the binomial, hypergeometric, and Poisson distributions, as well as nonparametric limits for the percentiles of a continuous distribution.

Definition

By definition, if $P(L \leq \theta \leq U) \geq 1 - \alpha$ then (L, U) is a confidence interval for θ with confidence coefficient at least $100 * (1 - \alpha) \%$. For the case of the binomial parameter p , estimated by the sample proportion x/n , the standard Clopper-Pearson limits are defined as:

$$L = p: \sum_{i=x}^n \binom{n}{i} p^i (1-p)^{n-i} = \frac{\alpha}{2}$$
$$U = p: \sum_{i=0}^x \binom{n}{i} p^i (1-p)^{n-i} = \frac{\alpha}{2}$$

Figure 1 illustrates this approach, which simply excludes values of p which are so low or so high as to show little probability of resulting in samples like that observed.

It is this Clopper-Pearson approach which is used in the macros described below; it is straightforward to program and guarantees the nominal confidence level. However, for discrete distributions, the Clopper-Pearson type interval can be quite conservative for small n , and a number of alternatives have been described (Blyth and Still 1983, Angus and Schafer 1984, Vollset 1993, Edwards 1994). In applications where the conservativeness presents a problem, other methods may be preferred.

SAS Macros

Discrete Distributions

The macro CIDISCRT (Appendix A) calculates exact confidence limits for binomial, Poisson, or hypergeometric distributions. Three positional parameters specify the distribution, sample size, and number observed in the sample. An additional parameter is required for the hypergeometric case to specify the finite population size. Optional parameters specify the desired confidence level and the number of decimal places for formatting the output. The macro calculates two-sided intervals, but one-sided intervals may be obtained simply by changing the confidence level.

For the binomial and Poisson cases, the macro relies on relations (Hogg and Craig 1978) with the beta and gamma distribution functions, respectively, through the SAS functions BETAINV and GAMINV. For the hypergeometric case, the macro achieves a solution iteratively using the SAS function PROBHYP.

Continuous Distributions

The macro CCONTIN (Appendix B) calculates

Figure 1

Clopper-Pearson 95% Confidence Limits for Binomial p
5 Successes Observed in 10 Trials

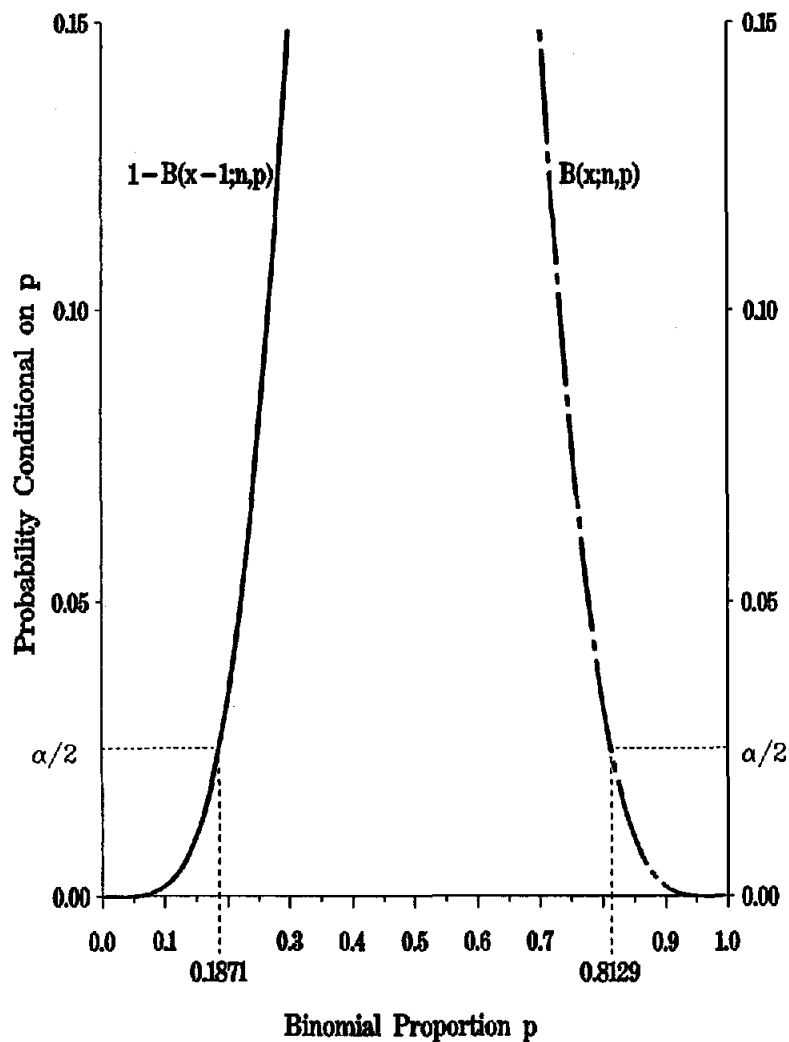
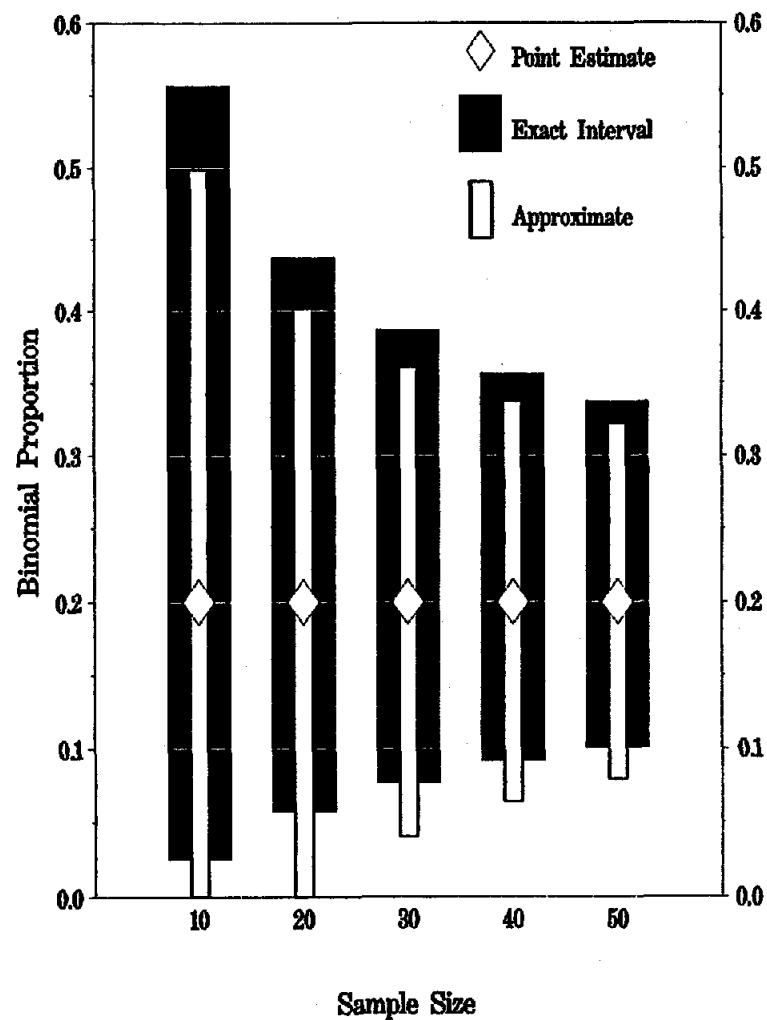


Figure 2

Exact and Approximate 95% Confidence Limits for Binomial p
By Sample Size for $\hat{p} = 0.2$



nonparametric confidence limits for a percentile of a continuous distribution. A single positional parameter specifies the variable of interest, and optional keyword parameters set the confidence level (95% by default), the percentile of interest (median by default), and allow calculation of lower, upper, or two-sided limits. Calculations are based upon the binomial distribution (Guenther 1973, Lehmann 1975, Campbell and Gardner 1988) through the SAS function BETAINV.

Example 1

For the binomial case with 5 successes observed in a sample of size 10, the command syntax to obtain a 95% confidence interval with four decimal places in the output would be the following:

```
%CIDISCRT(B,10,5,DEC=4);
```

and the output would appear as shown below.

95% Confidence Limits for Binomial p n = 10 x = 5

Estimate	Lower Limit	Upper Limit
0.5000	0.1871	0.8129

These results match the limits labelled "MAX" on Page 817 of Vollset (1993).

In Figure 2, the results from the macro have been displayed with SAS GRAPH together with results from a normal approximation (Snedecor and Cochran 1967). Results are presented for a binomial distribution with sample sizes from 10 to 50 and estimated proportion = 0.2.

Example 2

The use of the macro CICONIN to obtain confidence limits on a percentile is illustrated with an example from Campbell and Gardner (1988). Eleven observations are read into a data set and 95% confidence limits for the median are calculated with the statements shown below.

```
DATA TEST;
  INPUT CONC @@;
CARDS;
66 71.2 83 83.6 101 107.6 122 143 160 177 414
;
%CICONIN(CONC,TITL4=
  Exact Confidence Interval for the Median);
```

The resulting output follows.

95% Confidence Limits for 50th Percentile of CONC

Exact Confidence Interval for the Median

N	Estimate	Lower Limit	Upper Limit	Actual Confidence Level
11	107.6	71.2	177.0	98.8%

Note that, since the nonparametric results are based on binomial probabilities, the confidence level obtained will generally be somewhat larger than the nominal level.

Conclusion

The importance of confidence intervals in statistical applications is apparent in the wealth of literature on the subject. Increasing access to high speed computing has eliminated much of the need for cumbersome probability tables or reliance on approximate results. In particular, with the features of the SAS System, notably the MACRO facility and excellent probability functions, exact confidence limits are easily obtained.

References

- Angus, J.E. and Schafer, R.E. (1984). Improved confidence statements for the binomial parameter. *The American Statistician* 38, 189-191.
- Blyth, C.R. and Still, H.A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association* 78, 108-116.
- Blyth, C.R. (1986). Approximate binomial confidence limits. *Journal of the American Statistical Association* 81, 843-855.
- Campbell, M.J. and Gardner, M.J. (1988). Calculating confidence intervals for some non-parametric analyses. *British Medical Journal* 296, 1454-1456.
- Chung, J.H. and De Lury, D.B. (1950). *Confidence Limits for the Hypergeometric Distribution*. Univ. of Toronto Press, Toronto, Canada.
- Clopper, C.J. and Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 404-413.
- Edwardes, M. (1994). Confidence intervals for a

binomial proportion (letter to the editor with reply).
Statistics in Medicine 13, 1693-1698.

Ghosh, B.K. (1979). A comparison of some
approximate confidence intervals for the binomial
parameter. Journal of the American Statistical
Association 74, 894-900.

Guenther, W.C. (1973). Concepts of Statistical
Inference (2nd ed.), McGraw-Hill, New York.

Hogg, R.V. and Craig, A.T. (1978). Introduction to
mathematical statistics (4th edition), Macmillan
Publishing Co., New York.

Lehman, E.L. (1975). Statistical Methods Based on
Ranks. Holden-Day, San Francisco.

Lieberman, G.J. and Owen, D.B. (1961). Tables of
the Hypergeometric Probability Distribution. Stanford
Univ. Press, Stanford CA.

Pearson, E.S. and Hartley, H.O. (1954). Biometrika
Tables for Statisticians Vol. 1, Cambridge University
Press, Cambridge MA.

Samuels, M.L. and Lu, T-F. (1992). Sample size
requirements for the back-of-the-envelope binomial
confidence interval. The American Statistician 46, 228-
231.

Snedecor, G.W. and Cochran, W.G. (1967). Statistical
Methods (6th ed.) Iowa State Univ. Press, Ames, Iowa.

Vollset, S.E. (1993). Confidence intervals for a
binomial proportion. Statistics in Medicine 12, 809-
824.

SAS is a registered trademark or trademark of SAS
Institute, Inc. in the USA and other countries.

® indicates USA registration.

Contact Information:

Richard L. Berg, MS
Marshfield Medical Research Foundation
1000 N. Oak Avenue
Marshfield, WI 54449

bergr@mflclin.edu

Appendix A: SAS Macro CIDISCRT

Macro to compute exact confidence limits for binomial, poisson,
or hypergeometric distributions

%MACRO CIDISCRT(DIST,N,X,NPOP=., DEC=2,CL_PC=95);

3 Positional parameters must be specified:

DIST = B, P, or H to specify distribution

N = sample size

X = number observed

3 Keyword parameters (with defaults):

NPOP = size of population, required only for hypergeometric

DEC = number of decimals for estimates (2)

CL_PC = confidence level as % (95)

%GLOBAL TITL1 TITL2;

%LET ERR=0;

data ci;

a2 = (100-&CL_PC)/200;

%IF %UPCASE(&DIST)=B %THEN %DO;

%LET TITL1=Binomial p;

%LET TITL2=;

est = &X/&N;

if &X=&N then upper=1;

else upper=betainv(1-a2,&X+1,&N-&X);

if &X=0 then lower=0;

else lower=betainv(a2,&X,&N-&X+1);

%END;

%ELSE %IF %UPCASE(&DIST)=P %THEN %DO;

%LET TITL1=Poisson lambda;

%LET TITL2=;

est = &X/&N;

if &X=0 then lower=0;

else lower = gaminv(a2,&X)/&N;

upper = gaminv(1-a2,&X+1)/&N;

%END;

%ELSE %IF %UPCASE(&DIST)=H %THEN %DO;

%LET TITL1=Hypergeometric p;

%LET TITL2=%STR(N = &NPOP);

est=&X/&N;

if &NPOP ge &N then do;

khat=&NPOP*est;

khat1=int(khat);

if khat=khat1 then khat2=khat1;

else khat2=khat1+1;

alph1=probypr(&NPOP,khat1,&N,&X);

do while (alph1 > a2 and khat1 > &X);

khat1 = khat1 - 1;

alph1 = 1 - probypr(&NPOP,khat1,&N,&X-1);

end;

alph2=probypr(&NPOP,khat2,&N,&X);

do while (alph2 > a2 and khat2 < &NPOP-(&N-&X));

khat2 = khat2 + 1;

alph2 = probypr(&NPOP,khat2,&N,&X);

end;

lower=khat1/&NPOP;

upper=khat2/&NPOP;

end;

else do;

```

error // '** ERROR ** Specify population size NPOP at least'
/ '** ERROR ** as large as sample size. '//;
call symput('ERR','1');
end;
%END;

%ELSE %THEN %DO;
error // '** ERROR ** Specify distribution as B, P, or H' //;
call symput('ERR','1');
%END;

keep est lower upper;
run;

%IF &ERR=0 %THEN %DO;
proc print noobs label;
title "&CL_PC.% Confidence Limits for &TITL1";
title2 "n = &N x = &X &TITL2";
var est lower upper;
label est='Estimate' lower='Lower Limit' upper='Upper Limit';
format est lower upper 7.&DEC;
run;
%END;

%MEND CIDISCRT;
*****;

```

Appendix B: SAS Macro CCONTIN

```

*****
Macro which computes confidence limits for a
percentile of a continuous distribution
*****;

%MACRO CCONTIN(ANALVAR,DS=_LAST_,P=50,
CL_PC=95,SIDED=TWO,BYVAR=.,SORT=S,DEF=4,
VALFMT=6.1,TITL3=,TITL4=,TITL5=
,NAPPROX=50);

```

MACRO Parameters

1 Positional Parameter (must be specified in call):

ANALVAR = analysis variable

Keyword Parameters (with defaults):

DS = SAS data set name (_LAST_)

P = percentile of interest (50, the median)

CL_PC = confidence level as % (95)

SIDED = Lower, Upper, or Two-sided (T)

BYVAR = stats are by this categorical variable (' ')

SORT = 1 if DS not already sorted by BYVAR (1)

DEF = PCTLDEF for PROC UNIVARIATE (4)

TITL3 - TITL4 = optional titles

VALFMT = variable format (6.1)

NAPPROX = minimum sample size for
approximate results (50)

```

*****;
%LET SIDE=%UPCASE(%SUBSTR(&SIDED,1,1));
%IF &SIDE^=T %THEN %LET SIDETXT=%STR(One-sided);
%ELSE %LET SIDETXT=;
%LET ERR=0;

data ci_ ; set &DS;
where &ANALVAR > .Z;
byvar=&BYVAR;
keep &ANALVAR byvar;
proc sort; by byvar &ANALVAR;

data _NULL_; if 0 then set ci_ nobs=nobs;
if nobs ge &NAPPROX then
call symput('TITL2','Approximate Large Sample Method');
else call symput('TITL2','Exact Method');
stop;

proc univariate data=ci_ noprint pctldef=&DEF; by byvar;
var &ANALVAR;
output out=ci2_ n=n pctlpts=&P pctlpre=p;

data ci3_ ; set ci2_ ;
p = &P/100;
alpha=(100-&CL_PC)/100;
%IF &SIDE=T %THEN %DO;
alpha=alpha/2;
lowalpha = alpha;
upalpha = 1-lowalpha;
%END;
%ELSE %IF &SIDE=L %THEN %DO;
lowalpha = alpha;
upalpha = 0;
%END;
%ELSE %IF &SIDE=U %THEN %DO;
upalpha = 1-alpha;
lowalpha = 1;
%END;

if n ge &NAPPROX then do; ** Approx. large sample **;
term1 = .5 + n*p;
term2 = tinv(1-alpha,99999)*sqrt(n*p*(1-p));
%IF &SIDE^=U %THEN %DO;
yrank = int(term1 - term2); * Round down;
pforc1=probbnml(p,n,yrank-1);
output;
%END;
%IF &SIDE^=L %THEN %DO;
yrankr = (term1 + term2);
yrank = int(yrankr);
if yrankr > yrank then yrank + 1; * Round up;
pforc1=probbnml(p,n,yrank-1);
output;
%END;
end;

else do; ***** Exact method *****;
plast = probbnml(p,n,0);
if plast > lowalpha or probbnml(p,n,n-1) < upalpha then do;
error //
'** ERROR ** Sample size too small for given %tile/level'
' combination.' / '** ERROR ** You may try a one-sided'
' confidence interval.' //;
call symput('ERR','1');
end;
else do;
_x_=0; yrank=0;
do while (_x_ < n);
_x_ = _x_ + 1;
prob = probbnml(p,n,_x_);

```

```

%IF &SIDE^=U %THEN %DO;
  if yrank=0 and prob > lowalpha then do;
    yrank=_x_; pforcl=plast; output; end;
  %END;
%IF &SIDE=T %THEN %STR(else);
%IF &SIDE^=L %THEN %DO;
  if prob ge upalpha or _x_ = n-1 then do;
    yrank=_x_+1; pforcl=prob; output; _x_=n; end;
  %END;
  plast=prob;
end;
end;
retain plast;
keep byvar yrank pforcl;
run;
%IF &ERR NE 1 %THEN %DO;

```

```

  data ci_ ; set ci_ ; by byvar;
  if first.byvar then yrank = 1;
  else yrank+1;

```

```

  data ci3_ ; merge ci3_(in=in1) ci_ ; by byvar yrank;
  if in1;
  data ci3_ ; set ci3_(in=in1) ; by byvar;
  length upper lower 8.;
%IF &SIDE^=U %THEN %DO;
  if first.byvar then do;
    lower=&ANALVAR;
    lowp=pforcl;
  end;
%END;
if last.byvar then do;
  %IF &SIDE^=L %THEN %DO;
    upper=&ANALVAR;
  %END;
  %IF &SIDE=T %THEN
    %STR(truep=pforcl-lowp);
  %ELSE %IF &SIDE=L %THEN
    %STR(truep=1-pforcl);
  %ELSE %IF &SIDE=U %THEN
    %STR(truep=pforcl);
  output;
end;
retain lower lowp;
data ci3_ ; merge ci2_ ci3_ ;

```

```

title "&CL_PC.% &SIDETXT.Confidence Limits for"
      "&P.th Percentile of %UPCASE(&ANALVAR)";
title2 "&TITL2";
proc print noobs split='$';
  %IF &BYVAR NE . %THEN %DO;
    by byvar;
  %END;
  title3 " &TITL3 "; title4 " &TITL4 "; title5 " &TITL5 ";
  var n p&P lower upper truep;
  label n='N' p&P='Estimate' lower='Lower Limit'
        upper='Upper Limit'
        truep='Actual$Confidence$Level';
  format p&P lower upper &VALFMT. truep percent8.1;
run;
%END;
%MEND CCONTIN;
*****;

```