

Hodges-Lehmann confidence intervals for a shift in location using SAS® procedure SQL

Ann Olmsted, Dept. of Statistics, Texas A&M University,
College Station, Texas

Abstract

An inefficient but elegant way to compute Hodges-Lehmann nonparametric confidence intervals for a shift in location using SAS procedure SQL is described and illustrated. SAS procedure TIMEPLOT is used to display the results, and NPAR1WAY is used to obtain exact confidence levels. Similar code can be used to compute other nonparametric point and interval estimates, such as the corresponding Hodges-Lehmann CI for the median of a symmetric population.

Introduction

Let $X_1, \dots, X_m$ be a random sample from a continuous distribution F . Let $Y_1, \dots, Y_n$ be a random sample from a continuous distribution G . Assume $G(y) = F(y - \Delta)$; G is F shifted to the right by Δ units. The Hodges-Lehmann estimator of Δ associated with Wilcoxon's rank sum test is the median of the mn differences $Y_j - X_i$. If F has density f , the asymptotic efficiency of $\hat{\Delta}$ relative to $\bar{Y} - \bar{X}$ (ARE) is $12\sigma^2[\int f^2 dx]^2$, where $\sigma^2 = \text{Var}(F)$. The ARE is at least 0.864 and is $3/\pi$ or about 0.955 when F is normal. Let $Z_1, \dots, Z_n$ be a random sample from a continuous and symmetric distribution F with median θ . The Hodges-Lehmann estimator of θ associated with Wilcoxon's signed rank test is the median of the $\binom{n}{2} + n$ Walsh averages $(Z_i + Z_j)/2$, $j \leq i$. If F has density f , the AE of $\hat{\theta}$ relative to $\bar{Z}$ is identical to the value given above: $12\sigma^2[\int f^2 dx]^2$.

To illustrate, consider $\hat{\theta}$. Given the following 6 observations (Maritz 1981, Example 2.12):

3.38, 5.81, 4.46, 4.62, 4.15, 5.44

the 21 Walsh averages, in sorted order, are:

3.38, 3.765, 3.92, 4, 4.15, 4.305, 4.385, 4.41, 4.46, 4.54, 4.595, 4.62, 4.795, 4.95, 4.98, 5.03, 5.135, 5.215, 5.44, 5.625, 5.81

and their median, the estimate, is 4.595. Even for $n=6$, it was tedious to compute the estimate by hand.

Let SAS procedure SQL find the $\binom{n}{2} + n = n(n+1)/2$ pairwise averages and let SAS procedure UNIVARIATE find their median:

```
data WORK ;
  input x @@ ;
  id + 1 ;
cards ;
3.38 5.81 4.46 4.62 4.15 5.44
;
proc sql feedback ;
  create table WORK as
  select A.id as id1, B.id as id2, mean(A.x, B.x) as avg
  from WORK A, WORK B
  where A.id <= B.id
  order by avg, id1, id2 ;
proc univariate data=WORK ; var avg ;
+++++
Quantiles(Def=5)
50% Med      4.595
```

Shift CI example

Reduction in response to a painful stimulus after audio analgesia was measured by Petrie for 10 persons classified as augmenters (strong responders to stimuli) and 7 persons classified as reducers (weak responders):

```
CLASS      DIFF
2-augmenter 17.9 13.3 10.6 7.6 5.7 5.6 5.4 3.3 3.1 0.9
1-reducer   7.7 5.0 1.7 0.0 -3.0 -3.1 -10.5
```

Note: DIFF is the response without analgesia – the response with analgesia. Petrie's theory implied that differences should be larger for augmenters.

Although it may not be reasonable to assume that the augmenter and reducer distributions have the same shape, Lehmann (1975, Example 5, p. 92) used these data to illustrate shift CI calculations. The output of the shift CI macro Hodges2q is shown below. The first row shows the smallest and largest pairwise differences, the second row shows the second smallest and second largest, and so on.

```
data=/WORK/, class=/class/, id=/id/, out=/WORK/, var=/diff/
```

R	S	Group 2 diff -	Group 2 diff -	min	max
		group 1 diff	group 1 diff	-6.8	28.4
1	70	-6.80	28.40	L	H
2	69	-4.60	23.80	L	H
3	68	-4.40	21.10	L	H H
4	67	-4.10	21.00	L	H H
5	66	-2.30	20.90	L	H
6	65	-2.10	18.10	L	H H
7	64	-2.00	17.90	L	H
8	63	-1.90	16.40	L	H H
9	62	-1.70	16.30	L	H H
10	61	-0.80	16.20	L	H H
11	60	-0.10	16.20	L	H H
12	59	0.40	16.10	L	H H
13	58	0.60	15.90	L	H
14	57	0.70	13.80	L	H H
15	56	0.90	13.70	L	H H
16	55	1.40	13.60	L	H H
17	54	1.60	13.60	L	H H
18	53	2.60	13.30	L	H H
19	52	2.90	12.90	L	H
20	51	3.10	11.60	L	H
21	50	3.30	11.40	L	H H
22	49	3.70	10.70	L	H H
23	48	3.90	10.60	L	H H
24	47	3.90	10.60	L	H H
25	46	4.00	10.20	L	H
26	45	4.00	8.90	L	H
27	44	5.40	8.80	L H	
28	43	5.60	8.70	L H	
29	42	5.60	8.70	L H	
30	41	5.60	8.60	L H	
31	40	5.70	8.50	L H	
32	39	5.90	8.40	L H	
33	38	6.10	8.30	L H	
34	37	6.20	7.60	L H	
35	36	6.30	6.40	L H	

The shift CI macro (condensed version)

```
%macro Hodges2q(data=WORK, class=trt, id=eartag, out=WORK, var=x);
%let alpha = .10;
%global estimate;
title3 "data=/%data/, class=/%class/, id=/%id/, out=/%out/, var=/%var/" ;
(1)
proc sql feedback ;
  create table &out as
  select &id, &class, &var
  from &data
  where &class is not missing and .z < &var
  order by &class, &id ;
(2)
proc sql feedback ;
  create table &out as
  select
  A.&id as ida label="Group 1 &id",
  B.&id as idb label="Group 2 &id",
  B.&var - A.&var as diff label="Group 2 &var - group 1 &var"
  from &out A, &out B
  where A.&class < B.&class
  order by diff ;
(3)
* calc median diff & store in mac var ESTIMATE ;
proc univariate data=&out noprint ;
  var diff ;
  output out=Hodges2q median=estimate ;

proc sql feedback noprint ;
  select ESTIMATE into :ESTIMATE from Hodges2q ;
%let ESTIMATE = %cmpres(&ESTIMATE); %put ESTIMATE=/%ESTIMATE/;

  select count(*) as count into :count from &out ;
(4)
data &out ;
  set &out ;
  by diff ;
  retain iup 0 idown %eval(&count + 1) ;
  iup + 1 ;
  idown + (-1) ;
  if iup+idown ^= &count + 1 then do ;
    put iup= idown= ;
    abort return ;
  end ;
(5)
* fold the out file over on itself and proc timeplot the resulting intervals ;
proc sql feedback ;
  create table &out as
  select a.iup as r, b.iup as s, a.diff as lo, b.diff as hi
  from &out a, &out b
  where a.iup = b.idown and 2*a.iup <= &count
  order by r ;
(6)
proc timeplot data=&out ;
  id r s ;
  plot lo="L" hi="HI" / overlay ;
run ;
title3 ;
%mend Hodges2q;
```

How the shift CI macro works

Conceptually, if not literally, the FROM clause of the step 2 SELECT statement above combines every observation in the working file with every other observation and with itself, and the WHERE clause extracts the desired subset of all possible pairs of observations. The macro's 6 steps are annotated below:

- (1) Create a working file. The id variable should uniquely identify observations. In the one-sample example above, an id variable was essential to extract the desired subset of all possible pairs of observations. Here, it is not needed for this purpose and could be eliminated.
- (2) Combine every observation with every observation and extract the desired subset of all possible pairs, namely, all the pairs such that the first observation is one of the x's and the second is one of the y's. Calculate the y vs x differences and sort them from smallest to largest.
- (3) Use PROC UNIVARIATE to calculate the median y vs x difference, and use PROC SQL to store the value in macro variable ESTIMATE.
- (4) Number the y vs x differences, which have been sorted from smallest to largest, from 1 to mn (IUP) and from mn down to 1 (IDOWN). Since the next step folds the file over onto itself, matching the smallest difference to the largest, the second smallest to the second largest, and so on, it is convenient to have the observations numbered both ways.
- (5) Fold the file containing the mn y vs x differences onto itself, matching the smallest to the largest, the second smallest to the second largest, and so on. As above, this is done by using PROC SQL to combine every difference with every difference and extract the desired subset of the pairs using the index variables created in step 4.
- (6) Use PROC TIMEPLOT to display the order statistic CIs. Using this condensed version of the shift CI macro, it would be necessary to obtain the corresponding approximate or exact confidence levels from a table.

Obtaining approximate confidence levels

PROC SQL can be used as a calculator by creating a dummy file having 1 observation from which the calculated value is "selected". The dummy file must contain at least one variable, but none of the variables in the dummy file is actually used! The code below uses PROC SQL to obtain a pair of indices r, s , where $s = mn + 1 - r$, such that for large m and n the pair of order statistics $D_{(r)}, D_{(s)}$ has confidence level approximately equal to $1 - \alpha$ (Hollander and Wolfe 1973, p. 79):

```
%let alpha = .10;
data dummy ; retain x '1' ;
proc sql feedback ;
    select count(*) as n into :n1 - :n2
    from &out group by &class ;
%put n1 = /&n1/;
%put n2 = /&n2/;
select round(&n1*&n2/2 - probit(1 - &alpha/2) *
sqrt(&n1 * &n2 * (&n1 + &n2 + 1)/12)) as
approxr label="mn/2-probit(1-alpha/2)sqrt(mn(m+n+1)/12)"
into :approxr from dummy ;
%let approxr = %cmpres(&approxr);
title4 "Approx. r for >=1-&alpha conf lev CI for shift is &approxr"
```

Note: Lehmann (1975, p. 93) uses a continuity correction. To obtain Lehmann's version of the formula used above, replace $mn/2$ by $(mn + 1)/2$.

Obtaining exact confidence levels

In principle, it is simple to obtain exact confidence levels using PROC NPAR1WAY's "exact Wilcoxon;" statement. In practice, it's a bit tricky, because of numeric difficulties. For instance, using Lehmann's continuity-corrected version of the approximate procedure described above, $(D_{(19)}, D_{(52)})$ has confidence level $\approx 90\%$. The interval is (2.9, 12.9). Because of symmetry, the exact confidence level, $1 - .1088 = .891$, can be obtained by shifting the y's to the left by slightly less than 2.9 and running PROC NPAR1WAY to obtain the exact 2-sided Wilcoxon p-value:

```
Wilcoxon 2-Sample Test      S = 46.0000
Exact P-Values
(One-sided)  Prob <= S      = 0.0544
(Two-sided)  Prob >= |S - Mean| = 0.1088
```

The actual shift value used to obtain the output above was 2.899. (Note: Lehmann gives the confidence level as $1 - .108 = .892$). A program could be written to obtain the largest value of r such that $(D_{(r)}, D_{(mn+1-r)})$ has confidence level $\geq$ the desired value $1 - \alpha$:

1. Use the large-sample approximation formula to obtain a starting value of r .
2. If the exact confidence level is $\geq 1 - \alpha$, reduce r by 1 until the exact level is $< 1 - \alpha$, then increase r by 1. If the exact confidence level is $< 1 - \alpha$, increase r by 1 until the exact level is $\geq 1 - \alpha$.

Alternatives and recommendations

StatXact 3 (CYTEL Software Corporation 1996) computes exact Hodges-Lehmann CIs. Item 429 on the 1997 SAS Ballot (PART XVIII SAS/INSIGHT Software) was "provide Hodges-Lehmann estimate of location and confidence interval", so H-L estimates of and confidence intervals for the median of a symmetric distribution may soon be "built-in". A 1990 SUGI paper describes SAS macros MEDSYM (CIs for θ) and MEDIFF (CIs for Δ) (Wong and Lee). And there are published FORTRAN and C programs which calculate exact confidence levels. If you need a quick and simple method, adapting the code described in this paper or downloading Hodges1.sas (1-sample macro) or Hodges2.sas (2-sample macro) from <http://stat.tamu.edu/ftp/pub/> would be appropriate. If you need a full-featured pair of SAS macros, obtaining MEDSYM and MEDIFF would be worthwhile. If you are preparing a regulatory submission and have access to StatXact; use it; if efficiency is important, obtain one of the published FORTRAN or C programs.

Conclusions

By yoking SAS procedures SQL, UNIVARIATE, TIMEPLOT, and NPAR1WAY together, it is possible to do all of these things with less than a page of code:

- Calculate the 1-sample H-L estimate for the median of a symmetric distribution or the 2-sample estimate for a shift in location
- Calculate all possible symmetric-pair-of-order-statistics CIs
- Display the pairs graphically
- For large samples, find the pair with confidence level approximately equal to the desired value
- For small samples, find the narrowest interval with confidence level greater than or equal to the desired value

Hence, the Hodges-Lehmann estimates and CIs can be computed using SAS software almost as easily as the classical estimates and CIs. When should the H-L estimators be preferred to the corresponding classical estimators? Since this is a statistical computing paper, not a statistical paper, I'll leave the question open, except for one recommendation: if a thorough data-cleaning is not practical, play it safe

and use the H-L estimators.

References

- CYTEL Software Corporation (1996). *StatXact 3 for Windows User Manual*. Cambridge, MA: Author.
- Hodges, J.L., Jr. and Lehmann, E.L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics* 34, 598-611.
- [asymptotic results]
- Hollander, M. and Wolfe, D.A. (1973). *Nonparametric statistical methods*. New York: John Wiley & Sons.
- [formulas for large-sample approximations to the exact confidence levels]
- Lehmann, E.L. (1975). *Nonparametrics: statistical methods based on ranks*. San Francisco: Holden-Day.
- [formulas for large-sample approximations to the exact confidence levels; source of shift CI example]
- Maritz, J.S. (1981). *Distribution-free statistical methods*. London: Chapman and Hall.
- [source of one-sample example]
- Olmsted, Ann (1994). If I'm not a database user, what is the SQL procedure good for? *Proceedings of the Western Users of SAS Software Conference*, San Francisco, 1994.
- [SQL code for computing the Jonckheere statistic for testing for an upward or downward trend and Theil's nonparametric slope estimate]
- SAS Institute Inc. (1989). *SAS Guide to the SQL Procedure: Usage and Reference, Version 6, First Edition*. SAS Institute Inc., Cary, NC.
- SAS Institute Inc. (1990). *SAS Procedures Guide, Version 6, Third Edition*. SAS Institute Inc., Cary, NC.
- [proc timeplot]
- Wong, Rodney and Lee, Reiling (1990). Nonparametric estimates for the medians and difference in medians between two treatment groups. *Proceedings of the Fifteenth Annual SAS Users Group International Conference*. SAS Institute Inc., Cary, NC.
- [Several SAS macros are described but not listed. MEDSYM computes the Hodges-Lehmann estimate of the median of a symmetric distribution, confidence intervals for the median, and large-sample approximate confidence levels. MEDIFF computes the Hodges-Lehmann estimate of a shift in location, confidence intervals for the shift, and large-sample approximate confidence levels.]

Acknowledgements

I would like to thank Mike Abbott, for lending me his PROC SQL manual, and Andrew Cary, for showing by example that there is a way to do almost anything using SAS software.

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