

Having Confidence in Non-Parametric Data

John Salter, Oxford Pharmaceutical Sciences Ltd, Oxford, UK

ABSTRACT

Non-parametric models are such that the model structure is not specified a priori, but is instead determined from data. The term “non-parametric” is not meant to imply that such models completely lack parameters; rather, the number and nature of the parameters is flexible and not fixed in advance. Non-parametric models are therefore also called “distribution free”. Rather than quoting means and their confidence intervals, with non-parametric data, it may be considered more appropriate to present the median with confidence intervals.

This paper examines two techniques for calculating confidence intervals about the median – a simple (binomial) approximate method, and a more complex (signed rank) method, both using SAS®.

INTRODUCTION

Parametric inferential statistical methods are mathematical procedures for hypothesis testing which assume that the distributions of the variables being assessed belong to known parameterised families of probability distributions. While parametric techniques are robust – that is, they often retain considerable power to detect differences or similarities even when these assumptions are violated – some distributions violate the assumptions so markedly that a non-parametric alternative is more likely to detect a difference or similarity.

Non-parametric models differ from parametric models in that the model structure is not specified a priori, but is instead determined from data. The term “non-parametric” is not meant to imply that such models completely lack parameters; rather, the number and nature of the parameters is flexible and not fixed in advance. Non-parametric models are therefore also called “distribution free”.

Researchers sometimes quote means and their confidence intervals in situations where a median with confidence interval would be more appropriate (e.g. when outliers have a biasing effect on the mean but there is insufficient evidence to exclude them from the analysis, or – in this case – with non-parametric data).

This paper examines two techniques for calculating confidence intervals about the median – a simple (binomial) approximate method, and a more complex (signed rank) method, applying techniques derived from non-parametric methods of analysis, using SAS® to expedite the process.

METHODOLOGY

WHAT IS A CONFIDENCE INTERVAL?

A confidence interval (CI) is an interval between two numbers with an associated probability p which is generated from a random sample of an underlying population, such that if the sampling was repeated numerous times and the confidence interval recalculated from each sample according to the same method, a proportion p of the confidence intervals would contain the parameter in question. Confidence intervals are the most prevalent form of interval estimation.

When data follows a specific distribution (i.e. the data are parametric), then a confidence interval is usually presented about the mean of the data as follows:

$$\mu \pm Z_{1-\alpha/2} \sqrt{\frac{\sigma^2}{n}}$$

where μ is the population mean. This is used generally for large sample ($n > 30$) calculations; for smaller populations, the Z-statistic is replaced by the more robust t-statistic.

THE BINOMIAL APPROACH

This method is the simplest to perform and can provide a good approximation if the n is large enough. We can estimate confidence intervals for these using the Binomial distribution. The 95% confidence interval for the 50th quantile (i.e. the median) can be found by an application of the Binomial distribution (Conover 1980).

The number of observations less than a given quantile Q will be an observation from a Binomial distribution with parameters N and Q , with mean NQ and standard deviation $\sqrt{NQ(1-Q)}$. Taking the median as the 50th quantile, we can assume the “mean” in this case to be the 0.5^{Nth} observation and – applying the equation given above – the upper (UCL) and lower (LCL) confidence limits can be given as:

$$\begin{aligned} \text{LCL} &= NQ - Z_{1-\alpha} \sqrt{NQ(1-Q)} \\ &= 0.5N - Z_{1-\alpha} \sqrt{0.25N} \end{aligned}$$

$$\begin{aligned} \text{UCL} &= NQ + Z_{1-\alpha} \sqrt{NQ(1-Q)} \\ &= 0.5N + Z_{1-\alpha} \sqrt{0.25N} \end{aligned}$$

We round the values for the lower confidence limit (LCL) and upper confidence limit (UCL) up to the next integer. Then the 95% confidence interval is between these two “limits” in the ordered data.

EXAMPLE (1):

SUBJECT	1	2	3	4	5	6	7	8	9	10
PRE-DOSE	19	11	14	17	23	11	15	19	11	8
POST-DOSE	22	18	17	19	22	12	14	11	19	7
DIFF	-3	-7	-3	-2	+1	-1	+1	+8	-8	+1

Ranking the data differences gives us the following list:

-8, -7, -3, -3, -2, -1, +1, +1, +1, +8

As we have an even number of observations, our median is the midpoint of the two mid observations i.e. -1.5

Our confidence interval is:

$$0.5N \pm Z_{1-\alpha} \sqrt{0.25N} \approx 5 \pm 3.099$$

Therefore, we take the 1st (5-4) and 9th (5+4) observations as our confidence interval as (-8, 1).

The limitations of the Binomial method are that it is only really effective with large sample sizes – as with the example above, the small n shows a CI that encompasses all but one data point. Its strength lies in the fact that it can be applied to any quantile (or proportion) within a sample.

THE SIGNED RANK SUM APPROACH

This method is slightly more complex in that it involves further knowledge of statistical theory; the Wilcoxon Signed Rank Sum Test.

Here, observations are put in ascending order of magnitude (ignoring the sign) and given ranks of 1 to n' (zero values are ignored). Let T_+ be the sum of the ranks of the positive values and T_- be the sum of the ranks of the negative values. We then consider, as the null hypothesis, that T_+ and T_- will not differ greatly; the sum of T_+ and T_- is $\frac{1}{2}n'(n'+1)$ so an appropriate test would consist of evaluating the probability of a value of T_+ greater than or equal to the observed value. For large values of n' ($n' > 30$), the values of T_+ and T_- are approximately normally distributed as follows:

$$\approx N\left[0, \frac{n'(n'+1)(2n'+1)}{24}\right]$$

with a standardised normal deviate (with continuity correction) as follows:

$$\frac{\left| T_+ - \frac{1}{4} n'(n'+1)(2n'+1) \right| - \frac{1}{2}}{\sqrt{n'(n'+1)(2n'+1)/24}}$$

Supposing that the differences as detailed in Example (1) above are distributed symmetrically about the median; if subtracted from each observation, this would give the null expectation in the signed rank sum test. We can then take the aforementioned test statistic T_+ as the number of positive values amongst the “pair means” of the data (the mean of each pair of observations in the data).

Taking the total number of pair means in a population as $\frac{1}{2}n(n+1)$, we can take the median to be the midpoint of these observations, and applying the large sample approximation stated above, we have the following equation:

$$\frac{1}{4} [n(n+1)] - Z_{1-\alpha/2} \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

We therefore exclude the integer value of the above equation from either end of the ranked observations.

EXAMPLE (2):

Taking the same data as before:

SUBJECT	1	2	3	4	5	6	7	8	9	10
PRE-DOSE	19	11	14	17	23	11	15	19	11	8
POST-DOSE	22	18	17	19	22	12	14	11	19	7
DIFF	-3	-7	-3	-2	+1	-1	+1	+8	-8	+1

Ranking the data differences gives us the following list:

-8, -7, -3, -3, -2, -1, +1, +1, +1, +8

Taking the paired means as follows (55 observations = $\frac{1}{2} \times 10 \times 11$):

	-8	-7	-3	-3	-2	-1	+1	+1	+1	+8
-8	-8.0	-7.5	-5.5	-5.5	-5.0	-4.5	-3.5	-3.5	-3.5	0
-7		-7.0	-5.0	-5.0	-4.5	-4.0	-3.0	-3.0	-3.0	+0.5
-3			-3.0	-3.0	-2.5	-2.0	-1.0	-1.0	-1.0	+2.5
-3				-3.0	-2.5	-2.0	-1.0	-1.0	-1.0	+2.5
-2					-2.0	-1.5	-0.5	-0.5	-0.5	+3.0
-1						-1.0	0	0	0	+3.5
+1							+1.0	+1.0	+1.0	+4.5
+1								+1.0	+1.0	+4.5
+1									+1.0	+4.5
+8										+8.0

Applying the signed rank sum equation, we get the following confidence interval:

$$\frac{1}{4} [n(n+1)] - Z_{1-\alpha/2} \sqrt{\frac{n(n+1)(2n+1)}{24}} \approx 27.5 \pm 19.2$$

Therefore, we take the 9th observation and the 47th observation, which gives us a confidence interval of (-4.5, 1)

PROGRAMMING**BINOMIAL METHOD (BIN_CI.SAS)**

The following code can be used in SAS to derive the confidence intervals using the Binomial method:

```

/*****
** Macro name: BIN_CI.sas
**
** Variables : &DATASET - full dataset name (including libname)
**              &VAR      - variable of interest
**              &ALPHA    - confidence level (default=95)
**              &TRT      - treatment arm (numeric) [optional]
**
** Function   : Provides confidence limits about a median using the
**              Binomial approximation (Conover 1980)
**
** Outputs    : &LCLM     - lower confidence limit of median CI
**              &UCLM     - upper confidence limit of median CI
*****/

%macro bin_ci (dataset=,
               alpha=95,
               trt=,
               var=);

/*****
* Bring in non-missing data *
*****/
proc sort data=&DATASET out = setup;
  %if &TRT>0 %then %do;
    where &VAR ne . and trt=&TRT;
  %end;
  %else %do;
    where &VAR ne .;
  %end;
  by &VAR;
run;

data setup;
  set setup;
  rank=_n_;
run;

proc sql noprint;
  select count (&VAR) into: nn
  from setup;
quit;

/*****
* Create confidence interval based on Binomial approximation *
*****/
data bin_out;
  set setup;
  alpha=1-((100-&alpha)/200);
  zstat=probit(alpha);
  val = ceil(zstat*sqrt(0.25*&nn));
  call symput('low',(0.5*&nn - val));
  call symput('upp',(0.5*&nn + val));
run;

data final;
  set bin_out;
  if rank=&low then call symput("lclm",value);
  if rank=&upp then call symput("uclm",value);
run;

%mend;

```

SIGNED RANK SUM METHOD (WSRS_CI.SAS)

The following code can be used in SAS to derive the confidence intervals using the Signed Rank Sum method:

```

/*****
** Macro name: WSRS_CI.sas
**
** Variables : &DATASET - full dataset name (including libname)
**             &VAR      - variable of interest
**             &ALPHA    - confidence level (default=95)
**             &TRT      - treatment arm (numeric) [optional]
**
** Function   : Provides confidence limits about a median as described
**               in Armitage & Berry (4th Edition)
**
** Outputs    : &LCLM    - lower confidence limit of median CI
**               &UCLM    - upper confidence limit of median CI
*****/

%macro wsrs_ci(dataset=,
               var=,
               alpha=95,
               trt=);

/*****
* Bring in non-missing data *
*****/
data setup;
  set &DATASET;
  %if &TRT>0 %then %do;
    where &VAR ne . and trt=&TRT;
  %end;
  %else %do;
    where &VAR ne .;
  %end;
run;

data setup;
  set setup;
  nvar=_n_; * Mapping variable for merge later on;
run;

proc sql noprint;
  select count (&VAR) into: nn
  from setup;
quit;

/*****
* Output all variable data by subject *
* to create subject matrix
*****/
data matrix (keep=trt sid &VAR nvar);
  set setup (drop=nvar);
  do i=1 to &nn;
    nvar=i;
    output;
  end;
run;

proc sort data=matrix;
  by nvar;
run;

proc transpose data=matrix out=matrix2;
  by nvar;
  var &VAR;
run;

proc sort data=setup;
  by nvar;
run;

```

```

/*****
* Keep relevant information from merge *
*****/
data allin (drop=nvar _name_);
    merge setup (keep=trt sid &VAR nvar)
        matrix2;
    by nvar;
run;

%macro convert;
/*****
* Create paired means and remove duplicates *
*****/
data allin2 (where=(nmissx ne .));
    set allin;
    %let m=1;
    %do %while(&m le &nn);
        col&m=(&var+col&m)/2;
        %let m=%eval(&m+1);
    %end;
    %let q=1;
    %do %while (&q le &nn);
        nmissx=col&q;
        if _n_ gt &q then nmissx=.;
        output;
        %let q=%eval(&q+1);
    %end;
run;

%mend convert;

%convert;

title "Transposed output for Median CIs";
title2 "Armitage & Berry";
proc print data=allin2 ;
run;

/*****
* Calculate CIs *
*****/
data final;
    set allin2;
    alpha=1-((100-&alpha)/200);
    zstat=probit(alpha);
    wplow=1 + (int((&nn*(&nn+1)/4) - zstat*sqrt((&nn*(&nn+1)*(2*&nn+1))/24)));
    if wplow le 0 then wplow=1;
    wpupp=(&nn*(&nn+1)/2) - wplow;
    keep leg_sort treattxt nmissx wplow wpupp;
run;

proc sort data=final;
    by nmissx;
run;

title "Final output for median CIs";
proc print data=final;
run;

data conlim (keep=trt lclm uclm);
    set final end=eof;
    retain lclm uclm;
    if _n_=wplow then call symput("lclm",nmissx);
    if _n_=wpupp then call symput("uclm",nmissx);
    if eof;
run;

%mend;

```

CONCLUSION

We have examined two techniques for calculating confidence intervals about a median - the large sample Binomial method (that can be extended to cover any quantile in a specific dataset); and the more robust Signed Rank method that is structured towards the confidence interval about a median, but with greater precision and can be utilized by large and small samples alike.

As with all statistical techniques, we need to examine the data being analysed before coming to a decision about the methodological approach - the Binomial method is less user-intensive, but needs to be carefully considered with smaller studies.

REFERENCES

1. P. Armitage, G. Berry, J.N.S Matthews (2002); Statistical Methods in Clinical Research (Fourth Edition)
2. W.J. Conover (1980); Practical Non-Parametric Statistics

ACKNOWLEDGMENTS

Thanks to Katherine Hutchinson of Oxford Pharmaceutical Sciences Ltd. for her guidance and patience.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

John Salter
Oxford Pharmaceutical Sciences Ltd.
Lancaster House
Kingston Business Park
Kingston Bagpuize
OXFORDSHIRE
OX13 5FE
Work Phone: 01865 823823
Email: john.salter@ops-web.com
Web: www.ops-web.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.