Inference Procedures for One Sample and Paired-Data Location Problems

Robert J. Serfling – October, 2006

In §§1–4 we examine and compare three types of procedures for inference about *location* parameters, in the *one-sample* setting of an i.i.d. sample $X_1, \ldots, X_n$ from some cdf F, and with emphasis on the *nonparametric* approach. This will draw together facts and insights we have encountered previously and fill certain gaps as well.

In §5 we consider the closely related problem of inference about a *location shift* parameter associated with a set of "matched (paired) observations" (X_i, Y_i), not necessarily identically distributed, taken on a group of n subjects. This problem is treated by reduction to an application of the methods of §§1–3.

In §6 we illustrate with a Minitab session using a data set.

Lastly, a table of null hypothesis upper tail critical values for the Wilcoxon signed rank test is provided for selected sample sizes.

1. Procedures Based on the Sample Mean

- *assumptions*: a distribution F having mean μ and variance σ^2 (both unknown), and a random sample of independent observations $X_1, \ldots, X_n$ all having the distribution F.

- *parameter of interest*: μ

- *point estimator*: $\overline{X}$

- *confidence interval*: $\overline{X} \pm t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}}$ *or* $\overline{X} \pm z_{1-\alpha/2} \frac{s}{\sqrt{n}}$

  Here s is the usual sample standard deviation and for desired confidence $100(1-\alpha)\%$ the value $t_{n-1,1-\alpha/2}$ is selected from the appropriate t-distribution if F is normal, or otherwise in the large-sample case $z_{1-\alpha/2}$ is selected from the standard normal distribution.

- *hypothesis test for* $H_0 : \mu = \mu_0$

  Reject H_0 if

  $$\frac{\overline{X} - \mu_0}{s/\sqrt{n}} \text{ too large (1-sided test against } \mu > \mu_0)$$

  or if

  $$\left|\frac{\overline{X} - \mu_0}{s/\sqrt{n}}\right| \text{ too large (2-sided test against } \mu \neq \mu_0)$$

- *efficiency*: High if F is *truly normal* in shape. Very poor, however, even disastrous in some cases, if F is heavy-tailed (relative to normal).

- *robustness*: Poor, since extreme observations can pull $\overline{X}$ far away from μ.

- *nonparametric?*: No and Yes.
  The t-distribution procedure is a *parametric* approach, based on the normal model. However, the standard normal approximation, based on the CLT without knowledge of F other than an assumption of finite variance σ^2, is a *nonparametric* approach.

2. Procedures Based on the Sample Median, Order Statistics, Signs

- *assumptions*: a collection of independent observations $X_1, \ldots, X_n$, each having a *continuous* distribution (or at least a distribution attaching 0 probability to its median), with different distributions allowed for different observations, but all having the same value for their medians.

- *parameter of interest*: the value of the common median, ν

- *point estimator*: the sample median, $\widehat{\nu}$

- *confidence interval*: $(X_{(K)}, X_{(L)})$,
  where
  $$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$
  denote the *ordered* sample values and $L = n+1-K$, with K chosen to achieve desired confidence $100(1-\alpha)\%$, or confidence at least this value if not exactly, as follows. Note that
  $$\begin{aligned} P(X_{(K)} > \nu) &= 1 - P(\text{at least } K \text{ of the } n\ X_i\text{'s are} < \nu) \\ &= 1 - P(\text{bin}(n, 0.5) \geq K), \end{aligned}$$
  and thus it suffices to choose K small enough to make this (error) probability $\leq \alpha/2$. For small samples, use the table of the binomial$(n, 0.5)$ distribution. For large samples, use the corresponding normal approximation.

- *hypothesis test for* $H_0 : \nu = \nu_0$
  Reject H_0 if
  $$S = \sum_{i=1}^{n} Y_i \ \text{ is too far away from } \frac{n}{2},$$
  where $Y_i = 1$ or 0 according to whether $X_i > \nu_0$ or $X_i \leq \nu_0$. This is called the *sign test*, because the test statistic S is simply the number of deviations $X_i - \nu_0$ which have *positive sign*. The actual values of these deviations are ignored. Under H_0, the test statistic S is the number of "successes" in n independent trials with success probability 0.5 and hence has the bin$(n, 0.5)$ distribution with mean $\frac{n}{2}$.

- *efficiency relative to* $\overline{X}$: very high if F is *Laplace* (double exponential); rather low (64% for large samples) if F is actually *normal*; overall, not too bad.

- *robustness properties*: very favorable with respect to influence by *extreme* data values, but more sensitive than $\overline{X}$ to *small perturbations* ("wiggling") of the middle data values.

- *nonparametric?*: Yes. The confidence coefficients computed by the bin$(n, 0.5)$ distribution without knowledge of F, and the cutoff points for the sign test computed from this same distribution, are *exact* rather than approximate. For large samples the normal approximation to the bin$(n, 0.5)$ distribution may be employed.

3. Procedures Based on the Walsh Averages, Signed Ranks

- *assumptions*: a collection of independent observations $X_1, \ldots, X_n$, each having a *continuous* distribution (or at least a distribution attaching 0 probability to its median), with different distributions allowed for different observations, but all having the same value for their medians, and all distributions *symmetric* about their medians.

- *parameter of interest*: the common value of the point of symmetry (median ν) (which also equals the mean μ if the mean is finite)

- *point estimator*: median$\left\{\frac{X_i+X_j}{2},\ 1 \leq i \leq j \leq n\right\}$
  The quantities
  $$\frac{X_i + X_j}{2},\ 1 \leq i \leq j \leq n$$
  are called *Walsh averages*, and this point estimator is called the *Hodges-Lehmann location estimator*.

- *confidence interval*: $(W_{(K)}, W_{(L)})$,
  where
  $$W_{(1)} \leq W_{(2)} \leq \cdots \leq W_{(\frac{n(n+1)}{2})}$$
  denote the *ordered* Walsh averages, amd $L = \frac{n(n+1)}{2} + 1 - K$, with K chosen to achieve desired confidence (conservatively), as follows. Note that
  $$\begin{aligned} P(W_{(L)} < \nu) &= P(\text{at least } L \text{ of the } n(n+1)/2 \text{ Walsh averages are } < \nu) \\ &= P(\text{the number of Walsh averages} > \nu \text{ is} \leq K-1) \\ &= 1 - P_0(W^+ \geq K), \end{aligned}$$
  where the statistic W^+ is defined and discussed below and P_0 denotes the null hypothesis distribution. Thus it suffices to choose K small enough to make this probability $\leq \alpha/2$. For small samples, a special table for the *exact null hypothesis* distribution of W^+ is used. (The distribution of W^+ is known, whether or not F is known.) For large samples, use the *normal approximation* to this exact distribution, based on the null hypothesis mean and variance of W^+ (details below).

- *hypothesis test for $H_0 : \nu = \nu_0$*
  Reject H_0 if
  $$W^+ \text{ is too far away from } \frac{n(n+1)}{4},$$

where

$$W^+ = \sum_{i=1}^{n} Y_i\, R_i,$$

with again $Y_i = 1$ or 0 according to whether $X_i > \nu_0$ or $X_i \leq \nu_0$, and with R_i the rank of the ith *absolute* deviation $|X_i - \nu_0|$ among all the absolute deviations. Unlike the sign test, this test takes into account the sizes of the deviations $X_i - \nu_0$. That is, W^+ is the sum of *ranks* for those deviations with *positive signs*. It is called the *Wilcoxon signed rank test statistic*.

There is a close connection with the Hodges-Lehmann estimator given above, provided that no two sample values are equal (which is true with probability 1 if the distributions are continuous). Namely, we can rewrite W^+ as the number of Walsh averages greater than ν_0, i.e.,

$$W^+ = \sum_{i=1}^{n} \sum_{j=i}^{n} Z_{ij},$$

with $Z_{ij} = 1$ or 0 according to whether $\frac{X_i + X_j}{2} > \nu_0$ or $\frac{X_i + X_j}{2} \leq \nu_0$. When the null hypothesis H_0 is true, the statistic W^+ has mean

$$E(W^+) = \frac{n(n+1)}{4}$$

and variance

$$\mathrm{Var}(W^+) = \frac{n(n+1)(2n+1)}{24}\ .$$

It also is approximately normal with these mean and variance parameters, as the sample size n increases.

- *efficiency relative to* $\overline{X}$: Very high overall. Even 95.5% when the underlying distributions are all *normal*. Never below 86.4% for *any* continuous symmetric distribution F.

- *robustness properties*: Very favorable, due to limited influence by outliers and to some smoothing out of small perturbations of the data.

- *nonparametric?*: Yes. Can use the exact H_0-distribution (not depending on F), or the normal approximation in the large sample case.

4. Relative Efficiencies

The following table gives for several choices of distribution F the *asymptotic relative efficiency* (ARE) of the median estimator (sign test S) and Hodges-Lehmann estimator (Wilcoxon signed rank test W^+) relative to $\overline{X}$ (t-test). These ARE's represent the limiting ratio of sample sizes at which the two procedures being compared perform equivalently, in terms of the variance parameters of the associated approximating normal distributions. That is, if procedure A has variance parameter σ_A^2/n and procedure B has variance parameter σ_B^2/n, then the ratio of sample sizes at which these are equal is $\frac{\sigma_B^2}{\sigma_A^2}$, interpreted as the large-sample ARE of procedure A relative to procedure B.

cdf	S	W
Normal	$2/\pi = .637$	$3/\pi = .955$
Uniform	.333	1.000
Laplace	2.000	1.500
any cts sym F		$\geq .864$

It should be noted that ARE's do not necessarily give accurate indications for small samples. For example, the *exact* relative efficiency of the median estimator relative to $\overline{X}$ is actually .95 for $n = 5$, .80 for $n = 10$, and .70 for $n = 20$, decreasing to $2/\pi$ $= .637$ as $n \to \infty$.

5. Experiments Involving Matched (Paired) Data

Many experiments involve bivariate observations (X_i, Y_i), $1 \leq i \leq n$, taken for each of a group of n subjects. We might think of X_i as a "*before treatment*" observation and Y_i as an "*after treatment*" observation, for the ith subject. The target parameter is the "*treatment effect*", and the null hypothesis is "H_0: *no treatment effect.*"

A natural model may be formulated in terms of the differences $Z_i = Y_i - X_i$, by assuming that

$$Z_i = \theta + e_i \ , \quad 1 \leq i \leq n,$$

with $e_1, \ldots, e_n$ *independent* random variables (not necessarily identically distributed), and with θ representing an unknown additive "treatment effect". With appropriate additional assumptions, we may analyze this problem using methods similar to those of §§1-3 above.

(i) *Approach based on the sample mean.* Assume that the e_i's are *identically distributed* with mean 0 and use the approach of §1 with μ replaced by θ and the X_i's by the Z_i's.

(ii) *Approach based on the order statistics.* Here we need not assume identical distributions for the e_i's. However, we assume that the distributions of the e_i's are *continuous* and have *common median* 0. In this case, the approach and procedures of §2 remain valid, with the X_i's replaced by the Z_i's and ν by θ. Avoiding the identical distributions assumption allows for the possibility of subject-to-subject variation from factors other than the treatment effect, without these factors affecting value of the "treatment effect" θ.

(iii) *Approach based on the Walsh averages.* Here we adopt the same assumptions as in (ii), plus symmetry. (Note that the property of symmetry for the distribution of Z_i is satisfied, for example, if X_i and Y_i are independent and have a common distribution – which can be a different one for different i.) In this case the procedures of §3 remain in force.

6. Minitab Commands: Examples

Suppose that the data has been entered in column c1 of the Minitab Worksheet. The data values are

```
MTB >  print c1

Data Display

c1
     86      90      93      93      59       *      92      67      96      76
     92      92      95     102      87      73     104      80      90      92
      *       *      80      69      96      74      56      86      92      66
     65      77      84      89      90      78      77      98      98      81
    105      80     105      87      92      87      99      96      83      81
     68     101      92      80      77       *     108      92      87     102
     83      96      78     105       *      99      88      90      93      93
     99      80      87      84      84     105
```

Since there are 71 items in this data set, but 5 values are missing, the effective size of this data set is $n = 66$. (These data are actual test scores in the possible range of 0 to 110.)

6.1 t-test and related CI

A Minitab session for a two-sided *t test* of the hypothesis *mean = 80* for the parent population of the data in c1 has the following commands and output:

```
MTB > OneT c1;
SUBC>   Test 80.

One-Sample T: c1

Test of mu = 80 vs mu not = 80

Variable          N      Mean     StDev   SE Mean
c1               71     87.34     11.46      1.36

Variable             95.0% CI            T       P
c1             (   84.63,   90.05)     5.40  0.000
```

Note that the output includes the value of the test statistic, the p-value, and a 95% C.I. (by default).

If desired, one can choose a one-sided version and a different confidence level, for example:

```
MTB > OneT c1;
SUBC>   Test 80;
SUBC>   Confidence 99.0;
SUBC>   Alternative 1.

One-Sample T: c1

Test of mu = 80 vs mu > 80

Variable          N      Mean     StDev   SE Mean
c1               71     87.34     11.46      1.36

Variable      99.0% Lower Bound        T      P
c1                        84.10     5.40  0.000
```

(For a one-sided version in the other direction, replace 1 by -1 in the subcommand.)

Noe let's carry this out with the null hypothesis value of 80 replaced by 85, again for both two-sided and one-sided versions, but this time with 95% confidence in each version.

```
MTB > OneT c1;
SUBC>   Test 85.

One-Sample T: c1

Test of mu = 85 vs mu not = 85

Variable          N      Mean     StDev   SE Mean
c1               71     87.34     11.46      1.36

Variable             95.0% CI            T      P
c1            (   84.63,   90.05)     1.72  0.090

MTB > OneT c1;
SUBC>   Test 85;
SUBC>   Alternative 1.

One-Sample T: c1

Test of mu = 85 vs mu > 85

Variable          N      Mean     StDev   SE Mean
c1               71     87.34     11.46      1.36

Variable      95.0% Lower Bound        T      P
c1                        85.07     1.72  0.045
```

Note that the one-sided p-value is exactly one-half the two-sided p-value. Also, the one-sided p-value of 0.045 is just below the significance level 0.05 associated with the 95% choice of confidence level. This is consistent with

1) rejection of $H_0 : \mu = 85$ at level $\alpha = 0.05$, but acceptance at $\alpha = 0.01$
2) the H_0-value of 85 being below the 95% lower bound for μ.

6.2 Sign test and related CI

Now let's see what Minitab gives for a *sign test* of the hypothesis *median = 80* for the data in c1:

```
MTB > STest 80 c1;
SUBC>   Alternative 0.

Sign Test for Median: c1

Sign test of median = 80.00 versus  not =  80.00

                N    N*  Below  Equal  Above         P     Median
c1             71     5     15      5     51    0.0000      89.00

MTB > SInterval 95.0 c1.

Sign CI: c1

Sign confidence interval for median

                                           Achieved
     N    N*   Median   Confidence     Confidence interval   Position
c1  71     5    89.00      0.9424       (    86.00,   92.00)         28
                           0.9500       (    85.58,   92.00)        NLI
                           0.9673       (    84.00,   92.00)         27

MTB > SInterval 99.0 c1.

Sign CI: c1

Sign confidence interval for median

                                           Achieved
     N    N*   Median   Confidence     Confidence interval   Position
c1  71     5    89.00      0.9824       (    84.00,   92.00)         26
                           0.9900       (    84.00,   92.00)        NLI
                           0.9910       (    84.00,   92.00)         25
```

Note that for each level of confidence, *three intervals* are provided. For the 95% case, for example, the *first* interval gives the highest achievable confidence level (0.9424 for our sample size of 66) that is just below the requested confidence level (0.95). The *third* interval gives the lowest achievable confidence level (0.9673 for our sample size of 66) that is just above the requested level (0.95). These are calculated as described in §2 above. Also, the *ranks* of the left and right endpoints among the ordered sample values are provided.

Only rarely, however, can one achieve the requested confidence exactly with these intervals. Therefore, Minitab offers the middle confidence interval, which slightly adjusts the endpoints as needed to get the actual confidence level closer to the requested one. (This uses a nonlinear interpolation procedure of T.P. Hettmansperger and S.J. Sheather (1986), "Confidence Intervals Based on Interpolated Order Statistics", *Statistics and Probability Letters* **4** pp. 75-79.)

Now, for comparison, let's look at the p-values for testing a null hypothesis value of 85 instead of 80.

```
MTB > STest 85 c1;
SUBC>   Alternative 0.

Sign Test for Median: c1

Sign test of median = 85.00 versus  not =  85.00

               N    N*  Below  Equal  Above         P     Median
c1            71     5     27      0     44    0.0576      89.00

MTB > STest 85 c1;
SUBC>   Alternative 1.

Sign Test for Median: c1

Sign test of median = 85.00 versus  >  85.00

               N    N*  Below  Equal  Above         P     Median
c1            71     5     27      0     44    0.0288      89.00
```

6.3 Signed rank test and related CI

Since this test assumes *symmetry* of the parent population, let's first look at a histogram of the data.

```
MTB > Histogram c1.

Histogram of c1   N = 71   N* = 5

Midpoint        Count
      55            1  *
      60            1  *
      65            3  ***
      70            2  **
      75            6  ******
      80            9  *********
      85           12  ************
      90           14  **************
      95            9  *********
     100            8  ********
     105            5  *****
     110            1  *
```

Well, that's not too inconsistent with a symmetry assumption, so let's go ahead with *signed rank* tests for each of two null hypothesis values, 80 and 85, for the median, as well as 95% and 99% CI's. For testing 80, we obtain

```
MTB > WTest 80 c1;
SUBC>   Alternative 0.

Wilcoxon Signed Rank Test: c1

Test of median = 80.00 versus median not = 80.00

                  N     N for    Wilcoxon             Estimated
             N Missing   Test    Statistic        P      Median
c1          71        5     66      1826.5    0.000       88.00
```

The Wilcoxon test statistic of 1826.5 is the number of Walsh averages exceeding the H_0 value of 80. For the sample of 66 (non-missing) values in this data set, there are

$$\frac{66 \times 67}{2} = 2211$$

Walsh averages in all.

Since the p-value is 0.000, the null hypothesis can be rejected at all reasonable levels of significance.

The estimated median here, 88.0, is the median of the Walsh averages. This differs (as expected) from the median of the data values, which is 89 as seen above.

Now let's get 95% and 99% CI's.

```
MTB > WInterval 95.0 c1.

Wilcoxon Signed Rank CI: c1

                 Number  Estimated    Achieved
             N  Missing     Median  Confidence  Confidence Interval
c1          71        5      88.00        95.0  (   85.00,   91.00)

MTB > WInterval 99.0 c1.

Wilcoxon Signed Rank CI: c1

                 Number  Estimated    Achieved
             N  Missing     Median  Confidence  Confidence Interval
c1          71        5      88.00        99.0  (   84.00,   91.50)

MTB > WTest 85 c1;
SUBC>   Alternative 0.

Wilcoxon Signed Rank Test: c1

Test of median = 85.00 versus median not = 85.00

                  N     N for    Wilcoxon             Estimated
             N Missing   Test    Statistic        P      Median
c1          71       5     71      1630.0    0.044      88.00

MTB > WTest 85 c1;
SUBC>   Alternative 1.

Wilcoxon Signed Rank Test: c1

Test of median = 85.00 versus median  >  85.00

                  N     N for    Wilcoxon             Estimated
             N Missing   Test    Statistic        P      Median
c1          71       5     71      1630.0    0.022      88.00
```

7. Null Hypothesis Distribution of Wilcoxon Signed Rank Test Statistic

Table. Upper Tail Critical Values K and Probabilities P for the Null Hypothesis Distribution of the Wilcoxon Signed Rank Statistic W^+

(P denotes $P_0(W^+ \geq K)$)

n	K	P	n	K	P	n	K	P
3	6	.125	12	56	.102	16	93	.106
4	9	.125		60	.055		94	.096
	10	.062		61	.046		100	.052
5	13	.094		64	.026		106	.025
	14	.062		68	.010		112	.011
	15	.031		71	.005		113	.009
6	17	.109	13	64	.108		116	.005
	19	.047		65	.095	17	104	.103
	20	.031		69	.055		105	.095
	21	.016		70	.047		112	.049
7	22	.109		74	.024		118	.025
	24	.055		78	.011		125	.010
	26	.023		79	.009		129	.005
	28	.008		81	.005	18	116	.098
8	28	.098	14	73	.108		124	.049
	30	.055		74	.097		131	.024
	32	.027		79	.052		138	.010
	34	.012		84	.025		143	.005
	35	.008		89	.010	19	128	.098
	36	.004		92	.005		136	.052
9	34	.102	15	83	.104		137	.048
	37	.049		84	.094		144	.025
	39	.027		89	.053		152	.010
	42	.010		90	.047		157	.005
	44	.004		95	.024	20	140	.101
10	41	.097		100	.011		150	.049
	44	.053		101	.009		158	.024
	47	.024		104	.005		167	.010
	50	.010					172	.005
	52	.005						