Smoothing of sign test and approximation of its p-value

Mengxin LU

Kyushu Univ., Grad. Sch. Math.

Yoshihiko MAESONO

Kyushu Univ., Fac. Math.

1. Introduction

$X_1, X_2, \cdots, X_n$: i.i.d. $F(x-\theta)$

F is symmetric distr., i.e. $F(-x) = 1 - F(x)$ $(f(-x) = f(x))$

One sample testing problem:

Null-hypothesis: $H_0 : \theta = 0$ vs. Alternative: $H_1 : \theta > 0$

[Test statistics]

t-test

$$T = \frac{\sqrt{n}(\overline{X})}{\sqrt{V}}, \quad \overline{X} = \sum_{i=1}^{n} X_i, \quad V = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Sign test

$$S = \sum_{i=1}^{n} \psi(X_i), \quad \psi(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Mann-Whitney test (Wilcoxon's signed rank test)

$$W = \sum_{0 \leq i \leq j \leq n} \psi(X_i + X_j)$$

Under null hypothesis H_0, distributions of $S,\ W$ do not depend on F (distribution-free)

For observed values $s,\ w$ of $S,\ W$, we evaluate significance probabilities

$$P_0(S \geq s), \quad P_0(W \geq w)$$

If p-value is small enough, we reject H_0

For large n it is quite difficult to obtain exact distribution

Normal approximation

$$\frac{S - E_0(S)}{\sqrt{Var_0(S)}} \xrightarrow{L} N(0,1),$$

$$\frac{W - E_0(W)}{\sqrt{Var_0(W)}} \xrightarrow{L} N(0,1)$$

Using (continuity correction) p-vales are given by

$$\begin{aligned} P_0(S \geq s) &= P_0\left(\frac{2(S - \frac{n}{2} - 0.5)}{\sqrt{n}} \geq \frac{2(s - \frac{n}{2} - 0.5)}{\sqrt{n}}\right) \\ &\approx 1 - \Phi\left(\frac{2(s - \frac{n}{2} - 0.5)}{\sqrt{n}}\right), \end{aligned}$$

$$\begin{aligned} & P_0(W \geq w) \\ = \; & P_0\left(\frac{\sqrt{24}(W - \frac{n(n+1)}{4} - 0.5)}{\sqrt{n(n+1)(2n+1)}} \geq \frac{\sqrt{24}(w - \frac{n(n+1)}{4} - 0.5)}{\sqrt{n(n+1)(2n+1)}}\right) \\ \approx \; & 1 - \Phi\left(\frac{\sqrt{24}(w - \frac{n(n+1)}{4} - 0.5)}{\sqrt{n(n+1)(2n+1)}}\right) \end{aligned}$$

2. Smoothing sign based on kernel estimation

Purpose of this talk : constructing smoothing sign test

Brown, Hall, Young (2001, B.K.) : Smoothed sign test

Median $\tilde{\theta}$

$$\tilde{\theta} = argmin \sum_{i=1}^{n} |x_i - \theta|$$

Smoothed median:

$$\hat{\theta} = argmin \sum_{i<j} \{(x_i - \theta)^2 + (x_j - \theta)^2\}^{1/2}$$

Smoothed sign test : based on the smoothed median estimator

Proposed test has different Pitman A.R.E.

The proposed test statistic is not distribution-free

The asymptotic distribution depend on the underlying distribution

We have to make estimators

They also obtained an Edgeworth expansion,

but it contains unknown parameters

Let us consider empirical distribution

$$F_n(x_0) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x_0)$$

If $F(x)$ is symmetric around 0, $F(0) = \frac{1}{2}$

If alternative H_1 is true, $F_n(0)$ is an unbiased estimator of $F(-\theta)$

Under H_1, $F_n(0)$ takes smaller value than $\frac{1}{2}$, stochastically

Under H_0, $nF_n(0)$ is a binomial $B(n, \frac{1}{2})$

Normal approximation is valid, but the Edgeworth expansion is invalid

[Test statistic based on kernel estimator]

Let us assume a kernel function $k(u)$ is 2nd order

$$\int_{-\infty}^{\infty} k(u)du = 1, \quad \int_{-\infty}^{\infty} uk(u)du = 0$$

Let us define

$$W(t) = \int_{-\infty}^{t} k(u)du$$

Kernel estmator of $F(x_0)$ is given by

$$\tilde{F}_n(x_0) = \frac{1}{n}\sum_{i=1}^{n} W(\frac{x_0 - X_i}{h_n})$$

where h_n is bandwidth $h_n \to 0,\ nh_n \to \infty\ (n \to \infty)$

For testing H_0, we can use

$$\tilde{S} = \tilde{F}_n(0)$$

This test is regarded as smoothed test of S $(F_n(0))$

For observed value $\tilde{s}$, significance probability is

$$P_0(\tilde{S} \leq \tilde{s})$$

$\tilde{S}$ is not distribution-free.

In the sequel, we use the bandwidth $h_n = n^{-\frac{1}{4}}$, or $h_n = n^{-\frac{1}{3}}$

$\tilde{S}$ is a sum of *i.i.d.* random variables. If

$$0 < \lim_{n\to\infty} Var[W(\frac{-X_i}{h_n})] < \infty,$$

$\tilde{S}$ is asymptotically normal

$$\frac{\tilde{S} - E(\tilde{S})}{\sqrt{Var(\tilde{S})}} \xrightarrow{L} N(0,1)$$

Its asymptotic variance is given by

$$\lim_{n\to\infty} nVar_\theta(\tilde{S}) = F(-\theta)(1-F(-\theta)) = (1-F(\theta))F(\theta)$$

Further we have

$$\mu_{\tilde{S}_n}(\theta) = E_\theta(\tilde{S}) = F(-\theta) + O(h_n),$$

$$nVar_0(\tilde{S}) = F(0)(1-F(0)) + O(h_n) = \frac{1}{4} + O(h_n),$$

$$\lim_{n\to\infty} \frac{\mu'_{\tilde{S}_n}(0)}{\sqrt{nVar_0(\tilde{S}_n)}} = e(\tilde{S}) = -2f(0)$$

This efficacy coinsides to one of the sign test S.

Thus S and $\tilde{S}$ have same Pitman A.R.E.

3. Approximation of p-value based on Edgeworth

Edgeworth expansion for kernel type estimator $\tilde{F}_n(x_0)$ of $F(x_0)$

García-Soidán et al. (1997):

They showed the validity of the expansion, but not obtained an explicit form of the expansion

Huang and Maesono (2013)

Assuming $h_n = n^{-\frac{1}{4}}$ or $h_n = n^{-\frac{1}{3}}$, obtained the explicit form of the expansion

[When $h_n = n^{-\frac{1}{4}}$]

Under some regularity conditions

$$P\left(\frac{\tilde{F}_n(x_0) - F(x_0)}{\sqrt{n}\sqrt{Var(\frac{1}{n}W(\frac{x_0 - X_1}{h_n}))}} \leq y\right) = G(y) + o(n^{-\frac{1}{2}}) \tag{1}$$

where

$$G(y) = \Phi(y) + D_1 + n^{-\frac{1}{4}}D_2 + n^{-\frac{1}{2}}D_3. \tag{2}$$

$$\begin{aligned} D_1 = & -\phi(y)C_2 + \frac{1}{2}\phi'(y)C_2^2 - \frac{1}{6}\phi^{(2)}(y)C_2^3 + \frac{1}{24}\phi^{(3)}(y)C_2^4 \\ & -\frac{1}{120}\phi^{(4)}(y)C_2^5 + \frac{1}{720}\phi^{(5)}(y)C_2^6, \\ D_2 = & -\phi(y)C_2 + \phi'(y)C_2C_3 - \frac{1}{6}\phi^{(2)}(y)C_2^2C_3 + \frac{1}{24}\phi^{(3)}(y)C_2^3C_3, \end{aligned}$$

$$
\begin{aligned}
D_3 = & -\phi(y)C_4 + \frac{1}{2}\phi'(y)(C_3^2 + C_2C_4) - \frac{1}{6}B_{3,0}(\phi(y)(y^2-1) \\
& -C_2[\phi'(y)(y^2-1) + 2y\phi(y)] + C_2^2[\phi(y) + 2y\phi'(y) \\
& +\frac{1}{2}(y^2-1)\phi^{(2)}(y)] - C_2^3[\phi'(y) + y\phi^{(2)}(y) + \frac{1}{6}(y^2-1)\phi^{(3)}(y)])
\end{aligned}
$$

Further

$$
A_{i,j} = \int_{-\infty}^{\infty} W^i(u)k(u)u^j du,
$$

$$
B_{3,0} = \frac{1 - 2F(x_0)}{\sqrt{F(x_0)(1-F(x_0))}},
$$

$$B_{3,1} = \frac{3f(x_0)(A_{1,1} - A_{1,2})}{[F(x_0)(1 - F(x_0))]^{\frac{3}{2}}},$$

$$C_2 = \frac{f'(x_0)A_{0,2}}{\sqrt{F(x_0)(1 - F(x_0))}},$$

$$C_3 = -\frac{f''(x_0)A_{0,3}}{6\sqrt{F(x_0)(1 - F(x_0))}} + \frac{f(x_0)f'(x_0)A_{0,2}A_{1,1}}{2[F(x_0)(1 - F(x_0)]^{\frac{3}{2}}},$$

$$C_4 = \frac{f^{(3)}(x_0)A_{0,4}}{\sqrt{F(x_0)(1 - F(x_0))}} - \frac{f(x_0)f''(x_0)A_{0,3}A_{1,1}}{6[F(x_0)(1 - F(x_0))]^{\frac{3}{2}}}$$
$$- \frac{(f'(x_0))^2 A_{0,2}(A_{1,2} - F(x_0)A_{0,2})}{4[F(x_0)(1 - F(x_0))]^{\frac{3}{2}}} + \frac{f^2(x_0)f'(x_0)A_{0,2}A_{1,1}^2}{4[F(x_0)(1 - F(x_0)]^{\frac{5}{2}}}$$

When we use $\tilde{S}$, note that $x_0 = 0$

If $k(-u) = k(u)$ is symmetric, we have $W(u) = 1 - W(-u)$, and so

$$A_{1,2} = \int_{-\infty}^{\infty} W(u)k(u)u^2 du = \int_{-\infty}^{\infty} \{1 - W(-u)\}k(u)u^2 du$$

$$= \int_{-\infty}^{\infty} \{1 - W(t)\}k(-t)(-t)^2 dt \int_{-\infty}^{\infty} k(t)t^2 dt - \int_{-\infty}^{\infty} W(t)k(t)t^2 dt$$

$$= -\int_{-\infty}^{\infty} W(t)k(t)t^2 dt = -A_{1,2}$$

Thus $A_{1,2} = 0$. If the following conditions on the kernel

$$k(-u) = k(u), \quad \int_{-\infty}^{\infty} k(z)z^2 dz = \int_{-\infty}^{\infty} k(z)z^4 dz = 0, \tag{3}$$

$$\int_{-\infty}^{\infty} W(z)k(z)z dz = 0 \tag{4}$$

are satisfied, we have $C_2 = C_3 = C_4 = 0$. Thus we have the equation (1)

$$G(y) = \Phi(y) - n^{-\frac{1}{2}} \frac{B_{3,0}}{6} \phi(y)(y^2 - 1)$$

Further, since $F(0) = \frac{1}{2}$, we have $B_{3,0} = 0$

Also

$$E_0(\tilde{S}) = \frac{1}{2} + o(n^{-\frac{3}{2}}),$$
$$V_0(\tilde{S}) = \frac{1}{4n} + o(n^{-\frac{3}{2}})$$

Therefore the conditions (3) and (4) are satisfied, we have an improve-

ment of the normal approximation

$$P(2\sqrt{n}(\tilde{S} - \frac{1}{2}) \leq y) = \Phi(y) + o(n^{-\frac{1}{2}})$$

[Remark 1]

In addition to (3), (4), if the condition

$$A_{1,3} = \int_{-\infty}^{\infty} W(z)k(z)z^3 dz = 0$$

is satisfied, we have the Edgeworth expansion with remainder term $o(n^{-\frac{3}{4}})$, and the expansion does not include unknown parameters.

[**When** $h_n = n^{-\frac{1}{3}}$]

- If (3), (4) are satisfied, the residual term is $o(n^{-\frac{3}{4}})$
- If (3), (4) and $A_{1,3} = A_{2,2} = 0$ are satisfied, the residual term is $o(n^{-1})$. And the expansion does not have unknown parameters.

  The n^{-1} term is

$$\frac{1 - 3F(0) + 3F^3(0)}{F(0)(1 - F(0))}$$

[**When** $h_n = n^{-\delta}(\frac{1}{3} < \delta < \frac{1}{2})$ **のとき**]

- (3), (4) are satisfied, the residual term is $o(n^{-1})$, and the expansion does not have unknown parameters.

[Remark 2]

The kernel $k(u)$ which satisfies (3) and (4) is given by

$$k(u) = (a_0 + a_1|u| + a_2u^2 + a_3|u|^3)I(|u| \leq 1)$$

where

$$a_0 = \frac{45}{64}(1 + \sqrt{65}), \quad a_1 = \frac{9}{4}(5 - 3\sqrt{65}),$$
$$a_2 = \frac{105}{64}(-23 + 9\sqrt{65}), \quad a_3 = 9(3 - \sqrt{65})$$

It is possible to make a kernel which satisfies $A_{1,3} = A_{2,2} = 0$. The kernel is quite complicate.

[Sketch of the proofs]

Let us define

$$W_1 = W(\frac{x_0 - X_1}{h_n})$$

Evaluation of bias

Using Taylor expansion, we have

$$\begin{aligned} E(W_1) &= \int W(\frac{x_0 - y}{h_n}) f(y) dy \\ &= h_n \int W(z) f(x_0 - h_n Z) dz \\ &= h_n (-\frac{1}{h_n}) W(z) F(x_0 - h_n z)|_{-\infty}^{+\infty} \end{aligned}$$

$$
\begin{aligned}
& -h_n(-\frac{1}{h_n}) \int k(z)F(x_0 - h_n z)dz \\
= & \int k(z)F(x_0 - h_n z)dz \\
= & \int k(z)(F(x_0) - h_n z f(x_0) + \frac{1}{2}(h_n z)^2 f^{'}(x_0) + \cdots)dz \\
= & F(x_0) \int k(z)dz - f(x_0)h_n \int zk(z)dz \\
& + \frac{1}{2}f^{'}(x_0)h_n^2 \int z^2 k(z)dz + \cdots \\
= & F(x_0) + \frac{1}{2}h_n^2 f^{'}(x_0)A_{0,2} + \frac{1}{6}h_n^3 f^{''}(x_0)A_{0,3} \\
& + \frac{1}{24}h_n^4 f^{(3)}(x_0)A_{0,4} + O(h_n^5)
\end{aligned}
$$

If $k(u)$ is 4-th order kernel, bias is $O(h_n^4)$ and

$$A_{0,1} = A_{0,2} = A_{0,3} = 0$$

If $k(u)$ is 6-th order kernel, bias is $O(h_n^5)$ and

$$A_{0,1} = A_{0,2} = A_{0,3} = A_{0,4} = A_{0,5} = 0$$

If $k(u)$ is ℓ-th order kernel

$$\int_{-\infty}^{\infty} u^{m-1} k(u) du = 0 \ (2 \leq m \leq \ell - 1), \quad \int_{-\infty}^{\infty} u^{\ell} k(u) du \neq 0$$

Using Taylor expansion, we have

$$
\begin{aligned}
E(W_1^2) &= \int W^2(\frac{x_0 - y}{h_n}) f(y) dy \\
&= 2 \int W(z) k(z) F(x_0 - h_n z) dz \\
&= 2 \int W(z) k(z) (F(x_0) - h_n z f(x_0) + \frac{1}{2}(h_n z)^2 f^{'}(x_0) + \cdots) dz \\
&= 2(F(x_0) \int W(z) k(z) dz - \int W(z) k(z) h_n z f(x_0) dz \\
&\quad + \frac{1}{2} \int h_n^2 z^2 W(z) k(z) f^{'}(x_0) dz + \cdots) \\
&= F(x_0) - 2 h_n f(x_0) A_{1,1} + h_n^2 f^{'}(x_0) A_{1,2} \\
&\quad + \frac{1}{3} h_n^3 f''(x_0) A_{1,3} + O(h_n^5)
\end{aligned}
$$

Similarly

$$
\begin{aligned}
E(W_1^3) &= h_n \int W^3(z) f(x_0 - h_n z) dz \\
&= 3 \int W^2(z) k(z) F(x_0 - h_n z) dz \\
&= 3 \int W^2(z) k(z) (F(x_0) - h_n z f(x_0) + \frac{1}{2}(h_n z)^2 f^{'}(x_0) + \cdots) dz \\
&= 3F(x_0) \int W^2(z) k(z) dz - 3h_n f(x_0) A_{2,1} \\
&\quad + \frac{3}{2} h_n^2 f^{'}(x_0) A_{2,2} + O(h_n^3) \\
&= F(x_0) - 3h_n f(x_0) A_{2,1} + \frac{3}{2} h_n^2 f^{'}(x_0) A_{2,2} + O(h_n^3)
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
E(W_1^4) &= \int W^4(\frac{x_0 - y}{h_n}) f(y) dy \\
&= h_n \int W^4(z) f(x_0 - h_n z) dz \\
&= 4 \int W^3(z) k(z) F(x_0 - h_n z) dz \\
&= 4 \int W^3(z) k(z) (F(x_0) - h_n z f(x_0) + \cdots) dz \\
&= 4F(x_0) \int W^3(z) k(z) dz - 4h_n f(x_0) A_{3,1} + \cdots) \\
&= F(x_0) - 4h_n f(x_0) A_{3,1} + O(h_n^2)
\end{aligned}
$$

Using the above evaluations, we can obtain an approximation of $Var(\frac{1}{n}W(\frac{x_0-X_1}{h_n}))$ as follows:

$$
\begin{aligned}
& Var(\frac{1}{n}W(\frac{x_0-X_1}{h_n})) \\
= & \frac{1}{n^2}E(W_1^2) - \frac{1}{n^2}\{E(W_1)\}^2 \\
= & \frac{1}{n^2}(F(x_0) - 2h_n f(x_0)A_{1,1} + h_n^2 f^{'}(x_0)A_{1,2} + O(h_n^3)) \\
& -\frac{1}{n^2}(F^2(x_0) + h_n^2 F(x_0)f^{'}(x_0)A_{0,2} + O(h_n^3)) \\
= & \frac{1}{n^2}[F(x_0)(1-F(x_0)) - 2h_n f(x_0)A_{1,1} \\
& \quad +h_n^2 f'(x_0)(A_{1,2} - F(x_0)A_{0,2}) + O(h_n^3)]
\end{aligned}
$$

[Future problems]

- Simulation study
- Smoothing of the Wilcoxon's signed rank test (already done)
- Smoothing another rank statistics
- Best bandwidth h_n ?

• Brown, B.M., Hall, P. and Young, G.A. (2001). The smoothed median and the bootstrap, *Biometrika*, 88, 519-534.

• D'Abrera H.J.M. and Lehmann, E.L. (2006). *Nonparametrics: Statistical Methods Based on Ranks*, Springer.

• García-Soidán, P.H., González-Manteiga, W. and Prada-Sánchez, J.M. (1997). Edgeworth expansions for nonparametric distribution estimation with applications. *Jour. Stat. Plann. Inf.*, 65, 213-231.

• Hájek, J, Šidak, Z. and Sen, P.K. (1999). *Theory of Rank Tests*, Academic Press

• Huang, Z. and Maesono, Y. (2013). Improvement of the normal approximation for kernel distribution estimator. submitted

• Noether, G.E. (1955). On the theorem of Pitman, *Ann. Math. Statist.*, 26, 64-68.

Appendix. Pitman Asymptotic Relative Efficiency

Pitman's Asymptotic Relative Efficiency (A.R.E.)

Contiguous alternative $\{\theta_i\}$

$$\lim_{i\to\infty} \theta_i = \theta_0$$

$U_n,\ V_n$ satisfy

$$\lim_{n\to\infty} P_{\theta_0}(U_n \geq u_{n;\alpha}) = \lim_{n\to\infty} P_{\theta_0}(V_n \geq v_{n;\alpha}) = \alpha$$

$(0 < \alpha < 1)$. For natural positive numbers $\{m_i\},\ \{n_i\}$ $(i = 1, 2, \cdots)$, $U_n,\ V_n$ satisfy

$$\lim_{i\to\infty} P_{\theta_i}(U_{m_i} \geq u_{m_i;\alpha}) = \lim_{i\to\infty} P_{\theta_i}(V_{n_i} \geq v_{n_i;\alpha}) = \beta$$

$(0 < \beta < 1)$

$$ARE(U|V) = \lim_{i\to\infty} \frac{n_i}{m_i}$$

If $ARE(U|T)$ does not depend on $\alpha,\ \beta$, $ARE(U|T)$ is the Pitman's A.R.E. of U_n to V_n

Let us define

$$\lim_{n\to\infty} \frac{\mu'_{U_n}(\theta_0)}{\sqrt{n\sigma^2_{U_n}(\theta_0)}} = e(U), \qquad \lim_{n\to\infty} \frac{\mu'_{V_n}(\theta_0)}{\sqrt{n\sigma^2_{V_n}(\theta_0)}} = e(V)$$

Then Under some regularity conditions (Noether (1955)), we have

$$ARE(U|V) = [\frac{e(U)}{e(V)}]^2$$

A.R.E. of S and W

$$e(T) = \frac{1}{\sigma}, \qquad e(S) = 2f(0), \qquad e(W) = \sqrt{12}\int_{-\infty}^{\infty} f^2(x)dx$$

where $\sigma^2 = V(X_1)$

Pitman's A.R.E.

	Normal	Logistic	double exponential
$ARE(S\|T)$	$\frac{2}{\pi}$	$\frac{\pi^2}{12}$	2
$ARE(W\|T)$	$\frac{3}{\pi}$	$\frac{\pi^2}{9}$	$\frac{3}{2}$

double exponential $f(x) = \exp\{-\frac{1}{2}|x|\}$, S is better

logistic $f(x) = e^{-x}/(1+e^{-x})^2$ のとき, W is better

These results coincide with the theory of locally most power signed rank test

Some problems :

- Since S and W have discrete distributions, we cannot obtain critical point for any level α
- In many cases, the significance probability of Wilcoxon's signed rank test W is larger than that of the sign test S (when n is small)