Jackknife variance estimator for the sample median

Introduction

If $X_1, \cdots, X_n$ are independent continuous random variables with common density $f(x)$ and θ is the median of the distribution (that is, $F^{-1}(1/2) = \theta$) then we know that an approximation to the variance of the sample median (or more precisely, any order statistic $X_{(k)}$ with $k/n \approx 1/2$) is given by $1/(4nf^2(\theta))$. In order to use this approximation to obtain an estimate of the standard error of the sample median, we would need to obtain a good estimate of $f(\theta)$. However, densities are not that easy to estimate and so we might think to use the jackknife to estimate the variance of the sample median.

However, as mentioned in lecture, the jackknife variance estimator fails as an variance estimator for the sample median (and, in fact, for all sample quantiles). Essentially, the reason for this failure is the fact that the sample median is not sufficiently well-approximated by an average of random variables. In the next section, we will look at the behaviour of the jackknife variance estimator when the sample size is even.

The jackknife and spacings

To get a sense of why the jackknife fails, we will consider the sample median when the sample size is even, i.e. $n = 2m$. (A similar analysis can be done when n is odd but is somewhat messier!)

Define $\widehat{\theta}$ to be the usual sample median (that is, the average of the two middle order statistics) and $\widehat{\theta}_{-i}$ to the sample median with X_i deleted from the sample. Then it follows that the leave-one-out estimators satisfy $\widehat{\theta}_{-i} = X_{(m)}$ for m values of i and $\widehat{\theta}_{-i} = X_{(m+1)}$ for m values of i. Now

$$\widehat{\theta}_{\bullet} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{-i} = \frac{1}{2}(X_{(m)} + X_{(m+1)})$$

and so

$$\begin{aligned}\widehat{\mathrm{Var}}_{\mathrm{jack}}(\widehat{\theta}) &= \frac{n-1}{n}\sum_{i=1}^{n}(\widehat{\theta}_{-i} - \widehat{\theta}_{\bullet})^2 \\ &= \frac{n-1}{4}(X_{(m+1)} - X_{(m)})^2.\end{aligned}$$

If the jackknife variance estimator works then it should be approximately equal to $1/(4nf^2(\theta))$ for large n; in other words, $n\widehat{\mathrm{Var}}_{\mathrm{jack}}(\widehat{\theta})$ should converge to $1/(4f^2(\theta))$. However,

$$n\widehat{\mathrm{Var}}_{\mathrm{jack}}(\widehat{\theta}) = \frac{n-1}{4n}\left\{n(X_{(m+1)} - X_{(m)})\right\}^2.$$

Using our results on spacings, it follows that

$$n(X_{(m+1)} - X_{(m)}) \xrightarrow{d} \frac{1}{f(\theta)}W$$

where W has an Exponential distribution with mean 1. Thus

$$n\widehat{\mathrm{Var}}_{\mathrm{jack}}(\widehat{\theta}) \xrightarrow{d} \frac{1}{4f^2(\theta)}W^2$$

where the limit is a random variable (taking values on the positive real line) and not a constant. Note that $P(W^2 \leq 1) = P(W \leq 1) = 1 - \exp(-1) = 0.632$, which suggests that the jackknife variance estimator will tend to underestimate the variance of the median more than it will overestimate the variance of the median.

Example: Naive confidence intervals using the jackknife

Naively, we might use the interval

$$\left[\widehat{\theta} - 1.96 \times \left\{\widehat{\mathrm{Var}}_{\mathrm{jack}}(\widehat{\theta})\right\}^{1/2}, \widehat{\theta} + 1.96 \times \left\{\widehat{\mathrm{Var}}_{\mathrm{jack}}(\widehat{\theta})\right\}^{1/2}\right]$$

as an approximate 95% confidence interval for the population median $\theta = F^{-1}(1/2)$.

The following R code estimates the coverage of this naive 95% confidence interval in the case of a Logistic distribution with a sample size of $n = 50$ with data generated using $theta = F^{-1}(1/2) = 0$. In the simulation, we generate 10000 samples of size $n = 50$ from the Logistic distribution with $\theta = 0$ and for each sample, compute the upper and lower limits of the naive confidence interval; we can estimate the true coverage probability by

$$\text{estimated coverage probability} = \frac{\text{number of intervals containing 0}}{10000}.$$

(This estimate will be accurate to about 0.01.)

```
> cover <- 0
> for (i in 1:10000) {
+    x <- sort(rlogis(50))  # 50 observations from Logistic median 0
+    med <- median(x)
+    jackse <- sqrt(49*(x[26]-x[25])^2/4)  # jackknife standard error
+    lower <- med - 1.96*jackse # lower limit
+    upper <- med + 1.96*jackse # upper limit
+    if (upper*lower<=0) cover <- cover + 1
+    }
> cover/10000  # estimated coverage probability
[1] 0.6979
```

Thus the coverage of the naive 95% confidence interval is actually closer to 70%. This is consistent with our observation in the previous section that the jackknife tends to underestimate the variance of the sample median.

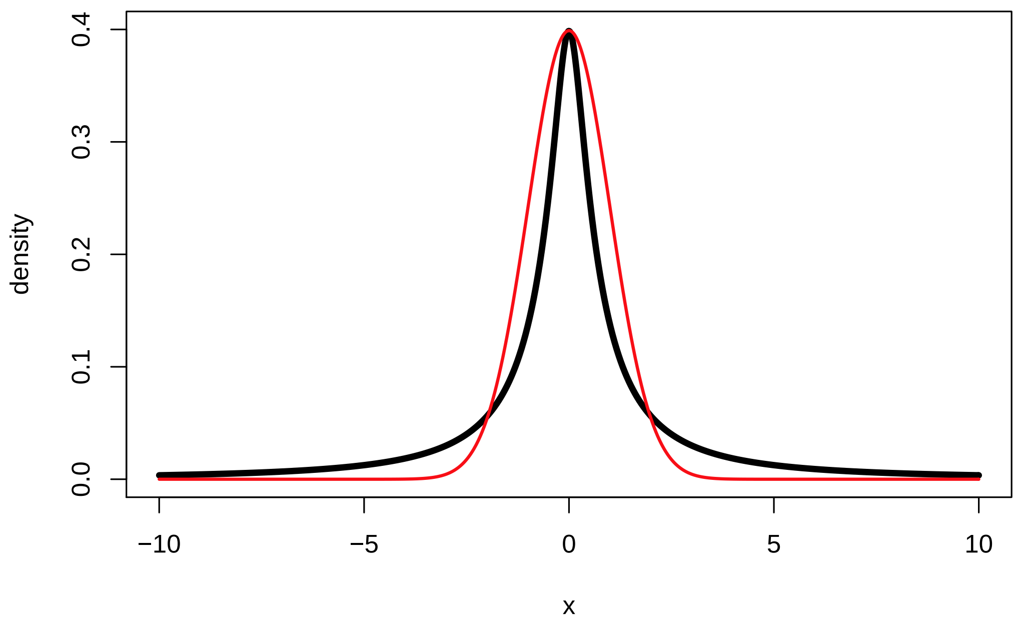

Figure 1: Density of V (black) with a $\mathcal{N}(0,1)$ density (red) for comparison

Some theory

Can the coverage probability in this example be explained theoretically? Note that this naive jackknife confidence interval uses the pivot

$$\frac{\widehat{\theta}-\theta}{\left\{\widehat{\mathrm{Var}}_{\mathrm{jack}}(\widehat{\theta})\right\}^{1/2}},$$

which converges in distribution (for even sample sizes $n = 2m$ as $m \to \infty$) to $V = Z/W$ where Z and W are independent with $Z \sim \mathcal{N}(0,1)$ and W Exponential with mean 1. The density of V is[1]

$$\begin{aligned} f_V(x) &= \frac{1}{\sqrt{2\pi}}\int_0^\infty w\exp(-w)\exp\left(-\frac{w^2x^2}{2}\right)dw \\ &= \frac{1}{x^2\sqrt{2\pi}} - \frac{\exp(1/(2x^2))[1-\Phi(1/|x|)]}{|x|^3} \end{aligned}$$

and the distribution function of V is

$$F_V(x) = \int_0^\infty \exp(-w)\Phi(wx)\,dw$$

[1]The conditional distribution of V given $W = w$ is $\mathcal{N}(0, 1/w^2)$ and so we can determine the density of V by integrating the conditional density multiplied by the marginal density of W.

where $\Phi(t)$ is the $\mathcal{N}(0,1)$ distribution function. Note that for larger values of $|x|$ (for example, $|x| > 4$),

$$f_V(x) \approx \frac{1}{x^2\sqrt{2\pi}} - \frac{1}{2|x|^3},$$

which indicates that the tails of V are very heavy; for example, $E(|V|) = E(|Z|)E(W^{-1}) = \infty$ since $E(W^{-1}) = \infty$. The density $f_V(x)$ is shown in Figure 1 with a $\mathcal{N}(0,1)$ density for comparison.

It is possible to numerically integrate to evaluate $F_V(x)$ at a given x but $F_V(x)$ can be approximated very easily by the simple Monte Carlo estimate

$$\widehat{F}_V(x) = \frac{1}{N}\sum_{i=1}^{N} \Phi(W_i x)$$

where $W_1, \cdots, W_N$ are independent Exponential random variables with mean 1. Likewise, we can estimate the density $f_V(x)$ by

$$\widehat{f}_V(x) = \frac{1}{N}\sum_{i=1}^{N} W_i \phi(W_i x) = \frac{1}{N}\sum_{i=1}^{N} \frac{W_i}{\sqrt{2\pi}} \exp\left(-\frac{W_i^2 x^2}{2}\right)$$

Using $N = 10^6$, we can estimate the large sample coverage probability, $P(-1.96 \le V \le 1.96)$, of the naive confidence interval using the following R code:

```
> w <- rexp(1000000)
> mean(pnorm(1.96,0,1/w)-pnorm(-1.96,0,1/w))
[1] 0.6947249
```

This estimate is consistent with the coverage probability that we saw in the Logistic example.

If we consider confidence intervals of the form

$$\left[\widehat{\theta} - t_p \times \left\{\widehat{\mathrm{Var}}_{\mathrm{jack}}(\widehat{\theta})\right\}^{1/2}, \widehat{\theta} + t_p \times \left\{\widehat{\mathrm{Var}}_{\mathrm{jack}}(\widehat{\theta})\right\}^{1/2}\right]$$

how large does t_p need to be in order for the confidence interval to have (approximately) $100p\%$ coverage? To do this, we simply need to find t_p such that $P(-t_p \le V \le t_p) = p$. Table 1 gives the values of t_p for various values of p.

p	0.50	0.70	0.80	0.90	0.95	0.99
t_p	0.92	2.01	3.35	7.34	15.33	79.16

Table 1: Values of t_p for various confidence levels p

We can now repeat the earlier experiment now using $t_{0.95} = 15.33$, which hopefully should give a coverage for the confidence interval closer to 0.95.

```
> cover <- 0
> for (i in 1:10000) {
+    x <- sort(rlogis(50))  # 50 observations from Logistic median 0
+    med <- median(x)
+    jackse <- sqrt(49*(x[26]-x[25])^2/4)  # jackknife standard error
+    lower <- med - 15.33*jackse # lower limit
+    upper <- med + 15.33*jackse # upper limit
+    if (upper*lower<=0) cover <- cover + 1
+    }
> cover/10000  # estimated coverage probability
[1] 0.9534
```

Towards a better confidence interval

The key to finding a better confidence interval lies in finding a better estimator of $1/f(\theta)$; in the jackknife with $n = 2m$, this is estimated by $n(X_{(m+1)} - X_{(m)})$.

Suppose that a and b are positive integers with $a < n/2 < b$ where $r = b - a$ is relatively small compared to n. Then

$$\frac{n}{r}(X_{(b)} - X_{(a)}) = \frac{1}{r}\sum_{k=a}^{b-1} n(X_{(k+1)} - X_{(k)}).$$

If a and b are "close" to $n/2$ then

$$\Big(n(X_{(a+1)} - X_{(a)}), \cdots, n(X_{(b)} - X_{(b-1)})\Big) \xrightarrow{d} (E_1, \cdots, E_r)$$

where $E_1, \cdots, E_r$ are independent Exponential random variables with mean $1/f(\theta)$. This in turn suggests that the distribution of $n(X_{(b)} - X_{(a)})/r$ can be approximated by the distribution of an average of r independent Exponential random variables with mean $1/f(\theta)$.

How can we use this to construct a confidence interval for θ, the population median? The idea is essentially to choose a and b so that

- $r = b - a$ is as large as possible, and
- the Exponential (with mean $1/f(\theta)$) approximations to the distributions of $n(X_{(a+1)} - X_{(a)}), \cdots, n(X_{(b)} - X_{(b-1)})$ are valid.

If these conditions hold then we can estimate the variance of $\widehat{\theta}$ by

$$\widehat{\mathrm{Var}}_r(\widehat{\theta}) = \frac{n(X_{(b)} - X_{(a)})^2}{4r^2}$$

and use the pivot

$$\frac{\widehat{\theta} - \theta}{\left\{\widehat{\mathrm{Var}}_r(\widehat{\theta})\right\}^{1/2}}$$

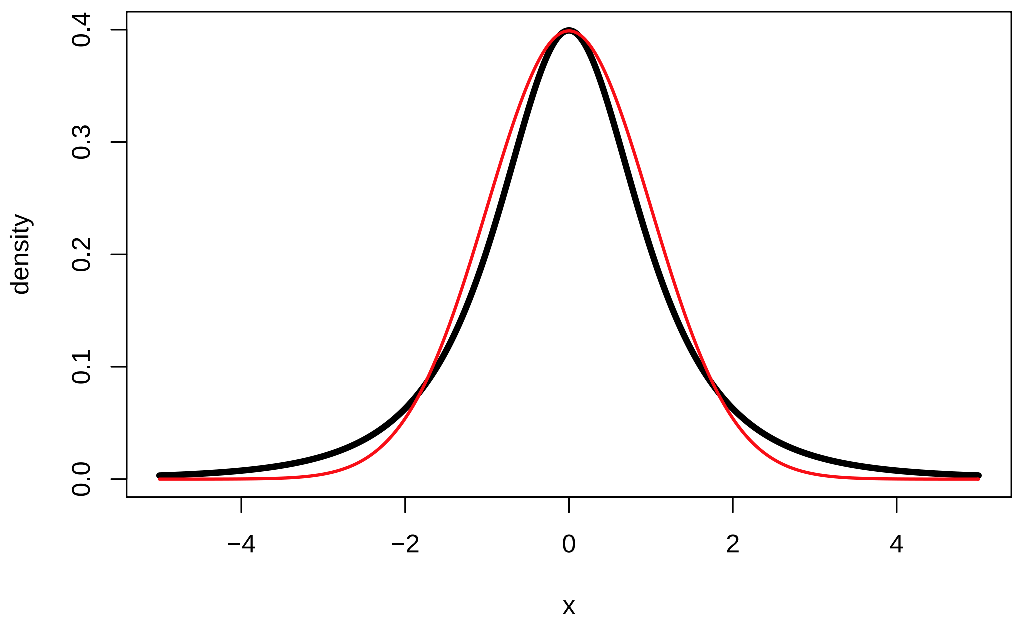

Figure 2: Density of V for $r = 5$ (black) with a $\mathcal{N}(0,1)$ density (red) for comparison

to construct a confidence interval for θ.

We can approximate the distribution of the pivot using a similar approach to that used earlier. As $n \to \infty$, the distribution of the pivot converges to the distribution of $V = Z/W$ where (as before) $Z \sim \mathcal{N}(0,1)$ and W (independent of Z) has a Gamma distribution with density

$$g_W(x) = \frac{r^r x^{r-1} \exp(-rx)}{\Gamma(r)} \quad \text{for } x \geq 0$$

(where $r = b - a$). From this, we can obtain the distribution and density functions of V as follows:

$$\begin{aligned} F_V(x) &= \int_0^\infty g_W(w)\Phi(wx)\, dw \\ f_V(x) &= \int_0^\infty w\, g_W(w)\phi(wx)\, dw. \end{aligned}$$

As r increases, the distribution of V gets closer a $\mathcal{N}(0,1)$ distribution. Figure 2 shows the density of V for $r = 5$ with a $\mathcal{N}(0,1)$ density for comparison.