Performance Evaluation of Stein's Two-Stage Estimation Procedure

Zhou Wang

Capstone Seminar

Department of Mathematical Sciences
Binghamton University

April 4, 2019

Background

Let $X_1, X_2, \cdots$ be a sample drawn from a normal population with mean μ and variance σ^2, where both μ and σ are unknown. We wish to construct a $(1-\alpha)\%$ CI, say I, for μ with a given length $2d > 0$, which means

$$\mathbb{P}_{\mu,\sigma}(\mu \in I) \geq 1-\alpha \quad \text{for all } \mu \text{ and } \sigma; \text{ length of } I \text{ is } 2d.$$

There is no fixed sample size estimation method to solve this problem but it can be solved sequentially.

Background

For fixed sample size n, we know

$$I = \bar{X}_n \pm t_{1-\frac{\alpha}{2}}(n-1)\frac{S_n}{\sqrt{n}},$$

but it cannot guarantee the length of I to be fixed because of unknown σ.

However, if we know σ, then for fixed CI length $2d$, the ideal (minimum) sample size is

$$n_{\text{opt}} = \left\lceil z^2_{1-\frac{\alpha}{2}}\sigma^2/d^2\right\rceil.$$

Background

Stage 1: Pick a pilot size $m \geq 2$ and sample $X_1, X_2, \cdots, X_m$. Then compute sample variance $S_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X}_m)$ and $\hat{n}_{\text{opt}} = \left\lceil t^2_{1-\frac{\alpha}{2}}(m-1)S_m^2/d^2 \right\rceil$.

If $m > \hat{n}_{\text{opt}} \Rightarrow$ Stop. Otherwise, go to

Stage 2: Sample $\hat{n}_{\text{opt}} - m$ more observations.

Background

Therefore, the total sample size

$$N = \max\left\{m, \left\lceil t^2_{1-\frac{\alpha}{2}}(m-1)S_m^2/d^2 \right\rceil\right\}.$$

Based on this method, $I_N = \bar{X}_N \pm d$ obviously has the length $2d$. Moreover,

$$P_{\mu,\sigma}(\mu \in I_N) \geq 1 - \alpha,$$

which was shown by Charles Stein in 1945.

Find the distribution for N

Characteristics of interest for two-stage estimation:

i) $\mathbb{E}_{\mu,\sigma}[N]$

ii) $\text{Var}_{\mu,\sigma}[N]$

iii) $\mathbb{P}_{\mu,\sigma}(\mu \in I_N)$

iv) $\mathbb{E}_{\mu,\sigma}[N]/n_{\text{opt}}$

Find the distribution for N

The potential value N takes are $m, m+1, m+2, \cdots$.

$$
\begin{aligned}
\mathbb{P}_{\mu,\sigma}(N=m) &= \mathbb{P}_{\mu,\sigma}\left(m \geq \left\lceil t^2_{1-\frac{\alpha}{2}}(m-1)S^2_m/d^2 \right\rceil\right) \\
&= \mathbb{P}_{\mu,\sigma}\left(m \geq t^2_{1-\frac{\alpha}{2}}(m-1)S^2_m/d^2\right) \\
&= \mathbb{P}_{\mu,\sigma}\left(\frac{(m-1)S^2_m}{\sigma^2} \leq \frac{m(m-1)d^2}{\sigma^2 t^2_{1-\frac{\alpha}{2}}(m-1)}\right) \\
&= \mathbb{P}_{\mu,\sigma}\left(\chi^2(m-1) \leq \frac{m(m-1)d^2}{\sigma^2 t^2_{1-\frac{\alpha}{2}}(m-1)}\right)
\end{aligned}
$$

Find the distribution for N

The potential value N takes are $m, m+1, m+2, \cdots$.

$$\begin{aligned}
\mathbb{P}_{\mu,\sigma}\left(N = m+k\right) &= \mathbb{P}_{\mu,\sigma}\left(m+k = \left\lceil t^2_{1-\frac{\alpha}{2}}(m-1)S_m^2/d^2 \right\rceil\right) \\
&= \mathbb{P}_{\mu,\sigma}\left(m+k-1 < t^2_{1-\frac{\alpha}{2}}(m-1)S_m^2/d^2 \leq m+k\right) \\
&= \mathbb{P}_{\mu,\sigma}\left(\frac{(m+k-1)(m-1)d^2}{\sigma^2 t^2_{1-\frac{\alpha}{2}}(m-1)} < \frac{(m-1)S_m^2}{\sigma^2} \leq \frac{(m+k)(m-1)d^2}{\sigma^2 t^2_{1-\frac{\alpha}{2}}(m-1)}\right) \\
&= \mathbb{P}_{\mu,\sigma}\left(\frac{(m+k-1)(m-1)d^2}{\sigma^2 t^2_{1-\frac{\alpha}{2}}(m-1)} < \chi^2(m-1) \leq \frac{(m+k)(m-1)d^2}{\sigma^2 t^2_{1-\frac{\alpha}{2}}(m-1)}\right),
\end{aligned}$$

where $k = 1, 2, 3, \cdots$.

Find the distribution for N

$$\mathbb{E}_{\mu,\sigma}[N] = \sum_{n=m}^{\infty} n\mathbb{P}_{\mu,\sigma}(N = n), \text{ Var}_{\mu,\sigma}[N] = \mathbb{E}_{\mu,\sigma}[N^2] - \mathbb{E}^2_{\mu,\sigma}[N].$$

Find the distribution for N

$\mathbb{E}_{\mu,\sigma}[N] = \sum_{n=m}^{\infty} n\mathbb{P}_{\mu,\sigma}(N = n)$, $\text{Var}_{\mu,\sigma}[N] = \mathbb{E}_{\mu,\sigma}[N^2] - \mathbb{E}^2_{\mu,\sigma}[N]$.

$$
\begin{aligned}
\mathbb{P}_{\mu,\sigma}(\mu \in I_N) &= \mathbb{P}_{\mu,\sigma}(|\bar{X}_N - \mu| \le d) \\
&= \sum_{n=m}^{\infty} \mathbb{P}_{\mu,\sigma}(|\bar{X}_n - \mu| \le d \bigcap N = n) \\
&= \sum_{n=m}^{\infty} \mathbb{P}_{\mu,\sigma}(|\bar{X}_n - \mu| \le d)\mathbb{P}_{\mu,\sigma}(N = n) \\
&= \sum_{n=m}^{\infty} \left[2\Phi(\sqrt{n}d/\sigma) - 1\right] \mathbb{P}_{\mu,\sigma}(N = n).
\end{aligned}
$$

Experimental Results

Compute $\mathbb{E}_{\mu,\sigma}[N]$, $\sqrt{\text{Var}_{\mu,\sigma}[N]}$ and $\mathbb{P}_{\mu,\sigma}(\mu \in I_N)$ for below 4 s-cenarios:

i) $\alpha = 0.05, \sigma = 1, d = 0.5$

ii) $\alpha = 0.05, \sigma = 2, d = 0.5$

iii) $\alpha = 0.1, \sigma = 1, d = 0.5$

iv) $\alpha = 0.1, \sigma = 1, d = 0.3$

Experimental Results

```r
twoStageSamp <- function(m, d, sigma, alpha){
  X <- rnorm(m, 0, sigma); nopt <- ceiling((qnorm(1 - alpha/2) * sigma/d)^2)
  Ntilde <- ceiling((qt(1 - alpha/2, m - 1)/d)^2 * var(X))
  N <- max(m, Ntilde); X <- c(X, rnorm(N - m, 0, sigma))

  Ncandi <- m:(50 * N)
  Qchi <- c(0, Ncandi) * (m - 1) * (d/(sigma * qt(1 - alpha/2, m - 1)))^2
  dist <- diff(pchisq(Qchi, m - 1))

  ch <- t(dist) %*% cbind(Ncandi, Ncandi^2, pnorm(d * sqrt(Ncandi)/sigma))
  list(DistN = dist, EN = ch[1], SigN = sqrt(ch[2] - ch[1]^2),
       CovProb = 2 * ch[3] - 1, Nopt = nopt)
}
```

Experimental Results

m	$\alpha = 0.05, \sigma = 1, d = 0.5$			$\alpha = 0.05, \sigma = 2, d = 0.5$		
	$\mathbb{E}[N]$	$\sqrt{\text{Var}[N]}$	$\mathbb{P}(\mu \in I_N)$	$\mathbb{E}[N]$	$\sqrt{\text{Var}[N]}$	$\mathbb{P}(\mu \in I_N)$
5	31.389	21.736	0.958	123.842	87.209	0.951
10	21.196	9.346	0.960	82.379	38.596	0.951
20	21.473	3.081	0.978	70.594	22.738	0.951
30	30.015	0.247	0.994	67.439	17.552	0.951
500	500	0	1	500	0	1
n_{opt}	16			62		

Experimental Results

m	$\alpha = 0.1, \sigma = 1, d = 0.5$			$\alpha = 0.1, \sigma = 1, d = 0.3$		
	$\mathbb{E}[N]$	$\sqrt{\text{Var}[N]}$	$\mathbb{P}(\mu \in I_N)$	$\mathbb{E}[N]$	$\sqrt{\text{Var}[N]}$	$\mathbb{P}(\mu \in I_N)$
5	18.822	12.690	0.918	51.019	35.680	0.904
10	14.717	5.520	0.929	37.864	17.556	0.904
20	20.093	0.635	0.975	33.970	10.398	0.906
30	30.000	0.005	0.994	34.698	6.242	0.919
500	500	0	1	500	0	1
n_{opt}	11			31		

Plot $\mathbb{P}_{\mu,\sigma}(N = m + k)$ for $k = 0, 1, 2, \cdots$.

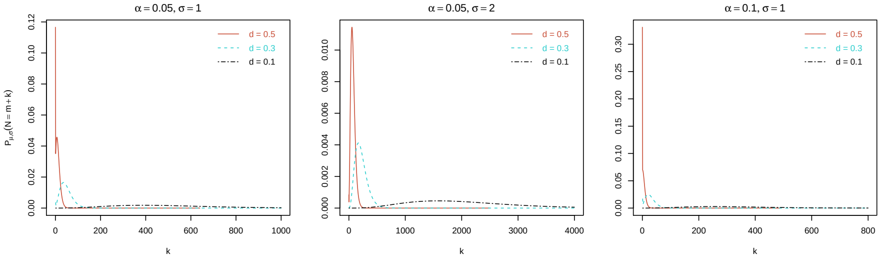

Figure: Plot for $P_{\mu,\sigma}(N = m + k)$ given $m = 10$

In each picture, the peak of probability and its corresponding k will drop and increase respectively as d decreases.

Plot $\mathbb{E}_{\mu,\sigma}[N]$ as a function of m.

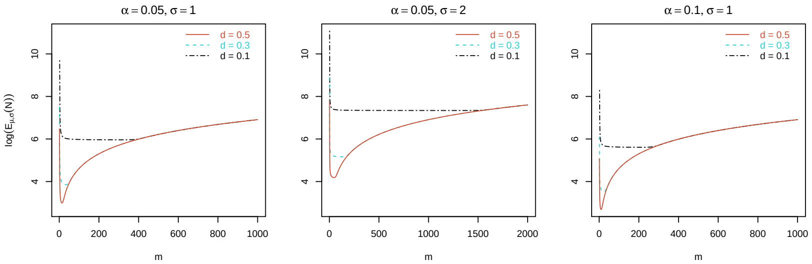

Figure: Plot for $\log \mathbb{E}_{\mu,\sigma}[N]$

All the curves for different scenarios go down first and then go up. For fixed α and σ, $\mathbb{E}_{\mu,\sigma}[N]$'s from different scenarios are going to be close when m is really large.

Plot $\inf \mathbb{E}_{\mu,\sigma}[N]$ and n_{opt} as functions of d.

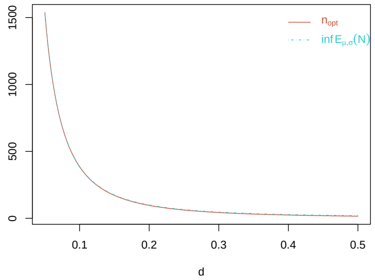

Figure: Plot for $\inf \mathbb{E}_{\mu,\sigma}[N]$ and n_{opt}

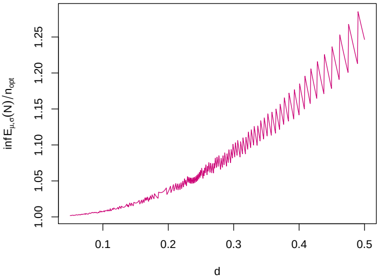

Figure: Plot for $\dfrac{\inf \mathbb{E}_{\mu,\sigma}[N]}{n_{\text{opt}}}$

The two figures assume that $\sigma = 1$, $\alpha = 0.05$ and $d \in (0.05, 0.5)$.

Inspect the behavior of $\lim_{d\to 0} \mathbb{P}_{\mu,\sigma}(\mu \in I_N)$.

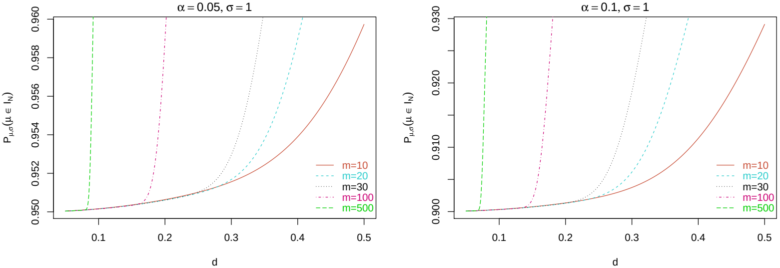

Figure: Plot for coverage probability

Both plots show that $\mathbb{P}_{\mu,\sigma}(\mu \in I_N)$ converges to $1-\alpha$ as d goes down to 0. Moreover, the larger m is, the faster coverage probability converges.

Conclusion

i) Stein's method tends to oversample.

ii) It is asymptotically $(d \downarrow 0)$ consistent.

Thanks!