# Asymptotic Normality of Maximum Likelihood Estimators and A Discussion on their Finite Normal Approximations

Naitong Chen

December 7, 2020

# 1 Body

First we set up the notaions and assumptions that are used throughout the report.

## 1.1 Notations and Assumptions

Let $X_1, \cdots, X_n$ be i.i.d. continuous random variables in $\mathbb{R}$ with probability density function $f(x;\theta_0)$, where $\theta_0 \in \Theta$ is a unknown parameter.

Write the likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta), \tag{1}$$

and subsequently the log-likelihood function

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta). \tag{2}$$

Then the maximum likelihood estimator can be computed as

$$\hat{\theta} = \arg\max_{\theta\in\Theta} l(\theta) = \arg\max_{\theta\in\Theta} L(\theta). \tag{3}$$

Note then since conditons , a necessary consequence is that

$$\left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\theta_0} = 0. \tag{4}$$

We use $\mathbb{E}_\theta[X]$ to denote expectation with respect to $X$ under $f(x;\theta)$. Using this notation we define the Fisher information

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right]. \tag{5}$$

Throughout the report, we assume the following regularity conditions.

- (R1) $f(x;\theta)$ is identifiable: $\theta_1 \neq \theta_2 \implies f(x;\theta_1) \neq f(x;\theta_2)$.
- (R2) $f(x;\theta)$ has common support for all $\theta \in \Theta$.
- (R3) $\theta_0$ is an iterior point in $\Theta$.
- (R4) $f(x;\theta)$ is twice differentiable in $\theta$.
- (R5) The integral $\int f(x;\theta)$ can be different twice in $\theta$ under the integral sign.
- (R6) $\hat{\theta}$ is the unique solution to $\frac{\partial l(\theta)}{\theta} = 0$.

## 1.2 Intermediate Results

We prove two lemmas that are essential to developing the main results of this report.

**Lemma 1**: Under regularity conditions, (define $P_{\theta_0}$),

$$\lim_{n\to\infty} P_{\theta_0}\left(L(\theta_0 \mid X) > L(\theta \mid X)\right) = 1, \forall\theta \neq \theta_0. \tag{6}$$

*Proof.* We begin by taking the log on both sides of the inequality on the LHS and rearrange.

$$L(\theta_0 \mid X) > L(\theta \mid X) \implies \sum_{i=1}^{n} \log f(X_i;\theta_0) > \sum_{i=1}^{n} \log f(X_i;\theta) \tag{7}$$

$$\implies \sum_{i=1}^{n} \left(\log f(X_i;\theta_0) - \log f(X_i;\theta)\right) < 0 \tag{8}$$

$$\implies \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) < 0. \tag{9}$$

Since $X_i$'s are i.i.d. , the summands are independent, and so by the Weak Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) \xrightarrow{p} \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]. \tag{10}$$

Note that $-\log(\cdot)$ is strictly convex, then by Jensen's inequality, we can establish, on the RHS of the above equation,

$$-\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] = \mathbb{E}_{\theta_0}\left[-\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] > -\log \mathbb{E}_{\theta_0}\left[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right]. \tag{11}$$

Now note

$$\log \mathbb{E}_{\theta_0}\left[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right] = \log \int_{\mathbb{R}} \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)dx = \log \int_{\mathbb{R}} \frac{f(x;\theta)}{d}x = \log 1 = 0. \tag{12}$$

Together, we have

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) \xrightarrow{p} \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] < 0. \tag{13}$$

To show the desired equation, it is equivalent to show, by Eq. (9),

$$\lim_{n\to\infty} P_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) < 0\right) = 1. \tag{14}$$

By Eq. (13), we know that for all $\epsilon > 0$,

$$\lim_{n\to\infty} P_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) - \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]\right| < \epsilon\right) = 1. \tag{15}$$

Again by rearranging the inequality inside, we get

$$\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] - \epsilon < \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) < \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] + \epsilon. \tag{16}$$

Note that the probability of event Eq. (16) is less than or equal to that of

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) < \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] + \epsilon. \tag{17}$$

Since $\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] < 0$, by fixing $\epsilon = -\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] > 0$, we have

$$\lim_{n\to\infty} P_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) - \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]\right| < \epsilon\right) \tag{18}$$

$$\leq \lim_{n\to\infty} P_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) < \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] + \epsilon\right) \tag{19}$$

$$= \lim_{n\to\infty} P_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) < 0\right) = 1. \tag{20}$$

Therefore, we conclude that

$$\lim_{n\to\infty} P_{\theta_0}\left(L(\theta_0 \mid X) > L(\theta \mid X)\right) = 1, \forall\theta \neq \theta_0. \tag{21}$$

□

**Lemma 2**: Under regularity conditions,

$$\mathbb{E}\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right) = 0, \text{ and } I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right] = \mathrm{Var}\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right). \tag{22}$$

*Proof.* We begin with differentiating both sides of $1 = \int_{\mathbb{R}} f(x;\theta)dx$.

$$0 = \frac{\partial}{\partial\theta}\int_{\mathbb{R}} f(x;\theta)dx \tag{23}$$

$$= \int_{\mathbb{R}} \frac{\partial}{\partial\theta} f(x;\theta)dx \tag{24}$$

$$= \int_{\mathbb{R}} \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} f(x;\theta)dx \tag{25}$$

$$= \int_{\mathbb{R}} \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx \tag{26}$$

$$= \mathbb{E}_\theta\left[\frac{\partial \log f(x;\theta)}{\partial\theta}\right]. \tag{27}$$

Differentiating with respect to $\theta$ again, we get

$$0 = \frac{\partial}{\partial\theta}\int_{\mathbb{R}} \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx \tag{28}$$

$$= \int_{\mathbb{R}} \frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} f(x;\theta)dx + \int_{\mathbb{R}} \frac{\partial \log f(x;\theta)}{\partial\theta}\frac{\partial f(x;\theta)}{\partial\theta}dx \tag{29}$$

$$= \int_{\mathbb{R}} \frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} f(x;\theta)dx + \int_{\mathbb{R}} \frac{\partial \log f(x;\theta)}{\partial\theta}\frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx \tag{30}$$

$$= \mathbb{E}_\theta\left[\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right] + \mathbb{E}_\theta\left[\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2\right], \tag{31}$$

where Eq. (30) used the same trick as Eq. (26). By Eq. (5), we have

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right]. \tag{32}$$

Finally, since $\mathbb{E}_\theta\left[\frac{\partial \log f(x;\theta)}{\partial\theta}\right]=0$,

$$\mathrm{Var}\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)=\mathbb{E}_\theta\left[\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2\right]-\left(\mathbb{E}_\theta\left[\frac{\partial \log f(x;\theta)}{\partial\theta}\right]\right)^2=I(\theta). \tag{33}$$

Together, we conclude that

$$\mathbb{E}\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)=0,\ \text{and}\ I(\theta)=-\mathbb{E}_\theta\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right]=\mathrm{Var}\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right). \tag{34}$$

□

## 1.3 Asymptotic Normality of MLE

We first establish the consistency of MLE.

**Lemma 3**: Under regularity conditions, $\hat{\theta}$ is a consistent estimator of $\theta_0$.

*Proof.* We begin by showing that the equation $\frac{\partial l(\theta|x)}{\partial\theta}=0$ has a solution $\hat{\theta}_n$ that converges in probability to $\theta_0$.

Since $\theta_0$ is an iterior point of $\Theta$, we can find $a>0$ such that $\theta_0\in(\theta_0-a,\theta_0+a)\subset\Theta$. Then define the event

$$S_n=\{x: l(\theta_0\mid x)>l(\theta_0-a\mid x)\cap l(\theta_0\mid x)>l(\theta_0+a\mid x)\}. \tag{35}$$

Lemma 1 says, under $P_{\theta_0}$, when $n$ approaches infinity, $\theta_0$ is the unique maximizer to $L(\theta\mid x)$, which implies that it is also the unique maximizer to $l(\theta\mid x)$. Then clearly $\lim_{n\to\infty}P_{\theta_0}(S_n)=1$.

Note since $l(\theta_0\mid x)>l(\theta_0-a\mid x)$ and $l(\theta_0\mid x)>l(\theta_0+a\mid x)$, with $f(\theta;x)$ being continous and differentiable, there must exists, for any $x\in S_n$, a local maximum in $(\theta_0-a,\theta_0+a)$. Denote this value $\hat{\theta}_n$, then $\frac{\partial l(\theta|x)}{\partial\theta}\Big|_{\theta=\hat{\theta}_n}=0$.

Then for all $a>0$ small enough, we can find a sequence of $\hat{\theta}$ such that

$$\lim_{n\to\infty}P_{\theta_0}\left(|\hat{\theta}_n-\theta_0|<a\right)=1. \tag{36}$$

By choosing $\hat{\theta}_n$ to be the one closest to $\theta_0$, denoted $\theta_n^*$, we have identified a sequence $(\theta_n^*)_{n\geq1}$, independent of $a$, such that

$$\lim_{n\to\infty}P_{\theta_0}\left(|\hat{\theta}_n-\theta_0|<a\right)=1,\forall a>0. \tag{37}$$

This precisely means that $\hat{\theta}_n\xrightarrow{p}\theta_0$.

Now note that under regularity conditions, $\hat{\theta}_n$ is the unique solution to $\frac{\partial l(\theta|x)}{\partial\theta}=0$. Therefore, we conclude that $\hat{\theta}$ is a consistent estimator of $\theta_0$. □

We now use this fact to prove the asymptotic normality of MLE.

**Theorem 1**: Under regularity conditions,

$$\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{d}\mathcal{N}\left(0,[I(\theta_0)]^{-1}\right), \tag{38}$$

where $I(\theta_0)=\mathbb{E}_{\theta_0}\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right)^2\right]$.

*Proof.* By the Mean Value Theorem, for $f : \mathbb{R} \to \mathbb{R}$ continuous on $[a, b]$ and differentiable on $(a, b)$, for all $c \in (a, b)$,

$$\frac{f(a) - f(b)}{a - b} = f'(c). \tag{39}$$

Let $f(\theta) = l'(\theta) = \frac{\partial l(\theta)}{\partial \theta}, a = \hat{\theta}, b = \theta_0, c = \theta_1 \in (\theta_0, \hat{\theta})$. Then with $l''(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta^2}$, the above equation becomes

$$\frac{l'(\hat{\theta}) - l'(\theta_0)}{\hat{\theta} - \theta_0} = l''(\theta_1). \tag{40}$$

We know $l'(\hat{\theta}) = 0$, then the equation above becomes

$$0 = l'(\theta_0) + (\hat{\theta} - \theta_0) l''(\theta_1) \implies \sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\sqrt{n} l'(\theta_0)}{l''(\theta_1)}. \tag{41}$$

We first look at the denominator of Eq. (41). By Lemma 3, $\hat{\theta} \xrightarrow{p} \theta_0$. Then since $\theta_1 \in (\theta_0, \hat{\theta})$, we must have $\theta_1 \xrightarrow{p} \theta_0$. Then by Proposition 10.7 from the lecture notes,

$$l''(\theta_1) \xrightarrow{p} l''(\theta_0). \tag{42}$$

Now by the Weak Law of Large Numbers and Lemma 2,

$$l''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(x_i; \theta) \Big|_{\theta=\theta_0} \xrightarrow{p} \mathbb{E}_{\theta_0} \left[ \frac{\partial^2}{\partial \theta^2} \log f(X_1; \theta) \Big|_{\theta=\theta_0} \right] = -I(\theta_0). \tag{43}$$

Now we look at the numerator of Eq. (41). By Lemma 2, $\mathbb{E}_{\theta_0} \left[ \frac{\partial}{\partial \theta} \log f(X_1; \theta) \big|_{\theta=\theta_0} \right] = 0$ and $I(\theta_0) = \mathrm{Var}\left( \frac{\partial}{\partial \theta} \log f(X_1; \theta) \big|_{\theta=\theta_0} \right)$. Then by the Central Limit Theorem,

$$\begin{aligned}
\sqrt{n} l'(\theta_0) &= \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i; \theta) \Big|_{\theta=\theta_0} && (44) \\
&= \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i; \theta) \Big|_{\theta=\theta_0} - \mathbb{E}_{\theta_0} \left[ \frac{\partial}{\partial \theta} \log f(X_1; \theta) \Big|_{\theta=\theta_0} \right] \right) && (45) \\
&\xrightarrow{d} \mathcal{N}\left( 0, \mathrm{Var}\left( \frac{\partial}{\partial \theta} \log f(X_1; \theta) \Big|_{\theta=\theta_0} \right) \right) && (46) \\
&= \mathcal{N}\left( 0, I(\theta_0) \right). && (47)
\end{aligned}$$

Together by Eqs. (41), (43) and (47) and Slutsky's Theorem,

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\sqrt{n} l'(\theta_0)}{l''(\theta_1)} \xrightarrow{d} \frac{\mathcal{N}\left( 0, I(\theta_0) \right)}{I(\theta_0)} = \mathcal{N}\left( 0, [I(\theta_0)]^{-1} \right). \tag{48}$$

Therefore we conclude that $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left( 0, [I(\theta_0)]^{-1} \right)$. □

Out of all possible estimators for $\theta_0$, how good is the MLE? The next result gives us some insight into the quality of MLE compared to others in terms of the variance of these estimators.

**Theorem 2**: Let $Y = u(X_1, \cdots, X_n)$ be an unbiased estimator of $\theta_0$ such that $\mathbb{E}_{\theta_0}[Y] = \theta_0$. Then under regularity conditions, $\mathrm{Var}(Y) \geq \frac{1}{n I(\theta_0)}$.

*Proof.* We first expand $E_{\theta_0}[Y]$.

$$\theta_0 = E_{\theta_0}[Y] = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} u(x_1, \cdots, x_n) \prod_{i=1}^{n} f(x_i; \theta_0) dx_1 \cdots dx_n. \tag{49}$$

Differentiating both sides with respect to $\theta_0$ gives

$$1 = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} u(x_1, \cdots, x_n) \left( \sum_{i=1}^{n} \frac{\partial f(x_i; \theta_0)}{\partial \theta_0} \frac{1}{f(x_i; \theta_0)} \right) \prod_{i=1}^{n} f(x_i; \theta_0) dx_1 \cdots dx_n \tag{50}$$

$$= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} u(x_1, \cdots, x_n) \left( \sum_{i=1}^{n} \frac{\partial \log f(x_i; \theta_0)}{\partial \theta_0} \right) \prod_{i=1}^{n} f(x_i; \theta_0) dx_1 \cdots dx_n \tag{51}$$

By writing $Z = \sum_{i=1}^{n} \frac{\partial \log f(X_i; \theta_0)}{\partial \theta_0}$, we get, with $\rho$ denoting the correlation coefficient between $Y$ and $Z$,

$$1 = \mathbb{E}_{\theta_0}[YZ] = \mathbb{E}_{\theta_0}[Y]\mathbb{E}_{\theta_0}[Z] + \rho\sqrt{\mathrm{Var}(Y)}\sqrt{\mathrm{Var}(Z)} \implies \rho = \frac{1}{\sqrt{\mathrm{Var}(Y)}\sqrt{\mathrm{Var}(Z)}}. \tag{52}$$

We note that since $X_i$'s are i.i.d. ,

$$\mathrm{Var}(Z) = \mathrm{Var}\left( \sum_{i=1}^{n} \frac{\partial \log f(X_i; \theta_0)}{\partial \theta_0} \right) \tag{53}$$

$$= \sum_{i=1}^{n} \mathrm{Var}\left( \frac{\partial \log f(X_i; \theta_0)}{\partial \theta_0} \right) \tag{54}$$

$$= n\mathrm{Var}\left( \frac{\partial \log f(X_1; \theta_0)}{\partial \theta_0} \right) \tag{55}$$

$$= nI(\theta_0). \tag{56}$$

Then

$$\rho = \frac{1}{\sqrt{nI(\theta_0)}\sqrt{\mathrm{Var}(Y)}}. \tag{57}$$

By definition, $rho^2 \leq 1$, then

$$\rho^2 = \frac{1}{nI(\theta_0)\mathrm{Var}(Y)} \leq 1 \implies \mathrm{Var}(Y) \geq \frac{1}{nI(\theta_0)}. \tag{58}$$

Therefore for any unbiased estimator $Y$ of $\theta_0$, we have $\mathrm{Var}(Y) \geq \frac{1}{nI(\theta_0)}$. □

Note the lower bound on the variance is exactly the asymptotic variance of $\hat{\theta}$. This means that asymptotically, the MLE achieves the smallest possible variance out of all unbiased estimators of $\theta$. While this is a nice property, it remains rather theoretical. In practice, when we work with finitely many observations, it is not clear how far away we are from asymptotic normality. And existing heuristics from undergraduate statistics courses such as calling samples larger than 25 or 30 large enough is far from satisfying. Fortunately, [AL15] have done some pioneer work aimed at answering this exact question. We turn in the next section for a brief illustration of their result on the the closeness to the asymptotic normal distribution from the MLE approximated using finitely many observations.

## 2 Open questions and research directions

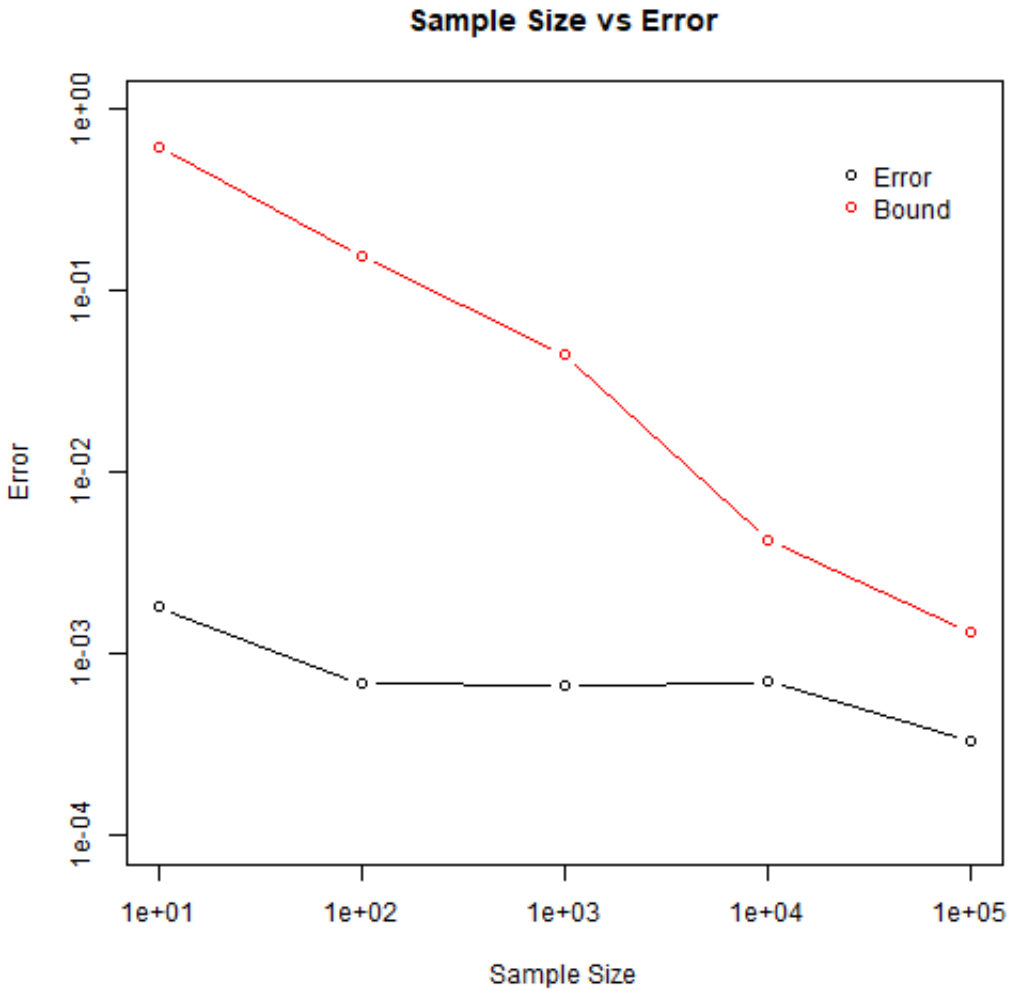


Figure 1: Simulation Gamma unknown scale

# References

[AL15] A. Anastasiou and C. Ley. "Bounds for the asymptotic normality of the maximum likelihood estimator using the Delta method". In: *arXiv preprint arXiv:1508.04948* (2015).