AIMS Mathematics, 7(10): 19083–19104.
DOI: 10.3934/math.20221048
Received: 13 June 2022
Revised: 27 July 2022
Accepted: 23 August 2022
Published: 29 August 2022

http://www.aimspress.com/journal/Math

Research article

Truncation point estimation of truncated normal samples and its applications

Shenglan Peng* **and Zikang Wan**

Computer Department, Jingdezhen Ceramic University, Jingdezhen 330403, China

* **Correspondence:** Email: solfix123@163.com.

Abstract: The moment estimates and maximum likelihood estimates of the truncation points in the truncated normal distribution are given, as well as the interval estimates for large samples. The estimation method of truncation point is applied to the assembly of DNA sequencing data, and moment estimation, maximum likelihood estimation and interval estimation of gap length are obtained. Monte Carlo simulations show that the experimental results are very close to the theoretical estimates. When the estimation method given in this paper is applied to a real DNA sequencing dataset, ideal estimation results are also obtained.

Keywords: truncated normal distribution; truncation point; moment estimation; maximum likelihood estimation; Lander-Waterman model; gap length
Mathematics Subject Classification: 62F10, 62P10

1. Introduction

Let $\varphi(z)$ and $\Phi(z)$ be the density and cumulative functions of a standard normal random variable, respectively.

$$\varphi(z) := \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}}, \quad \Phi(z) := \int_{-\infty}^{z} \varphi(t)dt.$$

Let X be a truncated normal random variable with truncation point θ on the left, where the mean and variance of the underlying normal distribution are denoted by μ and σ^2. The density function of X can be written as:

$$f(x\,|\,\mu,\sigma^2,\theta) = \begin{cases} \dfrac{\varphi[(x-\mu)/\sigma]}{\sigma[1-\Phi((\theta-\mu)/\sigma)]}, & x > \theta \\ 0, & \text{otherwise.} \end{cases} \tag{1.1}$$

Truncated normal distribution on the right can be similarly defined. Since the density function of the normal distribution is symmetrical about μ, truncation on the right at θ' is equivalent to truncation on

the left at $2\mu - \theta'$. Therefore, only truncation on the left is discussed in this paper.

It can be easily shown that

$$\mu_\theta := E(X \,|\, \mu, \sigma^2, \theta) = \sigma h(\theta^*) + \mu \tag{1.2}$$

$$\sigma_\theta^2 := V(X \,|\, \mu, \sigma^2, \theta) = \sigma^2[1 - h'(\theta^*)] \tag{1.3}$$

where $\theta^* = (\theta - \mu)/\sigma$, and $h(x) = \varphi(x)/[1 - \Phi(x)]$ is the hazard function of the standard normal distribution. The reciprocal of $h(x)$ is also known as Mills' ratio.

Other moments of singly truncated normal distribution such as skewness and kurtosis can be found in Horrace's work [1]. In addition, Pender [2] has presented the corresponding results for doubly truncated normal distribution.

A singly truncated normal sample from X with size n is denoted by $x_1, x_2, \cdots, x_n$. The likelihood function for the sample is:

$$L = \frac{1}{[1 - \Phi(\theta^*)]^n} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right]. \tag{1.4}$$

As for maximum likelihood estimation, it will be easier to work with the log-likelihood function:

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - n \ln [1 - \Phi(\theta^*)]\,. \tag{1.5}$$

For the inference of truncated normal distributions, most of the available literature tends to focus on the estimation of mean and standard deviation (or variance) of the underlying normal distribution when the truncation point θ is known. The initial results were given by Pearson and Lee [3], who employed the method of moments to obtain formulas for estimating the mean and standard deviation, and the computation of these estimates relied on certain special functions that can be evaluated by a pre-prepared table. Later, Fisher [4] gave a solution to this problem by the maximum likelihood method, and he proved that the maximum likelihood estimation is equivalent to the moment estimation. Stevens [5] discussed the estimation of doubly truncated normal samples, as well as censored samples in which the frequency of observations outside the limits is recorded but the individual values of these observations are not measured.

Based on the maximum likelihood principle proposed by Fisher, some other authors have obtained more results on the estimation problem of truncated or censored samples. Hald [6] and Gupta [7] studied the maximum likelihood estimation of the parameters of censored normal samples. Halperin [8] showed that maximum likelihood estimates for single-parameter truncated samples under mild conditions are consistent, asymptotically normally distributed, and of minimum variance for large samples. Halperin's results can be readily extended and applied to truncated normal samples. Cohen [9–11] developed simplified estimators for singly truncated normal samples as follows:

$$\hat{\mu} = \bar{x} - c(\bar{x} - \theta)$$

$$\hat{\sigma}^2 = s^2 + c(\bar{x} - \theta)^2$$

where $\bar{x}$ and s^2 are the sample mean and the sample variance, respectively, $c = h(\theta^*)/[h(\theta^*) - \theta^*]$ is an auxiliary function. The value of c can be determined directly from $s^2/(\bar{x} - \theta)^2$ by a table provided by Cohen.

On the other hand, there is very little literature on truncation point estimation. For general truncated samples, Robson and Whitlock [12] presented point estimators of a truncation point with bias of order n^{-k}. For example, estimator $x_{(1)} - (x_{(2)} - x_{(1)})$ is unbiased to order n^{-2}, where $x_{(1)}$ and $x_{(2)}$ are the order statistics of the truncated sample.

In this article, the truncated samples are different from those of Robson and Whitlock. Specifically, x_i $(i = 1, \cdots, n)$ in truncated samples cannot be measured directly, and $y_i = x_i - \theta$ can be observed instead. In such a case, all observations in the sample are subtracted by an unknown truncation point θ, and only observations greater than 0 are recorded. By (1.1), the density function of $Y = X - \theta$ is:

$$f(y) = \begin{cases} \dfrac{\varphi[(y+\theta-\mu)/\sigma]}{\sigma[1-\Phi((\theta-\mu)/\sigma)]}, & y > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1.6}$$

and the log-likelihood function for the truncated sample $y_1, y_2, \cdots, y_n$ is:

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i+\theta-\mu)^2 - n\ln\left[1-\Phi(\theta^*)\right]. \tag{1.7}$$

Next, we will discuss the estimation of the truncation point θ for sample $y_1, y_2, \cdots, y_n$. As mentioned earlier, the calculation of estimates of mean and variance for truncated normal samples requires some special functions, which can be evaluated by a pre-prepared table. The estimation of the truncation point also requires some auxiliary functions, and the hazard function $h(x)$ of the standard normal distribution plays an important role in these functions. The following lemma gives some important properties of $h(x)$, which we will use later.

Lemma 1.1. *Let $h(x)$ be the hazard function of the standard normal distribution, then:*

$$(\sqrt{8+x^2}+3x)/4 < h(x) < (\sqrt{4+x^2}+x)/2, \ x > 0 \tag{1.8}$$

$$0 < h'(x) = h(h-x) < 1, \ -\infty < x < \infty \tag{1.9}$$

$$h''(x) = h[(h-x)(2h-x)-1] > 0, \ -\infty < x < \infty. \tag{1.10}$$

The detailed proof of this lemma can be found in Sampford's work [14]. It should be noted that the inequality on the right in (1.8) is attributed to Birnbaum's work [13]. Moreover, Yang and Chu [15] provided tighter bounds for $h(x)$ than those in (1.8).

The estimation for the truncation point θ makes no sense when the mean μ is unknown. In Section 3, we will present an application where there is a normal sample from which good estimates of the mean and variance can be made. In addition, there are many truncated samples derived from the same normal population, and the truncation points of interest may be different.

2. Materials and methods

In this section, we will discuss truncation point estimation for two cases: The variance σ^2 is known or unknown. For each of the two cases, the point estimates of the truncation points are given using the method of moment and the maximum likelihood method, respectively, and the estimated variance of the point estimates and the confidence intervals of the truncation points are also given.

2.1. σ^2 is known

Theorem 2.1. *Let $y_1, y_2, \cdots, y_n$ be a sample from a truncated normal population with density function (1.6), where the mean μ and the variance σ^2 are known, and the truncation point θ is unknown, the moment estimator for θ is given as follows:*

$$\hat{\theta} = \mu + \sigma g_1^{-1}\left(\frac{\bar{y}}{\sigma}\right) \tag{2.1}$$

where auxiliary function $g_1(x) := h(x) - x$ is monotonically decreasing.

Proof. Since $E(X) = E(Y) + \theta$, replace the left side of Eq (1.2) with $\bar{y} + \theta$, and we have:

$$\bar{y} + \hat{\theta} = \sigma h(\hat{\theta}^*) + \mu$$

where $\hat{\theta}^* = (\hat{\theta} - \mu)/\sigma$. It is algebraically equivalent to:

$$\begin{aligned}\frac{\bar{y}}{\sigma} &= h(\hat{\theta}^*) - \frac{\hat{\theta} - \mu}{\sigma} \\ &= h(\hat{\theta}^*) - \hat{\theta}^* = g_1(\hat{\theta}^*).\end{aligned}$$

Note that (1.9) implies $g_1'(x) = h'(x) - 1 < 0$, that is, $g_1(x)$ is monotonically decreasing. Applying the inverse function of g_1 to the both sides yields:

$$g_1^{-1}\left(\frac{\bar{y}}{\sigma}\right) = \hat{\theta}^* = \frac{\hat{\theta} - \mu}{\sigma}.$$

The estimator for truncation point in (2.1) is immediately obtained from above. □

Theorem 2.2. *For large samples, the variance of the truncation point estimator in (2.1) is approximated by:*

$$V(\hat{\theta}) \approx \frac{\sigma^2/n}{1 - h'(\hat{\theta}^*)}. \tag{2.2}$$

Proof. With first-order Taylor series expansion of (2.1) at $\mu_\theta - \theta$, we obtain:

$$\hat{\theta} \approx \theta + (g_1^{-1})'\Big|_{y=(\mu_\theta - \theta)/\sigma} \cdot (\bar{y} - \mu_\theta + \theta).$$

For large samples, $\bar{y} \approx \mu_\theta - \theta$, hence:

$$V(\hat{\theta}) \approx \left[(g_1^{-1})'\right]^2_{y=\bar{y}/\sigma} V(\bar{y}). \tag{2.3}$$

Recall that the derivative of the inverse function is the reciprocal of the derivative of the original function, namely, $(g_1^{-1})' = (g_1')^{-1}$, it follows that:

$$\left[(g_1^{-1})'\right]^2_{y=\bar{y}/\sigma} = \frac{1}{[g_1'(\hat{\theta}^*)]^2} = \frac{1}{[h'(\hat{\theta}^*) - 1]^2}. \tag{2.4}$$

Since $\bar{y} = \bar{x} - \theta$, we have:

$$V(\bar{y}) = V(\bar{x}) = \frac{\sigma_\theta^2}{n} \approx \frac{1 - h'(\hat{\theta}^*)}{n}\sigma^2. \tag{2.5}$$

The approximate variance (2.2) follows from (2.3)–(2.5). □

Theorem 2.3. *Let $y_1, y_2, \cdots, y_n$ be a sample from a truncated normal population with density function (1.6), where the mean μ and the variance σ^2 are known, and the truncation point θ is unknown, the maximum likelihood estimator for θ is given as follows:*

$$\hat{\theta} = \mu + \sigma g_1^{-1}\left(\frac{\bar{y}}{\sigma}\right).$$

That is, the maximum likelihood estimator of θ is identical to the moment estimator of θ.

Proof. Differentiating the log-likelihood function $\ln L$ in (1.7) with respect to θ:

$$\begin{aligned}\frac{d}{d\theta}\ln L &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i+\theta-\mu) + n\frac{\varphi(\theta^*)}{1-\Phi(\theta^*)}\frac{d\theta^*}{d\theta} \\ &= -\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i - n(\theta-\mu) + \frac{n}{\sigma}h(\theta^*) \\ &= \frac{n}{\sigma}\left[-\frac{\bar{y}}{\sigma} - \theta^* + h(\theta^*)\right].\end{aligned} \tag{2.6}$$

Setting this derivative equal to zero, we obtain the same estimator as the moment estimator in (2.1). □

It can be shown that the log-likelihood function in (1.7) satisfies the mild regularity conditions given by Halperin [8]. Thus, the maximum likelihood estimator of the truncation point θ is consistent, asymptotically normal, and approximately minimum variance under large samples.

We will employ another approach based on large-sample theory to find the approximate variance of the maximum likelihood estimator of the truncation point θ. By large-sample theory, we have:

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right)$$

where $I(\theta)$ is the Fisher information.

For large samples,

$$I(\theta) = E\left[\frac{d^2}{d\theta^2}\ln f(Y;\theta)\right] \approx -\frac{1}{n}\frac{d^2}{d\theta^2}(\ln L).$$

Therefore, the variance of $\hat{\theta}$ can be approximated as follows:

$$V(\hat{\theta}) \approx \left[-\frac{d^2}{d\theta^2}(\ln L)\right]^{-1}\Bigg|_{\theta=\hat{\theta}}. \tag{2.7}$$

By (2.6), we have:

$$\begin{aligned}\frac{d^2}{d\theta^2}(\ln L) &= \frac{d}{d\theta}\left[\frac{d}{d\theta}(\ln L)\right] \\ &= \frac{n}{\sigma^2}\left[h^2(\theta^*) - \theta^* h(\theta^*) - 1\right] \\ &= \frac{n}{\sigma^2}\left[h'(\theta^*) - 1\right].\end{aligned}$$

Hence, substituting into (2.7), we obtain the same approximation as (2.2):

$$V(\hat{\theta}) \approx \frac{\sigma^2/n}{1-h'(\hat{\theta}^*)}.$$

According to the asymptotic normality for large samples, the approximate 95% confidence interval for the truncation point θ is:

$$\mu + \sigma g_1^{-1}\left(\frac{\bar{y}}{\sigma}\right) \pm 1.96 \times \frac{\sigma/\sqrt{n}}{\sqrt{1-h'(\hat{\theta}^*)}}. \tag{2.8}$$

2.2. σ^2 is unknown

Theorem 2.4. *Let $y_1, y_2, \cdots, y_n$ be a sample from a truncated normal population with density function (1.6), where the mean μ is known, the variance σ^2 and the truncation point θ are unknown, the moment estimator for θ is given as follows:*

$$\hat{\theta}^* = g_2^{-1}\left(s_y^2/\bar{y}^2\right) \tag{2.9}$$

$$\hat{\sigma} = \bar{y}/g_1(\hat{\theta}^*) \tag{2.10}$$

$$\hat{\theta} = \mu + \hat{\sigma}\hat{\theta}^* \tag{2.11}$$

where s_y^2 is the sample variance of the truncated sample, and auxiliary function

$$g_2(x) := \frac{1-h'(x)}{[h(x)-x]^2}, \quad x > 0 \tag{2.12}$$

is monotonically increasing.

First, we prove the following lemma:

Lemma 2.5. *Let $g_2(x)$ be the auxiliary function defined in (2.12), then for any $x > 0$, we have $0 < g_2(x) < 1$ and $g_2'(x) > 0$, i.e., $g_2(x)$ is non-negative and monotonically increasing.*

Proof. From (1.9), $g_2(x) > 0$, and we have

$$\begin{aligned} g_2(x) &= \frac{1-h'(x)}{[h(x)-x]^2} \\ &= \frac{1-h(x)[h(x)-x]-[h(x)-x]^2}{[h(x)-x]^2} + 1 \\ &= \frac{1-[h(x)-x][2h(x)-x]}{[h(x)-x]^2} + 1. \end{aligned}$$

By (1.10), we know that $1-[h(x)-x][2h(x)-x] < 0$, hence $g_2(x) < 1$. Decompose $g_2(x)$ as follows

$$g_2(x) = \frac{1}{[h(x)-x]^2} - \frac{h(x)}{h(x)-x}.$$

Differentiating $g_2(x)$, we obtain:

$$g_2'(x) = \frac{2[1-h'(x)]-[h(x)-xh'(x)][h(x)-x]}{[h(x)-x]^3}.$$

Let ξ denote the numerator of the above fraction, namely

$$\xi(x) = 2[1 - h'(x)] - [h(x) - xh'(x)][h(x) - x].$$

Therefore, for $x > 0$, in order to prove $g_2'(x) > 0$, it is sufficient to prove $\xi(x) > 0$.

Notice that the value of $\xi(x)$ at $x = 0$ is greater than 0:

$$\xi(0) = 2[1 - h^2(0)] - h^2(0) = 2 - \frac{6}{\pi} > 0.$$

As $x \to \infty$, it follows from (1.8) that

$$\begin{aligned} 1 > h'(x) =& h(x)[h(x) - x] \\ &> \frac{\sqrt{8 + x^2} + 3x}{4} \cdot \frac{\sqrt{8 + x^2} - x}{4} \\ &= \frac{\sqrt{8 + x^2} + 3x}{2(\sqrt{8 + x^2} + x)} \to 1. \end{aligned}$$

Thus, we have $1 - h'(x) \to 0$. Similarly, inequalities

$$0 < \frac{2}{\sqrt{8 + x^2} + x} < h(x) - x < \frac{2}{\sqrt{4 + x^2} + x} \to 0$$

imply that $h(x) - x \to 0$, and $x[h(x) - x] \to 1$. Hence, we have:

$$[h(x) - xh'(x)][h(x) - x] = [h(x) - x]^2 - [1 - h'(x)] \cdot x[h(x) - x] \to 0.$$

Therefore, we obtain that $\xi \to 0$.

We employ proof by contradiction to show $\xi(x) > 0$. If there exists some $x_1 > 0$ such that $\xi(x_1) \le 0$, then there exists $x_2 > 0$ such that

$$\begin{cases} \xi(x_2) \le 0 \\ \xi'(x_2) = 0. \end{cases}$$

In addition, we have

$$\begin{aligned} \xi'(x) &= -2h'' + xh''(h - x) - (h - xh')(h' - 1) \\ &= -2h[(h - x)^2 + h(h - x) - 1] + xh''(h - x) - h(h - x)(h - xh') + h - xh' \\ &= h[2 - 2h(h - x) - (h - xh')(h - x)] - 2h(h - x)^2 + xh''(h - x) + h - xh' \\ &= h\xi + xh''(h - x) - 2h(h - x)^2 + h - xh(h - x) \\ &= h\xi + xh''(h - x) - h'' \\ &= h\xi + h''[x(h - x) - 1]. \end{aligned}$$

Since $h''(x) > 0$ and $x[h(x) - x] < 1$, it follows that

$$\xi'(x) < h(x)\xi(x).$$

Hence, we obtain

$$\xi'(x_2) < h(x_2)\xi(x_2) \le 0.$$

This contradicts $\xi'(x_2) = 0$. We complete the proof the lemma. □

Proof of Theorem 2.4. Substituting $\bar{y}$ for μ_θ into (1.2) and rearranging the equation, we have

$$\bar{y} = \sigma\left[h(\hat{\theta}^*) - \hat{\theta}^*\right] = \sigma g_1(\hat{\theta}^*). \tag{2.13}$$

Similarly, substituting s_y^2 for σ_θ^2 into (1.3), we obtain

$$s_y^2 = \sigma^2[1 - h'(\hat{\theta}^*)]. \tag{2.14}$$

We eliminate σ between (2.13) and (2.14), hence we obtain:

$$\frac{s_y^2}{\bar{y}^2} = \frac{1 - h'(\hat{\theta}^*)}{[h(\hat{\theta}^*) - \hat{\theta}^*]^2} = g_2(\hat{\theta}^*).$$

From Lemma 2.5, $g_2(x)$ is invertible when $x > 0$, so we have:

$$\hat{\theta}^* = g_2^{-1}\left(\frac{s_y^2}{\bar{y}^2}\right) = g_2^{-1}(CV_y^2)$$

where $CV_y := s_y/\bar{y}$ is the coefficient of variation of the truncated sample.

The estimator of the standard deviation σ can be obtained directly from (2.13):

$$\hat{\sigma} = \frac{\bar{y}}{g_1(\hat{\theta}^*)}.$$

Since $\theta = \mu + \sigma\theta^*$, the estimator for truncation point is:

$$\hat{\theta} = \mu + \hat{\sigma}\hat{\theta}^*.$$

□

Theorem 2.6. *Let $y_1, y_2, \cdots, y_n$ be a sample from a truncated normal population with density function (1.6), where the mean μ is known, the variance σ^2 and the truncation point θ are unknown, the maximum likelihood estimators for θ and $\hat{\sigma}$ are given as follows:*

$$\hat{\theta} = \mu + \hat{\sigma}\hat{\theta}^* \tag{2.15}$$

$$\hat{\sigma} = \frac{\bar{y}}{g_1(\hat{\theta}^*)} \tag{2.16}$$

where

$$\hat{\theta}^* = g_2^{-1}\left(\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n\bar{y}^2}\right). \tag{2.17}$$

Proof. Taking partial derivative of $\ln L$ in (1.7) with respect to θ, we obtain:

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\sigma}\left[-\frac{\bar{y}}{\sigma} - \theta^* + h(\theta^*)\right].$$

Equating the partial derivative to zero, hence:

$$\frac{\bar{y}}{\hat{\sigma}} = h(\hat{\theta}^*) - \hat{\theta}^* = g_1(\hat{\theta}^*). \tag{2.18}$$

The partial derivative of $\ln L$ with respect to σ is:

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i + \theta - \mu)^2 + n\frac{\varphi(\theta^*)}{1-\Phi(\theta^*)}\frac{\partial \theta^*}{\partial \sigma} \\ &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}[(y_i - \bar{y}) + \sigma(\frac{\theta-\mu}{\sigma} + \frac{\bar{y}}{\sigma})]^2 - n\frac{\varphi(\theta^*)}{1-\Phi(\theta^*)}\frac{\theta-\mu}{\sigma}\frac{1}{\sigma} \\ &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}[(y_i - \bar{y}) + \sigma(\theta^* + \frac{\bar{y}}{\sigma})]^2 - \frac{n}{\sigma}\theta^* h(\theta^*).\end{aligned}$$

From (2.18), we know that $\hat{\theta}^* + \bar{y}/\hat{\sigma} = h(\hat{\theta}^*)$, it therefore follows:

$$\begin{aligned}\left.\frac{\partial \ln L}{\partial \sigma}\right|_{(\hat{\theta},\hat{\sigma})} &= -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}[(y_i - \bar{y}) + \hat{\sigma}h(\hat{\theta}^*)]^2 - \frac{n}{\hat{\sigma}}\hat{\theta}^* h(\hat{\theta}^*) \\ &= -\frac{n}{\hat{\sigma}}\left[1 - \frac{1}{n\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 - h(\hat{\theta}^*)[h(\hat{\theta}^*) - \hat{\theta}^*]\right] \\ &= -\frac{n}{\hat{\sigma}}\left[1 - \frac{1}{n\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 - h'(\hat{\theta}^*)\right].\end{aligned}$$

Setting the partial derivative with respect to σ^2 equal to zero, we obtain:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \hat{\sigma}^2[1 - h'(\hat{\theta}^*)]. \tag{2.19}$$

By eliminating $\hat{\sigma}$ between (2.18) and (2.19), we have:

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n\bar{y^2}} = \frac{1 - h'(\hat{\theta}^*)}{h(\hat{\theta}^*) - \hat{\theta}^*} = g_2(\hat{\theta}^*).$$

Since g_2 is monotonically increasing at $x > 0$ from Lemma 2.5, hence:

$$\hat{\theta}^* = g_2^{-1}\left(\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n\bar{y}^2}\right).$$

Now, the estimator of σ can be obtained by (2.18):

$$\hat{\sigma} = \frac{\bar{y}}{h(\hat{\theta}^*) - \hat{\theta}^*} = \frac{\bar{y}}{g_1(\hat{\theta}^*)}.$$

Since $\theta^* = (\theta - \mu)/\sigma$, it follows that $\hat{\theta} = \mu + \hat{\sigma}\hat{\theta}^*$. □

According to large-sample theory, the maximum likelihood estimators $(\hat{\theta}, \hat{\sigma}^2)$ is asymptotically normal and the estimated approximate variance-covariance matrix is $\mathbf{V} = (-\mathbf{H})^{-1}|_{(\hat{\theta},\hat{\sigma})}$, where $\mathbf{H}$ is the Hessian matrix of the log-likelihood function $\ln L$ in (1.7). Then

$$V(\hat{\theta}) \approx v_{11}, \quad V(\hat{\sigma}) \approx v_{22}, \quad Cov(\hat{\theta}, \hat{\sigma}) \approx v_{12}$$

where v_{11}, v_{12}, and v_{22} are the elements of $\mathbf{V}$.

Similar to the results given by Cohen [9, 11], the Hessian matrix $\mathbf{H}$ is obtained by:

$$h_{11} := \frac{n}{\hat{\sigma}^2}[h'(\hat{\theta}^*) - 1] \tag{2.20}$$

$$h_{12} := \frac{n}{\hat{\sigma}^2}h(\hat{\theta}^*)[1 - \hat{\theta}^*(h(\hat{\theta}^*) - \hat{\theta}^*)] \tag{2.21}$$

$$h_{22} := -\frac{2n}{\hat{\sigma}^2} - \hat{\theta}^* h_{12} \tag{2.22}$$

and the elements of the approximate variance-covariance matrix $\mathbf{V}$ are functions of h_{11}, h_{12}, and h_{22} as follows:

$$v_{11} = \frac{h_{22}}{h_{12}^2 - h_{11}h_{22}}, \quad v_{22} = \frac{h_{11}}{h_{12}^2 - h_{11}h_{22}}, \quad v_{12} = \frac{-h_{12}}{h_{12}^2 - h_{11}h_{22}}. \tag{2.23}$$

By asymptotic normality, An approximate 95 percent confidence interval of the trunction point θ is:

$$\hat{\theta} \pm 1.96 \times \sqrt{v_{11}}. \tag{2.24}$$

3. Application: Estimation of gap length in DNA assembly

The fundamental principle of DNA sequencing is that the target DNA is randomly sheared into small fragments and sequenced separately; then, utilizing overlapping information between adjacent fragments identified (called reads), assembly packages pieced them together to restore the original sequence. Due to low coverage, sequencing errors, or repetitive sequences, some regions of the target DNA were not completed, and we call them gaps.

To address the difficulties caused by the repetitive sequences, a technique known as paired-end sequencing was developed. By paired-end sequencing, reads in pairs are produced from both ends of each DNA fragment and they are called mates each other. Mates are in opposite directions, at approximately known distances. Randomness in mates' distance results from random fragmentation of target DNA. However, the size of these fragments can be controlled within a certain range by biological technology. That is, the length of the DNA fragments in sequencing follows a certain probability distribution, which is usually assumed to be a normal distribution known.

The distance between mate pair provides information for the estimation of gap length. If a DNA fragment spans the entire gap and its indentation lengths L and R at the left and right ends (denoted by L and R, respectively) are known (as shown in Figure 1). Let the fragment's length be $X \sim N(\mu, \sigma^2)$, where μ and σ^2 are known, thus the gap length $\theta = X - (L + R)$. Let $Y = L + R$, then $Y = X - \theta$ follows a truncated normal distribution, where θ is the truncation point.

Unlike the truncated normal distribution discussed earlier, the truncation variable Y is not only related to the truncation point but also affected by the start or end position of the fragment on the target DNA. Specifically, whether a fragment spans the gap depends on its length, as well as its starting position on the target DNA. In order to obtain more practical truncated distributions, we employ the Lander-Waterman model, a mathematical model for shotgun sequencing, and all estimates of gap length are based on this model.

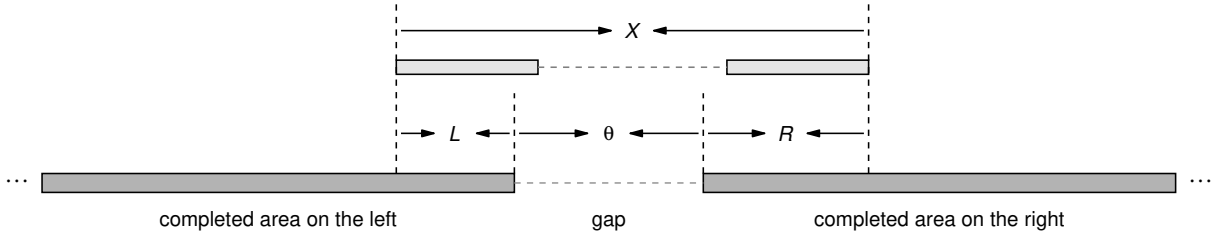

Figure 1. The information about the gap provided by the mate pair: the length of the gap θ is unknown, the DNA fragment where the mate pair is located spans the gap, its length X is unknown but the distribution is known, and the indent lengths L and R of mate pair at the left and right ends are able to be observed.

3.1. Lander-Waterman model

Lander and Waterman [16] developed a mathematical model for shotgun sequencing. Some key statistics of sequencing projects, such as coverage, the average number of gaps, etc., can be estimated by this model. Therefore, the Lander-Waterman model was often used to guide the plan of sequencing projects.

Let G be the length of the target DNA sequence and $\mathcal{P} = \{f_1, f_2, \cdots, f_N\}$ be a sequencing project for the target DNA, where $f_i = (S_i, X_i)$ are i.i.d. random vectors. $S_i \sim U(0, L)$ is the start position of the ith fragment located on the target DNA, $X_i \sim N(\mu, \sigma^2)$ is the length of f_i, S_i and X_i are mutually independent.

Roach et al. [17] introduced paired-end sequencing information into the Lander-Waterman model. A pair of reads named mate-pair is obtained by sequencing from both ends of each fragment in paired-end sequencing projects. If one read in mate-pair overlaps with a finished region (called a contig), and the other one overlaps with another contig, the mate-pair provides the information about the order of the contigs, as well as the information about the gap between the two adjacent contigs.

For a given gap in a sequencing project, it is assumed to be located in the region of the target sequence starting from a and of length θ. Therefore, the event that a fragment $f = (S, X)$ spans the entire gap can be expressed as $\{S < a;\ S + X > a + \theta\}$. All we can observe is that the mate-pair has an indent $L = a - S$ on the left neighboring contig and an indent $R = (S + X) - (a + \theta)$ on the right neighboring contig, respectively. Indeed, $Y = L + R = X - \theta$ is a truncated variable. However, the conditional event for Y is $\{S < a;\ S + X > a + \theta\}$ instead of $\{X > \theta\}$.

Let $\mathcal{G} = \{f = (S, X) \in \mathcal{P} \,|\, S < a;\ S + X > a + \theta\}$ be the set of all the fragments that can span the entire gap. Let L_i and R_i $(i = 1, \cdots, n)$ be the left indent and the right indent of the mate-pair of the ith fragment in $\mathcal{G}$. We define truncated sample as follows:

$$\mathcal{Y} := \{y_i = L_i + R_i,\ i = 1, 2, \cdots, n\} \tag{3.1}$$

Our inference about gap length θ is based on the truncated sample $\mathcal{Y}$. The conditional event is important for inference. Let A denote the conditional event $\{S < a;\ S + X > a + \theta\}$, The following lemma gives an approximation to the probability $P(A)$:

Lemma 3.1.

$$P(A) \approx \frac{\sigma}{G}[h(\theta^*) - \theta^*][1 - \Phi(\theta^*)] \tag{3.2}$$

where $\theta^* = (\theta - \mu)/\sigma$ is defined as before.

Proof. From the definition for event A, we know that:

$$\begin{aligned} P(A) &= P(0 < S < a,\ a + \theta < S + X < G) \\ &= P(0 < S < a; a + \theta - S < X < G - S) \\ &= P(\theta < X < G; a + \theta - X < S < (G - X) \wedge a). \end{aligned}$$

Let $f_X(x)$ be the density function of X, hence we have:

$$\begin{aligned} GP(A) &= G \int_{\theta}^{G} f_X(x)dx \int_{a+\theta-x}^{(G-x)\wedge a} \frac{1}{G} ds \\ &= \int_{\theta}^{G-a} f_X(x)dx \int_{a+\theta-x}^{a} ds + \int_{G-a}^{G} f_X(x)dx \int_{a+\theta-x}^{G-x} ds \\ &= \int_{\theta}^{G-a} (x - \theta) f_X(x)dx + (G - a - \theta) \int_{G-a}^{G} f_X(x)dx. \end{aligned}$$

The integral in the second term is:

$$\begin{aligned} (G - a - \theta) \int_{G-a}^{G} f_X(x)dx &= P(G - a < X < G) \\ &= (G - a - \theta)\left[\Phi\left(\frac{G - \mu}{\sigma}\right) - \Phi\left(\frac{G - a - \mu}{\sigma}\right)\right] \\ &\le (G - a - \theta)\left[1 - \Phi\left(\frac{G - a - \mu}{\sigma}\right)\right] \\ &\le \sigma \frac{G - a - \theta}{G - a - \mu} \varphi\left(\frac{G - a - \mu}{\sigma}\right). \end{aligned}$$

The last equality above is from $1 - \Phi(x) < x^{-1}\varphi(x)$ for $x > 0$. Except for the gaps at the edge of the genome, $G - a \gg \max(\mu, \sigma, \theta)$. For example, the genome size of *E. coli* is 4.4M, and the mean of insert size is generally less than 1K. Since $\varphi(x) \to 0$ as $x \to +\infty$, The second term is approximated as 0.

The first term is:

$$\begin{aligned} \int_{\theta}^{G-a} (x - \theta) f_X(x)dx &= \sigma \int_{\theta}^{G-a} \left(\frac{x - \mu}{\sigma} - \frac{\theta - \mu}{\sigma}\right) f_X(x)dx \\ &= \sigma \int_{(\theta-\mu)/\sigma}^{(G-a-\mu)/\sigma} \left(t - \frac{\theta - \mu}{\sigma}\right) \varphi(t)dt \\ &= \sigma \left[-\varphi(t) - \frac{\theta - \mu}{\sigma}\Phi(t)\right]_{(\theta-\mu)/\sigma}^{(G-a-\mu)/\sigma} \\ &\approx \sigma \left[-\varphi(t) - \frac{\theta - \mu}{\sigma}\Phi(t)\right]_{(\theta-\mu)/\sigma}^{+\infty} \\ &= \sigma \left[\varphi\left(\frac{\theta - \mu}{\sigma}\right) + \frac{\theta - \mu}{\sigma}\Phi\left(\frac{\theta - \mu}{\sigma}\right) - \frac{\theta - \mu}{\sigma}\right] \\ &= \sigma[\varphi(\theta^*) + \theta^*\Phi(\theta^*) - \theta^*]. \end{aligned}$$

Combining the two terms, we finally obtain:

$$P(A) \approx \frac{\sigma}{G}\left[\varphi(\theta^*) - \theta^*(1 - \Phi(\theta^*))\right] = \frac{\sigma[h(\theta^*) - \theta^*][1 - \Phi(\theta^*)]}{G}.$$

□

3.2. Moment estimators

Conditioning on that event A occurs, $Y = X - \theta$, we then have:

$$E(Y) = E(X - \theta|A) = E(X|A) - \theta. \tag{3.3}$$

In a similar way to the proof of Lemma 3.1, we obtain the conditional expectation $E(X|A)$, denoted by μ_θ. We substitute $\bar{y}$ for $E(Y)$ into (3.3), and we obtain a equation for the unknown parameter θ:

$$\bar{y} = \mu_\theta - \theta. \tag{3.4}$$

The solution of Eq (3.4) is the moment estimate of the truncation point θ.

Theorem 3.2. *Let $y_1, y_2, \cdots, y_n$ be a truncated sample as (3.1), the moment estimator for the truncation point θ is given as follows:*

$$\hat{\theta} = \mu + \sigma g_3^{-1}\left(\frac{\bar{y}}{\sigma}\right) \tag{3.5}$$

where auxiliary function

$$g_3(x) := \frac{1}{h(x) - x} - x \tag{3.6}$$

is monotonically decreasing.

First, we prove by a lemma that the auxiliary function g_3 is monotonically increasing.

Lemma 3.3. *The auxiliary function $g_3(x)$ defined in (3.6) is monotonically increasing.*

Proof. The derivative of $g_3(x)$ is:

$$\begin{aligned} g_3'(x) &= -\frac{h'(x) - 1}{[h(x) - x]^2} - 1 \\ &= \frac{1 - h(x)[h(x) - x] - [h(x) - x]^2}{[h(x) - x]^2} \\ &= \frac{1 - [h(x) - x][2h(x) - x]}{[h(x) - x]^2}. \end{aligned}$$

By (1.10), we know that $[h(x) - x][2h(x) - x] - 1 > 0$, it follows that $g_3'(x) < 0$. Therefore, $g_3(x)$ is monotonically increasing. □

Proof of Theorem 3.2. First, we calculate $E[(X-\mu)1_A]$ in a similar way in Lemma 3.1:

$$\begin{aligned} G \cdot E[(X-\mu)1_A] &= \int_{\theta}^{G-a} (x-\mu)(x-\theta) f_X(x)dx + (G-a-\theta)\int_{G-a}^{G} (x-\mu) f_X(x)dx \\ &\approx \sigma^2 \int_{\theta}^{G-a} \left(\frac{x-\mu}{\sigma}\right)\left(\frac{x-\mu}{\sigma} - \frac{\theta-\mu}{\sigma}\right) f_X(x)dx \\ &\approx \sigma^2 \int_{\theta^*}^{+\infty} \left(t^2 - \theta^* t\right)\varphi(t)dt \\ &= \sigma^2 \left[\Phi(t) - t\varphi(t) + \theta^*\varphi(t)\right]_{\theta^*}^{+\infty} \\ &= \sigma^2 \left[1 - \Phi(\theta^*)\right]. \end{aligned}$$

Next, we calculate $E(X|A)$ as follows:

$$\begin{aligned} \mu_\theta = E(X|A) &= \frac{1}{P(A)} E(X1_A) \\ &= \frac{G \cdot E[(X-\mu)1_A]}{G \cdot P(A)} + \mu \\ &\approx \frac{\sigma[1-\Phi(\theta^*)]}{\varphi(\theta^*) - \theta^*[1-\Phi(\theta^*)]} + \mu \\ &= \frac{\sigma}{\dfrac{\varphi(\theta^*)}{1-\Phi(\theta^*)} - \theta^*} + \mu. \end{aligned}$$

Hence, we obtain:

$$\mu_\theta = \frac{\sigma}{h(\theta^*) - \theta^*} + \mu. \tag{3.7}$$

Substituting $\sigma/(h(\theta^*) - \theta^*) + \mu$ for μ_θ in (3.4) and rearrange it, we obtain:

$$g_3(\theta^*) = \frac{1}{h(\theta^*) - \theta^*} - \theta^* = \frac{\bar{y}}{\sigma}. \tag{3.8}$$

We know that g_3 is monotonically decreasing from Lemma 3.3, hence:

$$\hat{\theta}^* = g_3^{-1}\left(\frac{\bar{y}}{\sigma}\right). \tag{3.9}$$

Since $\theta^* = (\theta - \mu)/\sigma$, we thus have:

$$\hat{\theta} = \mu + \sigma g_3^{-1}\left(\frac{\bar{y}}{\sigma}\right).$$

□

Lemma 3.4.

$$V(Y) = V(X|A) \approx \frac{\sigma^2 h''(\theta^*)}{h(\theta^*)[h(\theta^*) - \theta^*]^2}. \tag{3.10}$$

Proof. In the same way as for Theorem 3.2, we obtain:

$$\begin{aligned} G \cdot E[(X-\mu)^2 1_A] &= \int_{\theta}^{\infty} (x-\mu)^2 (x-\theta) f_X(x) dx \\ &\approx \sigma^3 \int_{\theta^*}^{+\infty} \left(t^3 - \theta^* t^2\right) \varphi(t) dt \\ &= \sigma^3 \left[-(t^2+2)\varphi(t) + \theta^* t \varphi(t) + \theta^* (1-\Phi(t))\right]_{\theta^*}^{+\infty} \\ &= \sigma^3 \left[2\varphi(\theta^*) - \theta^* (1-\Phi(\theta^*))\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} E(X^2|A) &= \frac{1}{P(A)} E(X^2 1_A) \\ &= \frac{1}{P(A)} \left\{E[(X-\mu)^2 1_A] + 2\mu E[(X-\mu)1_A] + \mu^2 P(A)\right\} \\ &= \frac{G \cdot E[(X-\mu)^2 1_A]}{G \cdot P(A)} + \frac{G \cdot E[(X-\mu)1_A]}{G \cdot P(A)} 2\mu + \mu^2 \\ &\approx \sigma^2 \frac{2h(\theta^*) - \theta^*}{h(\theta^*) - \theta^*} + \frac{2\sigma\mu}{h(\theta^*) - \theta^*} + \mu^2. \end{aligned}$$

Consequently, we have:

$$\begin{aligned} V(X|A) &= E(X^2|A) - [E(X|A)]^2 \\ &\approx \sigma^2 \frac{2h(\theta^*) - \theta^*}{h(\theta^*) - \theta^*} - \frac{\sigma^2}{[h(\theta^*) - \theta^*]^2} \\ &= \sigma^2 \frac{[2h(\theta^*) - \theta^*][h(\theta^*) - \theta^*] - 1}{[h(\theta^*) - \theta^*]^2} \\ &= \frac{\sigma^2 h''(\theta^*)}{h(\theta^*)[h(\theta^*) - \theta^*]^2}. \end{aligned}$$

The last equation is from (1.10) in Lemma 1.1. □

Theorem 3.5. *For large samples, the variance of $\hat{\theta}$ in (3.5) is approximated by:*

$$V(\hat{\theta}) \approx \frac{h(\theta^*)[h(\theta^*) - \theta^*]^2 \sigma^2}{n h''(\theta^*)}. \tag{3.11}$$

Proof. In a similar way as for Theorem 2.2, we have:

$$\begin{aligned} V(\hat{\theta}) &\approx \left[(g_3^{-1})'\right]^2_{y=\bar{y}/\sigma} \cdot V(\bar{y}) \\ &= \frac{1}{[g_3'(\hat{\theta}^*)]^2} \cdot \frac{1}{n} V(Y) \\ &= \frac{h(\theta^*)[h(\theta^*) - \theta^*]^2}{h''(\theta^*)} \cdot \frac{\sigma^2}{n}. \end{aligned}$$

Here we use the results for g_3' in Lemma 3.3. □

3.3. Maximum likelihood estimators

The conditional density of $Y = X - \theta$ under the condition that the event A occurs:

$$\begin{aligned} f_Y(y|\theta) &= \frac{1}{P(A)} \int_{a-y}^{a} f_{S,X}(s, y+\theta) ds \\ &= \frac{1}{P(A)} \int_{a-y}^{a} \frac{1}{G} f_X(y+\theta) ds \\ &= \frac{1}{P(A)} \frac{y}{G} f_X(y+\theta) \\ &= \frac{y/\sigma}{\varphi(\theta^*) - \theta^*[1-\Phi(\theta^*)]} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y+\theta-\mu)^2} \end{aligned}$$

where $f_{S,X}$ the joint probability density of S and X.

Therefore, the log-likelihood function for the truncated sample in (3.1) is:

$$\begin{aligned} \ln L(\theta) = &\sum_{i=1}^{n} \ln Y_i - \frac{n}{2} \ln(2\pi) - n \ln \sigma^2 \\ &- n \ln\left(\varphi(\theta^*) - \theta^*[1-\Phi(\theta^*)]\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i + \theta - \mu)^2. \end{aligned}$$

Differentiating $\ln L(\theta)$ with respect to θ, we have:

$$\begin{aligned} \frac{d \ln L}{d\theta} &= \frac{n}{\sigma} \cdot \frac{1-\Phi(\theta^*)}{\varphi(\theta^*) - \theta^*[1-\Phi(\theta^*)]} - \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i + \theta - \mu) \\ &= \frac{n}{\sigma} \cdot \frac{1}{h(\theta^*) - \theta^*} - \frac{n}{\sigma^2} (\bar{y} + \theta - \mu) \\ &= \frac{n}{\sigma} \left(\frac{1}{h(\theta^*) - \theta^*} - \frac{\bar{y}}{\sigma} - \theta^* \right). \end{aligned}$$

Setting this derivative equal to zero, we obtain the same equation as (3.8):

$$g_3(\theta^*) = \frac{1}{h(\theta^*) - \theta^*} - \theta^* = \frac{\bar{y}}{\sigma}.$$

Hence, we obtain the same estimator as (3.5).

3.4. Interval estimation for large samples

The second-order derivative of $\ln L(\theta)$ is:

$$\begin{aligned} \frac{d^2 \ln L}{d\theta^2} &= \frac{n}{\sigma^2} \cdot \frac{-h'(\theta^*) + 1}{[h(\theta^*) - \theta^*]^2} - \frac{n}{\sigma^2} \\ &= \frac{n}{\sigma^2} \cdot \frac{1 - [h(\theta^*) - \theta^*][2h(\theta^*) - \theta^*]}{[h(\theta^*) - \theta^*]^2} \\ &= -\frac{n}{\sigma^2} \cdot \frac{h''(\theta^*)}{h(\theta^*)[h(\theta^*) - \theta^*]^2}. \end{aligned}$$

By large sample properties of maximum likelihood estimates, we have:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left[0, \left(-\frac{d^2 \ln L}{nd\theta^2}\right)^{-1}\right].$$

The variance of the maximum likelihood estimate of the gap length $\hat{\theta}$ can be approximated by:

$$\mathrm{Var}(\hat{\theta}) \approx \left(-\frac{d^2 \ln L}{d\theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}} = \frac{h(\hat{\theta}^*)[h(\hat{\theta}^*) - \hat{\theta}^*]^2}{h''(\hat{\theta}^*)} \cdot \frac{\sigma^2}{n}.$$

Based on the asymptotic normality for large samples, the approximate 95% confidence interval for the truncation point θ is:

$$\hat{\theta} \pm 1.96 \times \frac{\sqrt{h(\hat{\theta}^*)}[h(\hat{\theta}^*) - \hat{\theta}^*]}{\sqrt{h''(\hat{\theta}^*)}} \times \frac{\sigma}{\sqrt{n}}. \tag{3.12}$$

4. Results

Since the analytic form of the auxiliary functions $g_1(x)$, $g_2(x)$, and $g_3(x)$ cannot be given, we usually use numerical search algorithms, such as the Newton-Raphson method, to obtain approximate solutions to the maximum likelihood estimate of the truncation point or gap length θ. In practice, another simple approach is to construct the inverse function table based on the monotonicity of the auxiliary function and combine it with interpolation techniques, which can quickly obtain the inverse function values. In the following, we perform Monte Carlo simulations for the truncated normal model and the Lander-Waterman model, respectively, to evaluate the estimates and the corresponding confidence intervals given earlier. Finally, our estimation methods are tested on actual DNA sequencing dataset.

4.1. Results for truncated normal model

Using Monte Carlo simulations, we evaluated the large-sample nature of the point estimates of the truncation point θ in (2.1) and the interval estimates in (2.8). In our experiments, the parameters of the population X are set as $\mu = 180$ and $\sigma^2 = 37^2$. The combinations of truncation point $\theta = 30, 90, 150$ and sample size $n = 50, 500$ were simulated respectively, and each combination was sampled 1000 times.

It can be seen from Figure 2 that the kernel densities of the truncated point estimates in (2.1) for different combinations are very close to the corresponding curves of theoretical normal densities. The numbers in the figure indicate the proportion that the true parameters are covered by the approximate 95% confidence intervals, and these numbers show that the true confidence coefficients of the confidence intervals given by (2.8) are quite close to 95%. Therefore, the simulation results of the truncation point estimators given in this paper are completely consistent with the theoretical analysis results.

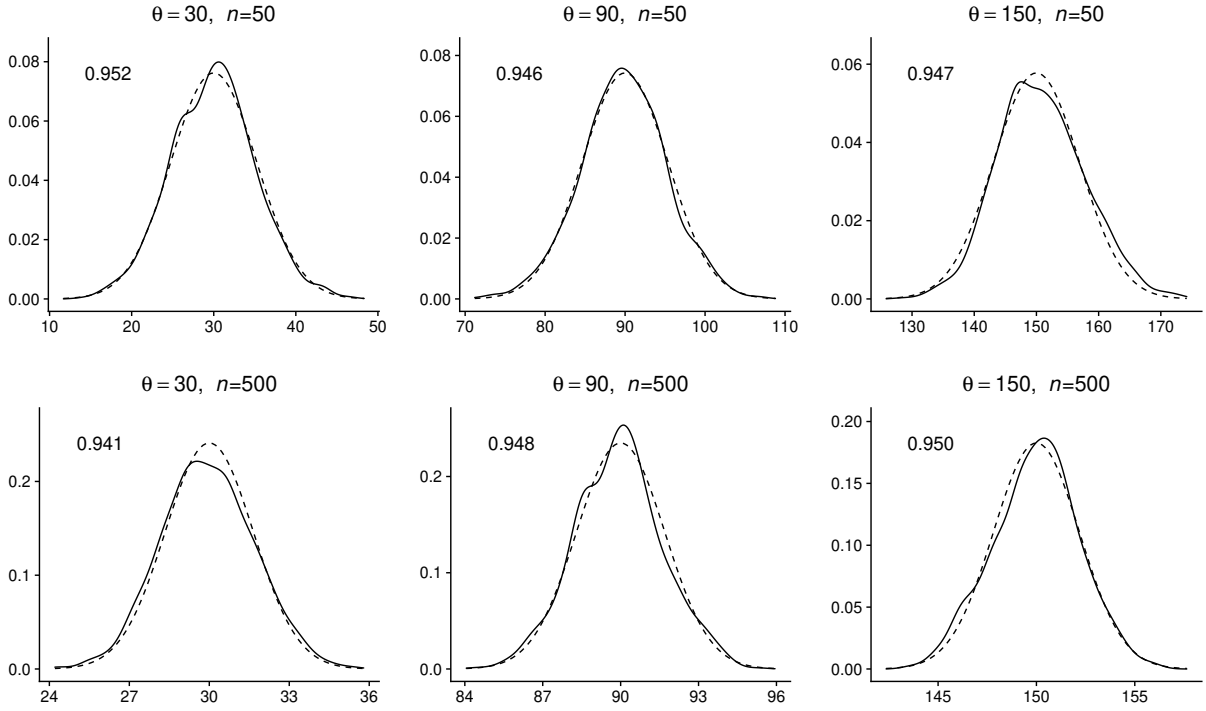

Figure 2. The simulated and theoretical densities estimated by the truncation points under different truncation points and different sample sizes. The solid lines are the densities of the simulation results, and the dashed lines are the theoretical densities. The numbers in the figure give the proportion that the true parameters are covered by the approximate 95% confidence intervals in the experiment.

4.2. Results for Lander-Waterman model

In this experiment, we randomly generated a DNA sequence of 316K bp in length with 500 gaps, and randomly generated 1.58M DNA fragments from it. The length of DNA fragment is normally distributed: $X \sim N(180, 37^2)$. In the same sequencing project, the larger the gap length θ, the less the number of fragments n spanning the gap, and there is roughly linear association between θ and n (see Figure 3).

In order to evaluate the estimates in (3.5) at different gap lengths, studentized errors were computed as follows:

$$t = \frac{\theta - \hat{\theta}}{\hat{\sigma}_\theta}$$

where the standard deviation $\hat{\sigma}_\theta$ was estimated by the square root of $V(\hat{\theta})$ in (3.11).

From this experiment, we found that 6.2% of the studentized error falls outside the approximate 95% confidence limit (see Figure 4). This shows that the true confidence coefficient of the approximate confidence interval given by (3.12) is very close to the nominal confidence coefficient of 0.95.

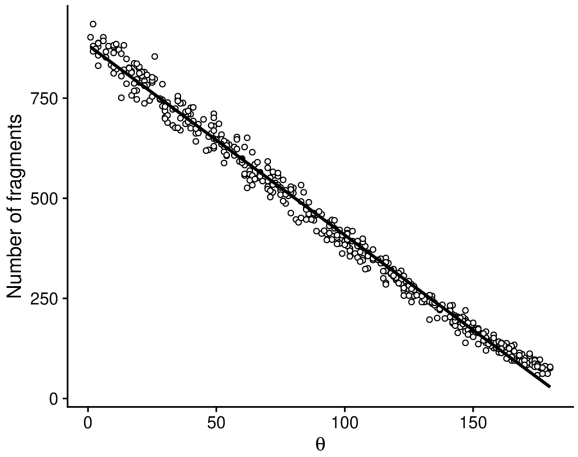

Figure 3. Number of fragments spanning gaps with different length.

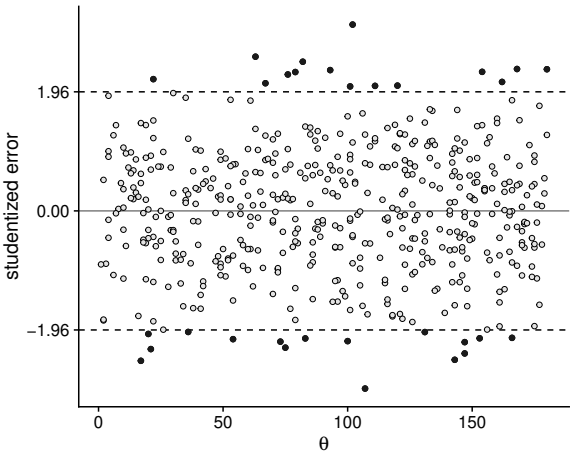

Figure 4. Studentized error in the estimation of gap length.

4.3. Actual DNA sequencing dataset

The Illumina whole-genome paired-end sequencing dataset SRR001665 of *E. coli* was used in this experiment. The total length of the target sequence was 4,639,675 bp, and a total of 20,816,448 pairs of read sequences of length 36 were found in this sequencing project. The dataset can be accessed online from https://www.ncbi.nlm.nih.gov/sra.

By mapping read pairs to reference sequences of the *E. coli* genome, we obtain the distribution of insert size for the SRR001665 dataset, which follows a normal distribution with mean $\mu = 215.43$ and standard deviation $\sigma = 10.51$ almost perfectly (see Figure 5).

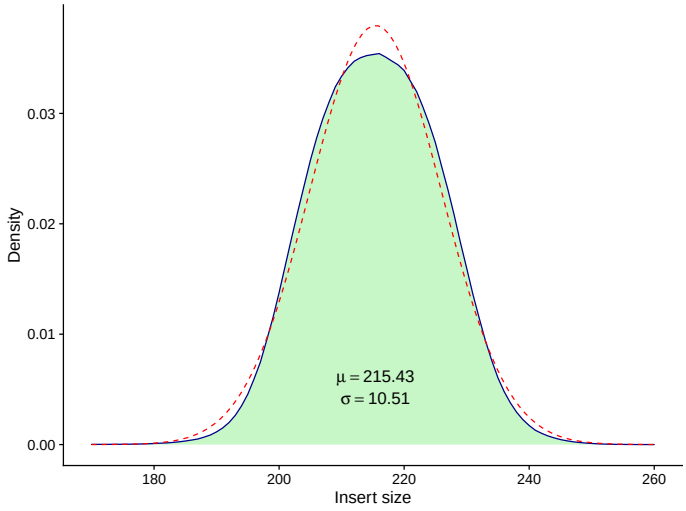

Figure 5. The distribution of insert size for the SRR001665 dataset.

First, we used the velvet [18] assembly software to assemble the reads of the sequencing dataset into contigs, and the paired-end information was not used in the assembly phase. Then, we utilized the paired-end information to find adjacent contigs by mapping mate pairs onto contigs. That is, one of

a read pair is on one contig and the other one is on some other contig, and so this read pair provides evidence that the two contig are adjacent. Furthermore, under the assumption of normality for the distribution of insert size, the length of the gap between two adjacent contigs can be estimated by (3.5) in Theorem 3.2.

For the large sample nature, we impose a limit on the number of mate pairs spanning the gap, i.e., we require the size of the truncated sample $\mathcal{Y}$ in (3.1) to be no less than 50. In the SRR001665 dataset, there are 542 gaps satisfying the condition $n \geq 50$, and the estimation results are shown in Figure 6.

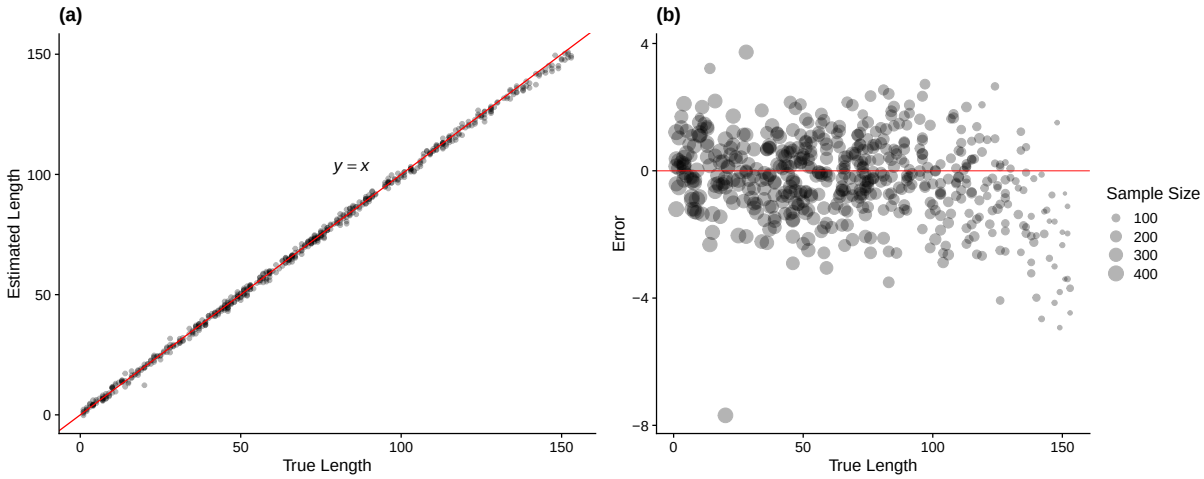

Figure 6. The estimation results for the gaps found in the SRR001665 dataset.

The experimental results show that the estimated values are very close to the true values (see Figure 6a). Most of the estimates have errors within 4, with one outlier that has an error close to 8 (see Figure 6b). In fact, by the estimate variance in (3.11), outliers can be used to diagnose whether the connection of two contigs considered to be adjacent is reliable.

It should be noted that the performance of the estimator (3.5) is sensitive to the normality. Small departures from normality do not create any serious problems in the estimates of gap length. However, if the distribution of insert size depart far from the normality, the estimator may have notable bias.

In fact, the length distribution of inserts is mainly affected by different platforms and their operational parameters, and in general, the insert size in a specific project approximately follow a normal distribution, which is generally slightly right-skewed [19].

5. Conclusions

In this paper, we study the truncation point estimation for truncated normal distribution, where the mean μ is known and the actual sample observations are the original observations after subtracting the truncation point. The cases in which the variance σ^2 is known and unknown are discussed, and the moment method and the maximum likelihood method are employed to give the estimators of the unknown parameters. When σ^2 is known, the two methods obtain the same estimator of the truncation point. When σ^2 is unknown, the two methods yield similar results, and the key difference lies in the estimation of the sample variance. Approximate confidence intervals for truncation point θ are given for large samples. Based on the Lander-Waterman model, the method of estimating truncation

points is extended to the estimation of gap length in DNA sequencing, and the point estimation and interval estimation of gap length under the assumption of normality are given. The Monte Carlo experimental results show that the estimators given in this paper are consistent with the results of the simulation experiments, whether it is the truncated normal model or the Lander-Waterman model. Experimental results from actual DNA sequencing dataset show that accurate estimates of gap length can be obtained by the estimation method proposed in this paper when the insert size is approximately normally distributed.

Acknowledgments

This work was supported by the Foundation of Jiangxi Educational Committee under Grant GJJ150926; Jingdezhen Ceramic University National Level Student Innovation and Entrepreneurship Training Program Project under Grant 202110408025.

Conflict of interest

The authors declare no conflict of interest.

References

1. W. C. Horrace, Moments of the truncated normal distribution, *J. Prod. Anal.,* **43** (2015), 133–138. http://dx.doi.org/10.1007/s11123-013-0381-8
2. J. Pender, The truncated normal distribution: Applications to queues with impatient customers, *Oper. Res. Lett.,* **43** (2015), 40–45. https://doi.org/10.1016/j.orl.2014.10.008
3. K. Pearson, A. Lee, On the generalized probable error in multiple normal correlation, *Biometrika,* **6** (1908), 59–68. http://dx.doi.org/10.1093/biomet/6.1.59
4. R. A. Fisher, *Properties and applications of Hh functions*, in Mathematical Tables, British Association for the Advancement of Science, 1931.
5. C. I. Bliss, W. L. Stevens, The calculation of the time mortality curve, *Ann. Appl. Biol.,* **24** (1937), 815–852. http://dx.doi.org/10.1111/j.1744-7348.1937.tb05058.x
6. A. Hald, Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point, *Scand. Actuar. J.,* **1** (1949), 119–134. http://dx.doi.org/10.1080/03461238.1949.10419767
7. A. K. Gupta, Estimation of the mean and standard deviation of a normal population from a censored sample, *Biometrika,* **39** (1952), 260–273. http://dx.doi.org/10.2307/2334023
8. M. Halperin, Maximum likelihood estimation in truncated samples, *Ann. Math. Stat.,* **23** (1952), 226–238. http://dx.doi.org/10.2307/2236448
9. A. C. Cohen, Simplified estimators for the normal distribution when samples are singly censored or truncated, *Technometrics,* **1** (1959), 217–237. http://dx.doi.org/10.1080/00401706.1959.10489859
10. A. C. Cohen, Tables for maximum likelihood estimates: Singly truncated and singly censored samples, *Technometrics,* **3** (1961), 535–541. http://dx.doi.org/10.1080/00401706.1961.10489973

11. A. C. Cohen, *Truncated and censored samples theory and applications*, New York: Marcel Dekker, 1991.

12. D. S. Robson, J. H. Whitlock, Estimation of a truncation point, *Biometrika,* **51** (1964), 33–39. http://dx.doi.org/10.2307/2334193

13. Z. W. Birnbaum, An inequality for Mill's ratio, *Ann. Math. Stat.,* **13** (1942), 245–246. http://dx.doi.org/10.1214/aoms/1177731611

14. M. R. Sampford, Some inequalities on Mill's ratio and related functions, *Ann. Math. Stat.,* **24** (1953), 130–132. http://dx.doi.org/10.2307/2236360

15. Z. H. Yang, Y. M. Chu, On approximating Mills ratio, *J. Inequal. Appl.,* **273** (2015), 273. http://dx.doi.org/10.1186/s13660-015-0792-3

16. E. S. Lander, M. S. Waterman, Genomic mapping by fingerprinting random clones, *Genomics,* **2** (1988), 231–239. http://dx.doi.org/10.1016/0888-7543(88)90007-9

17. J. C. Roach, C. Boysen, K. Wang, L. Hood, Pairwise end sequencing: A unified approach to genomic mapping and sequencing, *Genomics,* **26** (1995), 345–353. http://dx.doi.org/10.1016/0888-7543(95)80219-C

18. D. R. Zerbino, E. Birney, Algorithms for de novo short read assembly using de Bruijn graphs, *Genome Res.,* **18** (2008), 821–829. http://dx.doi.org/10.1101/gr.074492.107

19. J. Foox, S. W. Tighe, C. M. Nicolet, J. M. Zook, M. Byrska-Bishop, W. E. Clarke, et al., Performance assessment of DNA sequencing platforms in the ABRF next-generation sequencing study, *Nat. Biotechnol.,* **39** (2021), 1129–1140. http://dx.doi.org/10.1038/s41587-021-01049-5

AIMS Press