Estimating Sigma for the Distribution of Zebra Mussels in Lake Bergen

Helen Moses and Kaitlyn Peterson

March 8th, 2023

Introduction

In order to study the inner-workings of ecosystems, it is important to understand population abundance of the species present in these ecosystems. Often, it is far more feasible to obtain this information through sampling and estimation rather than attempting to collect data about the entire population. One way to estimate population abundance is through the utilization of distance sampling. Distance sampling involves an observer, who traverses established lines (called transects), recording the locations of observed organisms along those transects. In one study of zebra mussels in Minnesota lakes, scuba divers collected data through distance sampling in Lake Bergen (data shown in Figure 1).

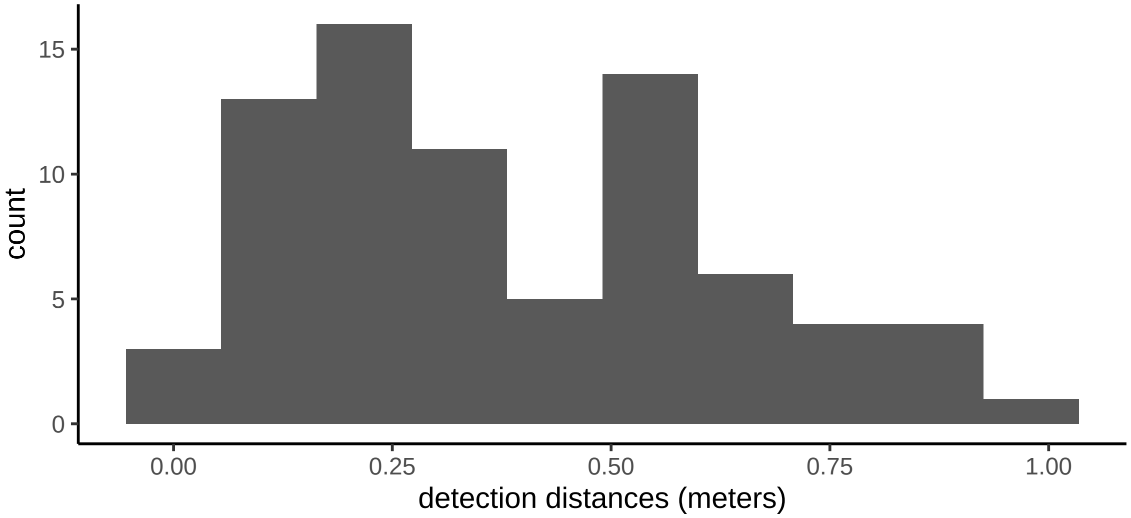


Figure 1.

Based on population data, researchers are able to discern probability models and probability density functions. These models and detection probabilities inform population estimation. Previous research found that individual values from random samples of observed distances of zebra mussels follow a half-normal distribution. However, the population parameter sigma squared (σ^2) is unknown. The value of σ^2 is needed in order to calculate detection probabilities, which are then used to estimate how many mussels are in the lake.

This paper thus derives and compares competing estimators of σ^2, with the ultimate goal of selecting the best estimator. This estimator will help describe the population of zebra mussels and allow researchers to obtain information about population abundance in Lake Bergen.

Methods

As previously stated, past research found that zebra mussel distances from the transect in Lake Bergen follow a half normal distribution. Thus, the model can be defined as $X_i \sim$ half normal, where each X_i represents an individual detection distance. This distribution has the following pdf, expected value, and variance:

$f(x) = \frac{2}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}$ for $x \geq 0$
$E(X) = \sigma\sqrt{\frac{2}{\pi}}$
$V(X) = \sigma^2(1 - \frac{2}{\pi})$

Using this data model, estimators of σ^2 were constructed using both Maximum Likelihood Estimation (MLE) and Method of Moments Estimation (MOM). See Appendix 1.b for the calculation of the MOM estimator.

$\hat{\sigma}^2_{MLE} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2$
$\hat{\sigma}^2_{MOM} = \frac{\pi(\bar{X})^2}{2}$

To evaluate these estimators, we derived the bias by subtracting σ^2 from the expected value of each estimator. We found that the MLE estimator is unbiased, but that the MOM estimator has positive but asymptotic bias that follows the formula $Bias[\hat{\sigma}^2_{MOM}] = \frac{\sigma^2(\frac{\pi}{2}-1)}{n}$. See Appendix 1.c for calculations of biases.

Another technique to evaluate estimators is comparing the confidence intervals. To aid with these calculations, it is helpful to know the distribution of the estimator. We found the distribution of the MLE estimator to be $\hat{\sigma}^2_{MLE} \sim Gamma(\frac{n}{2}, \frac{n}{2\sigma^2})$ through the distribution function method (See Appendix 1.a). With the quantiles from the pivotal stat distribution: $\frac{\hat{\sigma}^2_{MLE}}{\sigma^2} \sim Gamma(\frac{n}{2}, \frac{n}{2})$, we can calculate the lower and upper bounds of a 95% confidence interval (See Appendix 1.d). We found the following equations for the upper ($U(x)$) and lower ($L(x)$) bounds:

$L(x) = \dfrac{\frac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2}{q_{0.975}}$
$U(x) = \dfrac{\frac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2}{q_{0.025}}$

To evaluate these equations, we constructed a simulation study. As the distribution of each individual X_i is known, we can pull random samples from a half normal distribution after fixing arbitrary values for n and σ. The results for this simulation will allow us to conclude which estimator is the most accurate with the least amount of variability.

After fixing values of σ and sample size (n), we drew 10,000 random samples from the half normal distribution. For each randomly generated sample, we calculated an estimation of σ^2 based on the derived estimator formulas. Because we had found the distribution of the MLE estimator, we were also able to calculate the CI coverage and length for each random sample, based on the lower and upper bound calculations described above. Then for each sample, we ran a bootstrap simulation, calculating the estimation of σ^2 based on the bootstrap resample, for each of the 10,000 resamples. With the results of each simulation, we calculated the bias, mean square error (MSE), and bootstrap CI coverage and interval length for each of the estimators.

We then repeated this process across various values of n and σ. To test a wide variety of estimations, we chose small, medium, and large values of both σ and n (σ = 0.5, 4, 15; n=50, 300, 800). This allows us to conclude, regardless of the true values of n and σ, which estimator should be used to inform zebra mussel population abundance.

Note: An example of one run of this simulation can be found in the R Markdown Appendix.

Results

As predicted by the calculations of the theoretical bias of each estimator, the simulations found that $\hat{\sigma}^2_{MLE}$ has a smaller bias than $\hat{\sigma}^2_{MOM}$ (Figure 2). Regardless of the values of σ and n, this holds true. The bias of the MLE estimator is negligible across all simulations ($-0.0353 \leq Bias[\hat{\sigma}^2_{MLE}] \leq 0.438$) and any values other than 0 reflect simulation error.

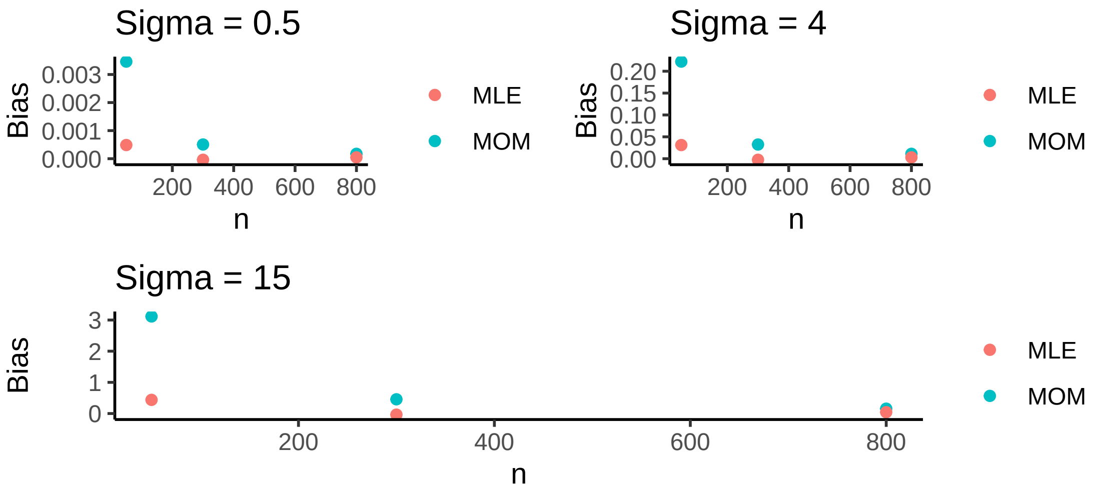


Figure 2.

Similarly, in every simulation, the values of $MSE[\hat{\sigma}^2_{MLE}]$ are smaller than those for the $MSE[\hat{\sigma}^2_{MOM}]$ (Tables 1 and 2). For example, with n = 50 and $\sigma = 4$, $MSE[\hat{\sigma}^2_{MLE}] = 10.197 \leq MSE[\hat{\sigma}^2_{MOM}] = 11.87$. These results indicate that the MLE estimator is more accurate and less variable.

Table 1: MOM estimator MSE

	n = 50	n = 300	n = 800
$\sigma = 0.5$	0.00290	0.000475	0.000173
$\sigma = 4$	11.87	1.946	0.710
$\sigma = 15$	2347.304	384.924	140.432

Table 2: MLE estimator MSE

	n = 50	n = 300	n = 800
$\sigma = 0.5$	0.00249	0.000411	0.000150
$\sigma = 4$	10.197	1.6842	0.615
$\sigma = 15$	2016.582	333.066	121.645

Complicating these findings is the bootstrap coverage of each estimator. For smaller sample sizes, the coverage rate for each estimator fell short of the desired 95% (for n of 50, MOM coverage is 0.9424 and MLE coverage is 0.9197) (Table 3). While in this instance the MOM coverage is closer to the desired value, as sample size becomes very large, both estimators have values closer to the desired coverage (for n of 800, MOM coverage is 0.9507 and MLE coverage is 0.9504) (Table 3). On the other hand, the formula based coverage for the MLE estimator is more consistently closer to the desired value across all sample sizes (0.9508, 0.9506, and 0.9536 for n=50, 300, and 800 respectively) (Table 3). However, it is also important to consider interval length. The MOM estimator has the longest CI interval lengths (Table 4). For the MLE estimator, the bootstrapped CI interval length is smaller than the formula based CI interval length (Tables 5 and 6).

Table 3: Coverage by estimation method

	n = 50	n = 300	n = 800
MOM Estimator Bootstrap Coverage	0.9424	0.9451	0.9507
MLE Estimator Bootstrap Coverage	0.9197	0.9454	0.9507
MLE Estimator Formula Based Coverage	0.9508	0.9506	0.9536

Table 4: MOM estimator bootstrap confidence interval length

	n = 50	n = 300	n = 800
$\sigma = 0.5$	0.209	0.0854	0.0523
$\sigma = 4$	13.364	5.467	3.350
$\sigma = 15$	187.926	76.849	47.105

Table 5: MLE estimator bootstrap confidence interval length

	n = 50	n = 300	n = 800
$\sigma = 0.5$	0.187	0.0793	0.0489
$\sigma = 4$	11.989	5.076	3.127
$\sigma = 15$	168.598	71.379	43.974

Table 6: MLE estimator confidence interval length

	n = 50	n = 300	n = 800
$\sigma = 0.5$	0.212	0.0810	0.0492
$\sigma = 4$	13.549	5.184	3.151
$\sigma = 15$	190.532	72.902	44.313

Based on these results, it is clear that the MLE estimator is more accurate and less variable than the MOM estimator. Therefore, we used the MLE estimator on the Lake Bergen data to obtain the point estimate of $\sigma^2 = 0.2075$.

For n = 50, this formula CI has a coverage rate closer to the desired 95% (95.08%) than the bootstrap method (91.97%) (Table 3). Even though the intervals for the bootstrap CIs are smaller, it is more important to achieve the desired confidence level. Due to the fact that the Lake Bergen data has a sample size of 77, which is closest to our simulations where n = 50, we found the confidence interval using the formula based CI calculation. This resulted in the confidence interval of 0.1549 to 0.2925. Thus we are 95% confident that the true value of σ^2 is between 0.1549 and 0.2925.

Discussion

With these point and interval estimates of σ^2, we were then able to calculate point and interval estimates for p_x, the probability of detecting a mussel that is x meters off the transect. We did so by substituting the point and interval estimates into the following equation for p_x (the probability density function of the half normal distribution): $p_x = e^{-x^2/(2\sigma^2)}$. This resulted in the point estimates of 0.0899 for x = 1 meter and 0.5475 for x = 0.5 meters. The interval estimates are (0.0396, 0.181) and (0.4462, 0.652) respectively.

These probabilities provide two useful insights. First, the probability of detection just one meter off the transect is extremely low (8.99%), and the probability of detection a half meter off the transect is just over

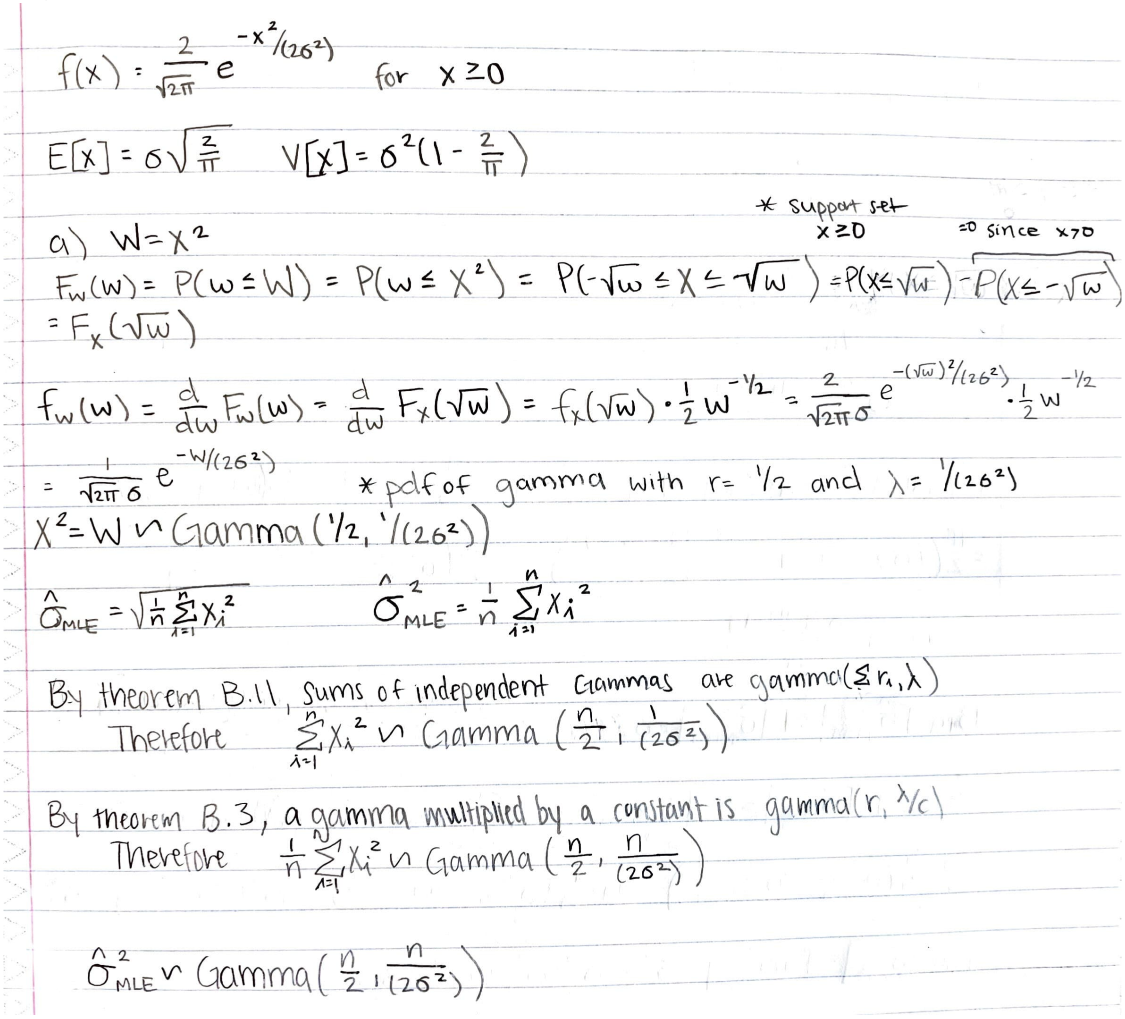
50% (54.75%). This represents the difficulty of surveying the exact counts of zebra mussels and motivates our method of estimating the population abundance. Second, these probabilities, in conjunction with the data on the observed muscles, allow researchers to obtain a close estimation of the population abundance. For example, if scuba divers detect y mussels with a 0.5 probability of detection, it follows that y mussels were missed, leading to a population estimation of 2y.

While our estimations are helpful in estimating population abundance, it is important to note the limitations of our simulations. Our biggest limitation was computing power. A replication value of 10,000 is fairly large, but still leads to non-negligible simulation error. This stands out with the bias of the MLE estimator. Our simulated values are non-zero despite our theoretical calculations that show the MLE estimator is unbiased (Appendix 1.c). Furthermore, unlimited computing power would allow for the simulation of a wider variety of values for n and σ. More values could have better informed the choices of estimator and CI calculation. Future research should explore the impact of greater computing capabilities.

Finally, with a data set of only 77 entries, the size of the Lake Bergen data set informed our choice of CI calculation. With larger values of n, we would have chosen the bootstrap method, which would have resulted in a shorter confidence interval than the formula method. As such, we hope that future research will be able to gather more data points from Lake Bergen to allow for a more precise estimation of σ.

Appendix

1.a) Calculation of distribution of MLE estimator



$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/(2\sigma^2)}$ for $x \geq 0$

$E[x] = \sigma\sqrt{\frac{2}{\pi}}$ $V[x] = \sigma^2(1 - \frac{2}{\pi})$

* support set $x \geq 0$

a) $W = X^2$

$F_W(w) = P(w \leq W) = P(w \leq X^2) = P(-\sqrt{w} \leq X \leq \sqrt{w}) = P(X \leq \sqrt{w}) - \overbrace{P(X \leq -\sqrt{w})}^{=0 \text{ since } x>0}$

$= F_X(\sqrt{w})$

$f_W(w) = \frac{d}{dw} F_W(w) = \frac{d}{dw} F_X(\sqrt{w}) = f_X(\sqrt{w}) \cdot \frac{1}{2} w^{-1/2} = \frac{2}{\sqrt{2\pi}\sigma} e^{-(\sqrt{w})^2/(2\sigma^2)} \cdot \frac{1}{2} w^{-1/2}$

$= \frac{1}{\sqrt{2\pi}\sigma} e^{-W/(2\sigma^2)}$ * pdf of gamma with $r = 1/2$ and $\lambda = 1/(2\sigma^2)$

$X^2 = W \sim \text{Gamma}(1/2, 1/(2\sigma^2))$

$\hat{\sigma}_{MLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$ $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

By theorem B.11, sums of independent Gammas are gamma$(\sum r_i, \lambda)$

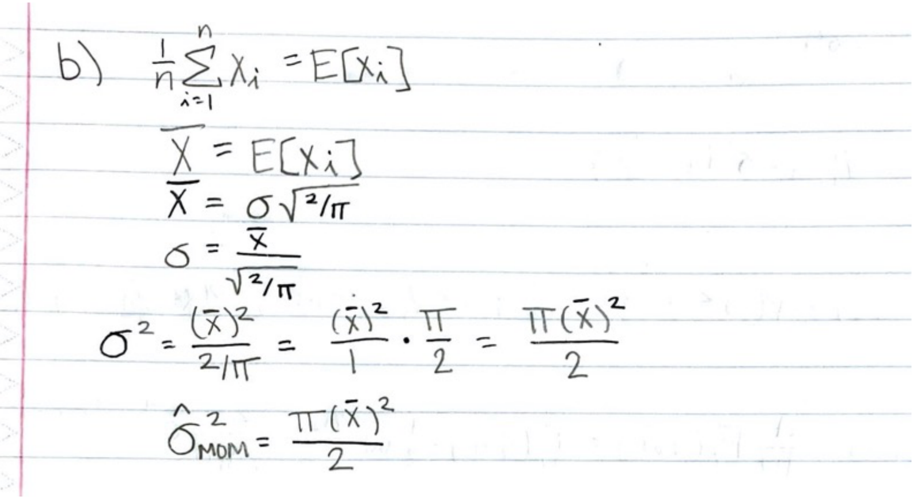
Therefore $\sum_{i=1}^{n} X_i^2 \sim \text{Gamma}(\frac{n}{2}, \frac{1}{(2\sigma^2)})$

By theorem B.3, a gamma multiplied by a constant is gamma$(r, \lambda/c)$

Therefore $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \sim \text{Gamma}(\frac{n}{2}, \frac{n}{(2\sigma^2)})$

$\hat{\sigma}^2_{MLE} \sim \text{Gamma}(\frac{n}{2}, \frac{n}{(2\sigma^2)})$

1.b) Calculation of MOM estimator



b) $\frac{1}{n}\sum_{i=1}^{n} X_i = E[X_i]$

$\overline{X} = E[X_i]$

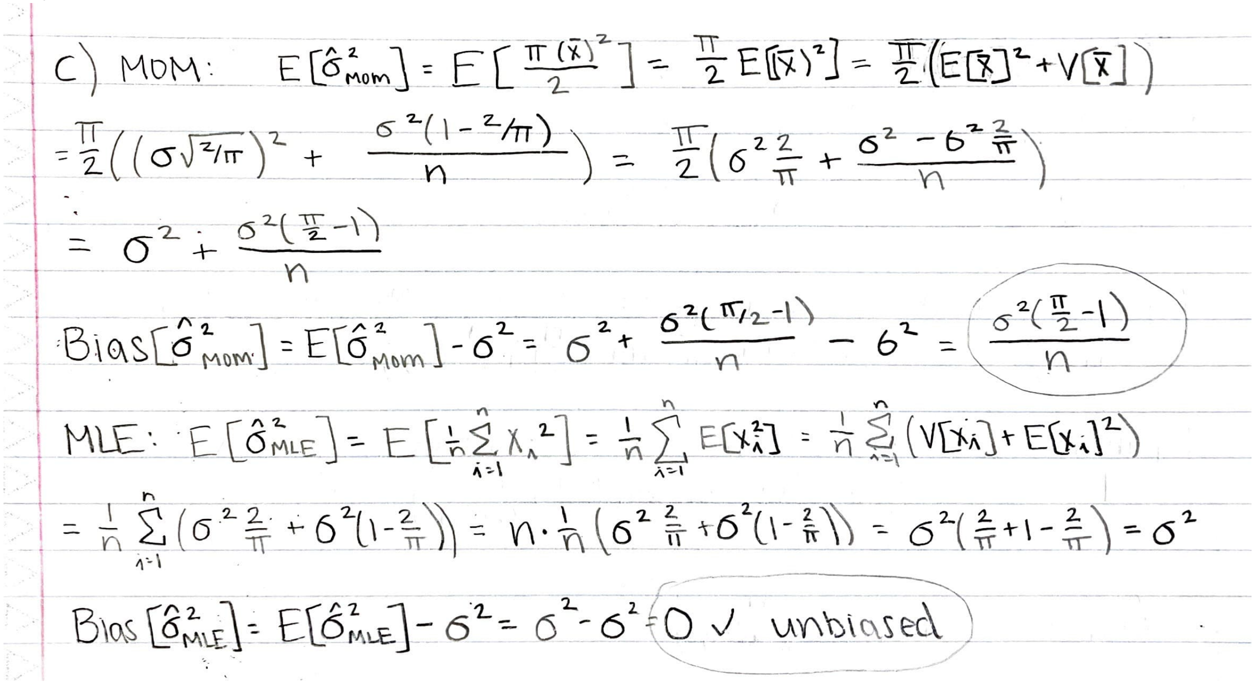
$\overline{X} = \sigma\sqrt{2/\pi}$

$\sigma = \frac{\overline{X}}{\sqrt{2/\pi}}$

$\sigma^2 = \frac{(\bar{x})^2}{2/\pi} = \frac{(\bar{x})^2}{1} \cdot \frac{\pi}{2} = \frac{\pi(\bar{x})^2}{2}$

$\hat{\sigma}^2_{MOM} = \frac{\pi(\bar{x})^2}{2}$

1.c) Calculation of biases for both estimators



c) MOM: $E[\hat{\sigma}^2_{Mom}] = E\left[\frac{\pi(\bar{X})^2}{2}\right] = \frac{\pi}{2}E[(\bar{X})^2] = \frac{\pi}{2}(E[\bar{X}]^2 + V[\bar{X}])$

$= \frac{\pi}{2}\left((\sigma\sqrt{2/\pi})^2 + \frac{\sigma^2(1-2/\pi)}{n}\right) = \frac{\pi}{2}\left(\sigma^2\frac{2}{\pi} + \frac{\sigma^2 - \sigma^2\frac{2}{\pi}}{n}\right)$

$= \sigma^2 + \frac{\sigma^2(\frac{\pi}{2}-1)}{n}$

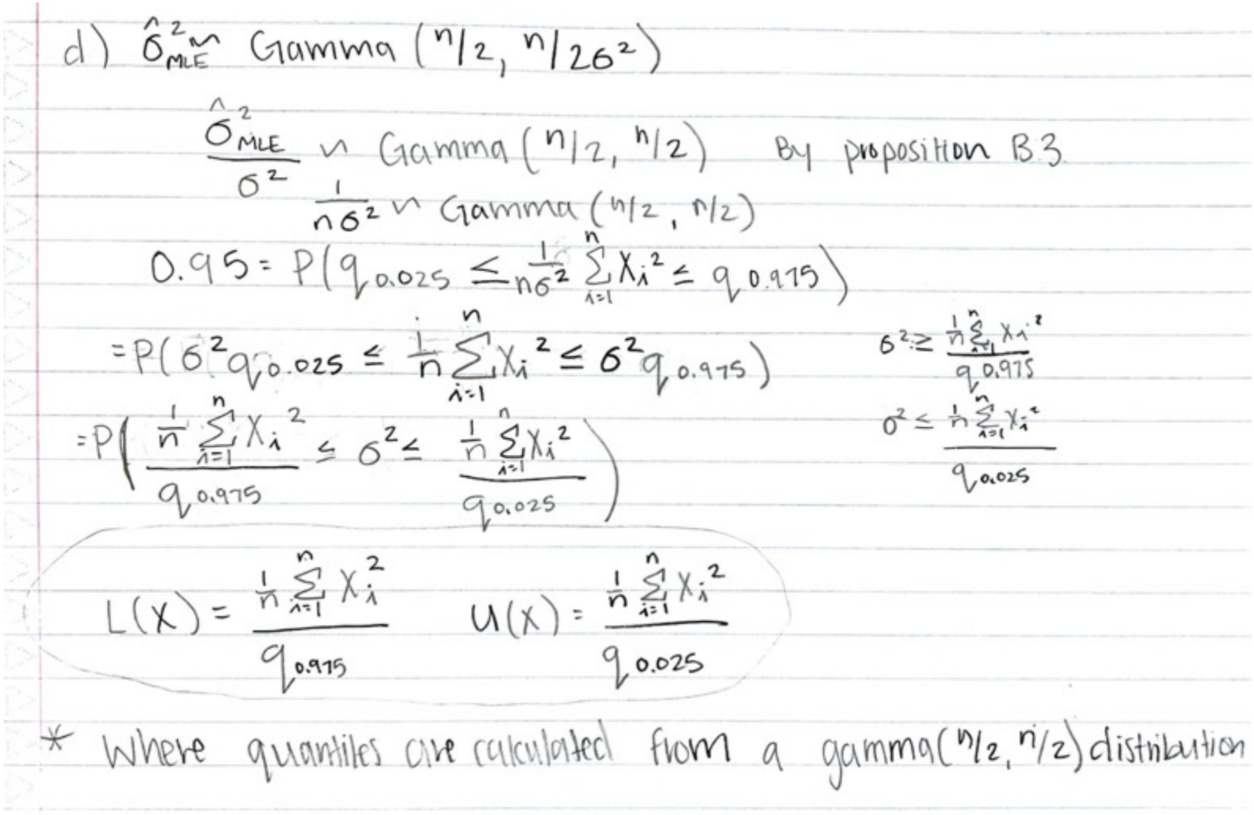
$\text{Bias}[\hat{\sigma}^2_{MOM}] = E[\hat{\sigma}^2_{Mom}] - \sigma^2 = \sigma^2 + \frac{\sigma^2(\pi/2-1)}{n} - \sigma^2 = \boxed{\frac{\sigma^2(\frac{\pi}{2}-1)}{n}}$

MLE: $E[\hat{\sigma}^2_{MLE}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] = \frac{1}{n}\sum_{i=1}^{n}(V[X_i] + E[X_i]^2)$

$= \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2\frac{2}{\pi} + \sigma^2(1-\frac{2}{\pi})\right) = n\cdot\frac{1}{n}\left(\sigma^2\frac{2}{\pi} + \sigma^2(1-\frac{2}{\pi})\right) = \sigma^2\left(\frac{2}{\pi}+1-\frac{2}{\pi}\right) = \sigma^2$

$\text{Bias}[\hat{\sigma}^2_{MLE}] = E[\hat{\sigma}^2_{MLE}] - \sigma^2 = \sigma^2 - \sigma^2 = \boxed{0 \checkmark \text{ unbiased}}$

1.d) Calculation of formula MLE confidence interval



d) $\hat{\sigma}^2_{MLE} \sim \text{Gamma}(n/2, n/2\sigma^2)$

$\frac{\hat{\sigma}^2_{MLE}}{\sigma^2} \sim \text{Gamma}(n/2, n/2)$ By proposition B3

$\frac{1}{n\sigma^2} \sim \text{Gamma}(n/2, n/2)$

$0.95 = P\left(q_{0.025} \le \frac{1}{n\sigma^2}\sum_{i=1}^{n} X_i^2 \le q_{0.975}\right)$

$= P\left(\sigma^2 q_{0.025} \le \frac{1}{n}\sum_{i=1}^{n} X_i^2 \le \sigma^2 q_{0.975}\right)$

$= P\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{q_{0.975}} \le \sigma^2 \le \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{q_{0.025}}\right)$

$\sigma^2 \ge \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{q_{0.975}}$

$\sigma^2 \le \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{q_{0.025}}$

$$\boxed{L(X) = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{q_{0.975}} \qquad U(X) = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{q_{0.025}}}$$

* Where quantiles are calculated from a gamma($n/2$, $n/2$) distribution

R Markdown Appendix

Simulations

```r
# Making the results reproducible
set.seed(647465357)
# Sample size
n <- 50
# Set value of sigma
sigma <- 15

# Number of replications
N <- 10000

# Setting up empty vectors
est_mle <- rep(NA, N)
est_mom <- rep(NA, N)
mle_lower <- rep(NA, N)
mle_upper <- rep(NA, N)
mle_boot_lower <- rep(NA, N)
mle_boot_upper <- rep(NA, N)
mom_boot_lower <- rep(NA, N)
mom_boot_upper <- rep(NA, N)

for (i in 1:N) {
  # Generating a half normal random sample
  x <- abs(rnorm(n, 0, sigma))

  # Computing point estimates based on the formulas for mle and mom
  est_mle[i] <- (1/n)*(sum(x^2))
  est_mom[i] <- (pi/2)*(mean(x)^2)

  # Computing lower and upper bounds based on the exact mle CI (Appendix 1. d)
  mle_lower[i] <- (est_mle[i])/ qgamma(.975, n/2, n/2)
  mle_upper[i] <- (est_mle[i])/ qgamma(.025, n/2, n/2)

  # Setting up empty vector
  boot_mle <- rep(NA, N)
  boot_mom <- rep(NA, N)

  # Setting up the for loop to bootstrap each sample
  for (j in 1:N) {
    resample <- sample(x, n, replace = TRUE)
    boot_mle[j] <- (1/n)*(sum(resample^2))
    boot_mom[j] <- (pi/2)*(mean(resample)^2)
  }

  # Storing the bootstrapped values
  mle_boot_lower[i] <- quantile(boot_mle, .025)
  mle_boot_upper[i] <- quantile(boot_mle, .975)

  mom_boot_lower[i] <- quantile(boot_mom, .025)
  mom_boot_upper[i] <- quantile(boot_mom, .975)
}
```

```r
# Calculating the means of all of the estimates
avg_est_mle <- mean(est_mle)
avg_est_mle
```

```
## [1] 225.4385
```

```r
avg_est_mom <- mean(est_mom)
avg_est_mom
```

```
## [1] 228.115
```

```r
avg_mle_lower <- mean(mle_lower)
avg_mle_upper <- mean(mle_upper)
avg_mle_boot_lower <- mean(mle_boot_lower)
avg_mle_boot_upper <- mean(mle_boot_upper)
avg_mom_boot_lower <- mean(mom_boot_lower)
avg_mom_boot_upper <- mean(mom_boot_upper)

# Calculating bias of the point estimates
mom_bias <- avg_est_mom - sigma^2
mom_bias
```

```
## [1] 3.115014
```

```r
mle_bias <- avg_est_mle - sigma^2
mle_bias
```

```
## [1] 0.4384716
```

```r
# Calculating the percent bias for the point estimates
mom_perc_bias <- (100*mom_bias)/(sigma^2)
mle_perc_bias <- (100*mle_bias)/(sigma^2)

# Calculating the mse for the point estimates
mom_mse <- mean(((est_mom) - (sigma^2))^2)
mom_mse
```

```
## [1] 2347.304
```

```r
mle_mse <- mean(((est_mle) - (sigma^2))^2)
mle_mse
```

```
## [1] 2016.582
```

```r
# Calculating the coverage and length of the interval for mle CI and the bootstrapped CIs
mle_cover <- mean((mle_lower <= sigma^2) & (sigma^2 <= mle_upper))
mle_cover
```

```
## [1] 0.9508
```

```r
mle_mean_length <- mean(mle_upper - mle_lower)
mle_mean_length
```

```
## [1] 190.5319
```

```r
mle_boot_cover <- mean((mle_boot_lower <= sigma^2) & (sigma^2 <= mle_boot_upper))
mle_boot_cover
```

```
## [1] 0.9197
```

```r
mle_boot_mean_length <- mean(mle_boot_upper - mle_boot_lower)
mle_boot_mean_length
```

```
## [1] 168.5978
```

```r
mom_boot_cover <- mean((mom_boot_lower <= sigma^2) & (sigma^2 <= mom_boot_upper))
mom_boot_cover
```

```
## [1] 0.9424
```

```r
mom_boot_mean_length <- mean(mom_boot_upper - mom_boot_lower)
mom_boot_mean_length
```

```
## [1] 187.9264
```

Estimating with the Data from Lake Bergen

```r
# Loading the collected data
lake <- read.csv("http://math.carleton.edu/kstclair/data/BergenData.csv")
```

```r
# Using the MLE estimator to get an estimation for sigma based on the real data
real_estimation <- (1/length(lake$distance))*(sum(lake$distance^2))
real_estimation
```

```
## [1] 0.2075276
```

```r
# Using the formula based confidence interval
real_mle_lower <- (real_estimation)/ qgamma(.975, n/2, n/2)
real_mle_lower
```

```
## [1] 0.1452864
```

```r
real_mle_upper <- (real_estimation)/ qgamma(.025, n/2, n/2)
real_mle_upper
```

```
## [1] 0.3206806
```

```r
# Estimation of detection probabilities in Lake Bergen
x <- c(.5, 1)

# Point estimate
prob_x <- exp(((-x^2)/(2*real_estimation)))
prob_x
```

```
## [1] 0.54753465 0.08987656
```

```r
# Confidence interval estimate
prob_lower <- exp(((-x^2)/(2*real_mle_lower)))
prob_lower
```

```
## [1] 0.42300560 0.03201728
```

```r
prob_upper <- exp(((-x^2)/(2*real_mle_upper)))
prob_upper
```

```
## [1] 0.6771951 0.2103077
```