Analysis of Size-biased Mitochondria Data

Yin-Ting Chou

Advisor: Aaron Rendahl

May 2nd 2017

Abstract

The goal of this project is to find out whether properties (area, perimeter, circularity and aspect ratio) of mitochondria are different by locations (proximal end, middle and distal end) in a single muscle fiber cell from a young mouse. However, the observed data is not a random sample but a size-biased sample, so the Arithmetic Mean is biased as an estimator for population mean. I instead use the concept of the Weighted Distribution to explore both parametric and nonparametric estimators of the population mean in the case of a size-biased sample. I first did a simulation study, which confirmed that arithmetic mean was inappropriate but that estimators based on the weighted distribution were appropriate with the parametric version having more assumptions but smaller standard deviation. I then applied these estimators in a Permutation Test, and constructed Bootstrapped confidence intervals for the parameters. I found mitochondria at the Middle have significantly larger Area and Perimeter and ones close to Distal end have significantly smaller Area, Perimeter and Circularity. Although the data in this project is from a single muscle fiber cell, the analysis results provide a strong foundation for future research on more cells.

[Intentionally blank page]

1 Introduction

This thesis is an extended study of my summer statistical consulting project which I worked in a pair with a PhD student, Ming Guo, at the Statistical Consulting Center in 2015. The client for this project was Professor Arriaga, the head of the Organelle Research Group in the University of Minnesota. The mission of his group is to answer complex questions related to the aging process, diabetes, obesity and neurodegeneration by starting with single cell and subcellular studies. This project is a preliminary study about the aging process and the main goal of this project is to understand whether properties (area, perimeter, circularity and aspect ratio) of mitochondria are different by locations (proximal, middle and distal end) in a single muscle fiber cell from a young mouse. After the main question is answered in this one cell, Prof. Arriaga team will start to test on many other cells. So, a secondary goal is to evaluate this data collecting procedure and provide suggestions on sampling methods. Due to the relationship between mitochondria and cell aging, the results of this project may be crucial to future research on the aging process, as described by Bratic and Larsson (2013).

The data of this project was collected by using the Transmission Electron Microscopy (TEM) technique to obtain super-resolution images of the single muscle fiber cell and then to manually measure properties (perimeter, area, circularity and aspect ratio) of sampled mitochondria from these images. Instead of choosing samples by using simple random sampling method, the sampled mitochondria were chosen if their area in the photo included one or more generated two-dimensional coordinates. A list of random coordinates was generated at the beginning of the sampling procedure. As a result of this procedure, each mitochondrion in the cell had a different probability of being picked as the sample and its sampling probability was proportional to its size (PPS), or its Area in

this case. In this situation, if Arithmetic Mean was used as our estimator for the population mean Area, it would be overestimated even though it's widely used as the best estimator for population mean. To deal with this problem, Cox (1962) first proposed the concept of the Weighted Distribution to adjust the current sampling density function when samples are size-biased rather than random, and then Patil and Ord (1976) further extended this concept on parametric distributions.

To find out the appropriate estimator of population mean as samples are size-biased, a simulation study was employed to test performance of the candidate estimators based on the selecting criteria, Root of MSE. The candidate estimators for the population mean Area are Arithmetic Mean (AM), Weighted Mean (WM) and Maximum Likelihood Estimator (MLE). Though we already knew that Arithmetic Mean would be overestimated, we were still interested in how bad it can be compared to other estimators. I have tried to add one more estimator which was based on the idea of Jones (1991) about estimating Weighted Distribution by kernel density. However, the result of this estimator was so close to Weighted Mean that I decided to drop this estimator on the candidate list. As for Perimeter, as we can see in Figure 1.0.1, Area and Perimeter had a strong positive relation which means that mitochondria with large size also have large perimeter. If mitochondria with larger size are more easily picked then it is also true to mitochondria with large perimeter. So, we can say Perimeter in this case is Area-biased. Hence, our candidate estimators for Perimeter mean are also not only Arithmetic Mean (AM), but also Weighted Mean (WM), Delta Method Estimator (DME) and 2nd Order Taylor's Approximation Estimator (2TAE), all of which consider the relationship between Area and Perimeter. For Circularity and Aspect Ratio, from Figure 1.0.1, we found that mitochondria with larger Area do not mean to have larger Circularity and Aspect Ratio. In other words, though the mitochondria with larger size are more easily picked, it will not influence the random sampling process of Circularity and Aspect Ratio. That is, the samples of Circularity and

Aspect Ratio can still be random samples under this size-biased sampling process. This case can also be explained as the following scenario. Suppose we have many different size of balls with number on them and the numbers are nothing to do with the size of balls. In this situation, even though the sampling probability of balls being picked is proportional to its size, the numbers on the chosen balls can still be treated as random numbers. Therefore, it is reasonable for us to use Arithmetic Mean (AM) as a reasonable estimator for the population means of Circularity and Aspect Ratio.

After figuring out the best estimator for each property, I used these estimators to do hypothesis tests. Because of the violation of ANOVA assumption of normality, Permutation Test was used for the hypothesis test and Bootstrapping technique was also conducted for confidence intervals of mean properties and mean difference with Bonferroni Correction. In the end, a suggestion on sampling procedure for future study on many other cells was provided based on the findings in the simulation study.

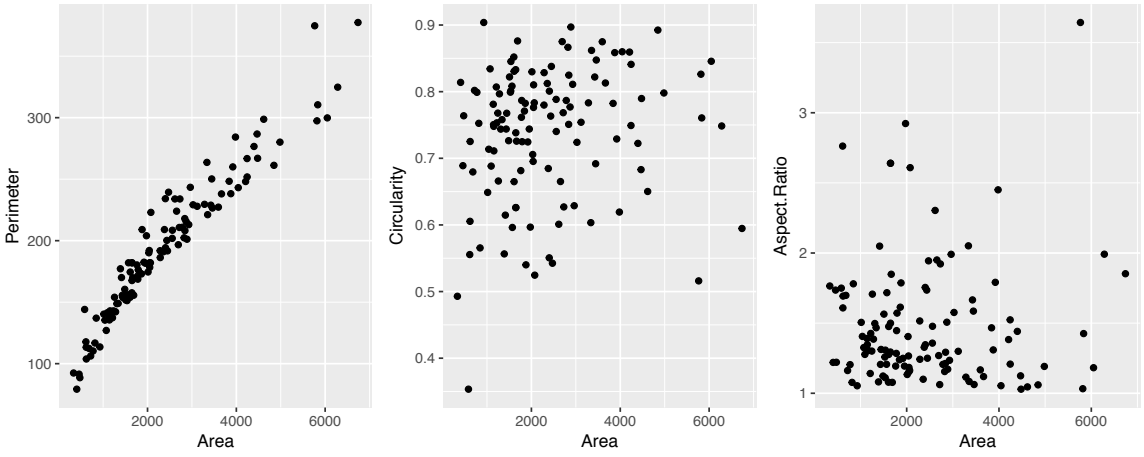

Figure 1.0.1: Scatter plots for Area vs. Perimeter, Area vs. Circularity and Area vs. Aspect Ratio

[Intentionally blank page]

2 Scientific Background

Mitochondria, also known as the "energy factories" of a cell, is a popular research topic in the fields of biology, chemical biology and biomedical engineering. One of the reasons for its popularity is that the mechanism of how mitochondria produce energy and its dysfunction have a strong relationship to the aging process, as described by Bratic and Larsson (2013).

The role of mitochondria in muscle contraction can provide some insight into cellular degeneration and aging. There are two ends on a single fiber cell, one is called a proximal point and the other is called a distal point. When a muscle contracts, the proximal point fixes at the same spot and the distal point is pulled to the proximal point. Given the different functions of these two ends, Professor Arriaga hypothesized that the required energy for muscle contraction varies within a single muscle fiber cell. Therefore, he proposed a hypothesis for this project: Properties (defined as Perimeter, Area, Circularity and Aspect Ratio) of mitochondria within a single fiber cell are significantly different by Locations (defined as Proximal end, Middle, or Distal end).

According to Prof. Arriaga, the result of the above hypothesis will be an important base for his future research on the aging process. As a muscle fiber cell ages, its mitochondria also degenerate which means their properties may change and may not be able to produce enough energy to facilitate muscle contractions as powerful as they once were. Hence, by observing the changes of mitochondria in muscle fiber cells over time, more ideas about the aging process may emerge.

3 Data

3.1 Sampling Procedure

A young mouse muscle fiber cell was picked and then magnified to 166 different images by using Transmission Electron Microscope (TEM) (Figure 3.1.1).

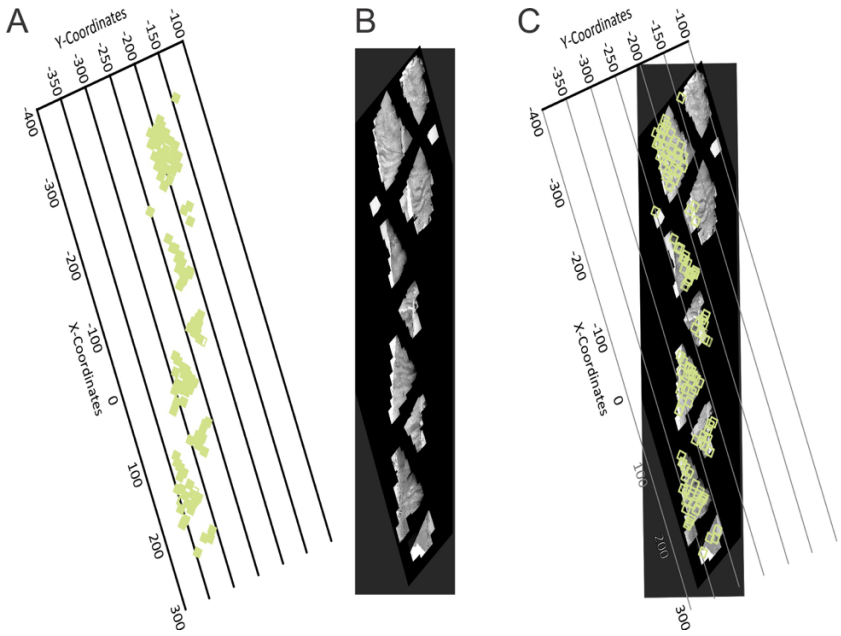

Figure 3.1.1: A young mouse muscle fiber cell. These graphs is about how the 166 different images were defined on coordinate.

For each location (Figure 3.1.2), images were divided into two groups, Subsarcolemmal group and Interfibrillar group. In each group, one image was randomly picked and 20 mitochondria were chosen from the image. Due to high costs of labor on doing simple random sampling, mitochondria were chosen by other sampling method. A list of random two-dimensional coordinates was gener-

ated and the mitochondria were chosen as sample if their area in the photo included one or more generated coordinates (Figure 3.1.3).

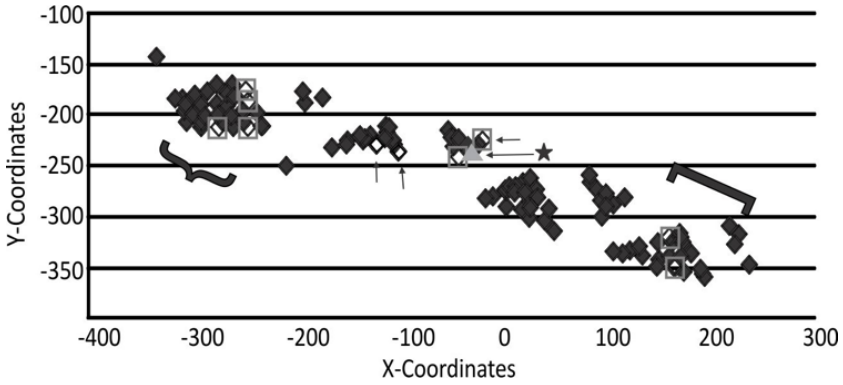

Figure 3.1.2: Definition of Locations. Those falls in " { " are defined as being in Proximal end, in " [" are being in Distal end, and the rest are being in Middle part.

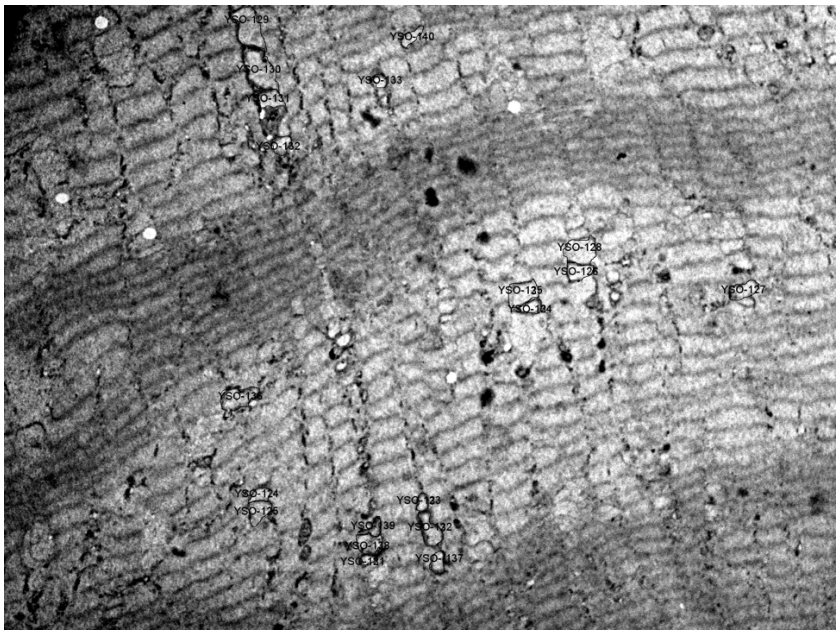

Figure 3.1.3: A sample of Mitochondria

3.2 Data Description

The following are the brief introduction about the attributions of each mitochondrion in this set of data.

About locations and groups

- Locations:

  The levels are Proximal end, Middle and Distal end.

- Groups:

  The levels are Subsarcolemmal and Interfibrillar groups.

- Image ID number:

  From which image the mitochondrion is selected.

- Mitochondrion ID number:

  The ID number of a mitochondrion in an image.

About the morphological properties

- Area (μm^2):

  The area occupied by a mitochondrion in an image.

- Perimeter (μm):

  The length of the boundary of a mitochondrion in an image.

- Circularity:

  Circularity is equal to $\frac{4\pi Area}{Perimeter^2}$. Measuring the resemblance of a mitochondrion to a circle.

  The range of circularity is between 0 and 1. 1 means a perfect circle.

- Aspect Ratio:

  Aspect Ratio is equal to $\frac{Length}{Width}$. If $AR \leq 2$, it is considered short; if $2 < AR \leq 4$, intermediate; if $AR > 4$, long.

3.3 Descriptive Statistics

To have a rough idea about how Properties of mitochondria differ over Locations in the set of data, the following four summary tables and eight figures provide information of descriptive statistics and distribution for each Property.

As can be seen in Table 3.3.1, Figure 3.3.1, Table 3.3.2, Figure 3.3.2, Table 3.3.3, Figure 3.3.3, the mitochondria at Distal part have the leftmost distributions of Area, Perimeter and Circularity; on the contrary, the ones at the Middle part have the rightmost distributions. This can be explained that generally Area, Perimeter and Circularity of mitochondria at Distal part are the smallest and the ones at the Middle part are the largest compared to the other two locations. Figure 3.3.4, Table 3.3.4 show that different from other Properties, the mitochondria in Middle part generally have the lowest Aspect Ratio than ones in Proximal and Distal part.

PMD	Mean	Sd	min	Q1	Median	Q3	Max
Proximal	2443.18	1309.09	845	1528.25	2056.5	3159.50	6740
Middle	2957.50	1335.37	817	1831.00	2812.0	3848.25	6049
Distal	1697.58	1262.37	335	760.50	1423.5	2042.50	6283

Table 3.3.1: Summary table for Area

PMD	Mean	Sd	min	Q1	Median	Q3	Max
Proximal	199.44	59.04	113.59	155.53	186.10	230.64	377.37
Middle	214.27	47.31	116.81	179.25	210.82	248.78	310.46
Distal	165.08	56.37	79.27	130.99	155.29	191.17	324.78

Table 3.3.2: Summary table for Perimeter

PMD	Mean	Sd	min	Q1	Median	Q3	Max
Proximal	0.74	0.11	0.52	0.67	0.78	0.82	0.90
Middle	0.77	0.08	0.55	0.73	0.78	0.83	0.90
Distal	0.70	0.10	0.35	0.64	0.73	0.77	0.84

Table 3.3.3: Summary table for Circularity

PMD	Mean	Sd	min	Q1	Median	Q3	Max
Proximal	1.53	0.52	1.05	1.16	1.35	1.73	3.64
Middle	1.37	0.36	1.03	1.14	1.27	1.48	2.64
Distal	1.53	0.41	1.10	1.24	1.40	1.72	2.92

Table 3.3.4: Summary table for Aspect Ratio

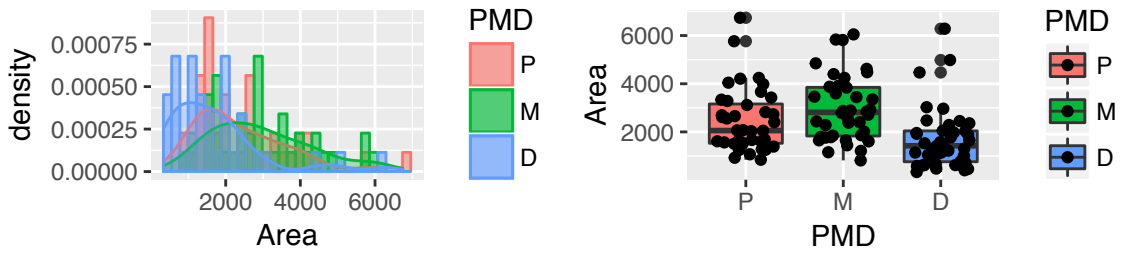

Figure 3.3.1: Histogram and Boxplot for Area by Locations

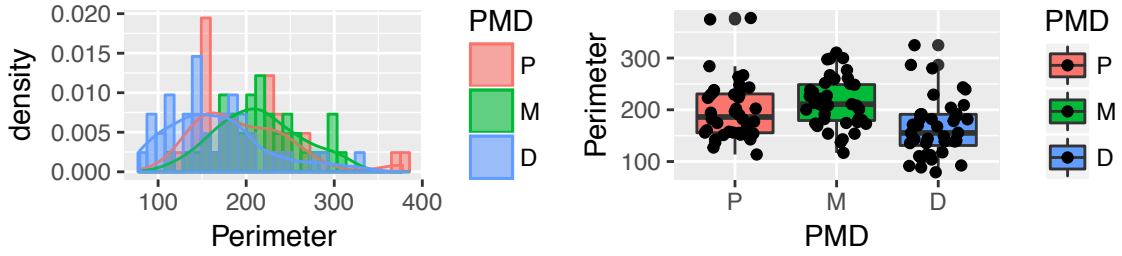

Figure 3.3.2: Histogram and Boxplot for Perimeter by Locations

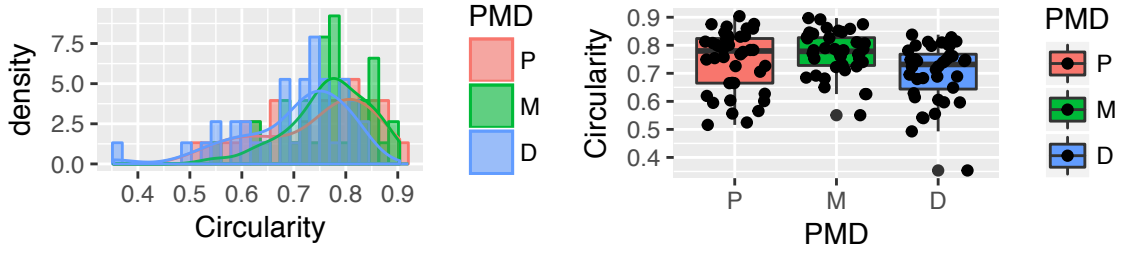

Figure 3.3.3: Histogram and Boxplot for Circularity by Locations

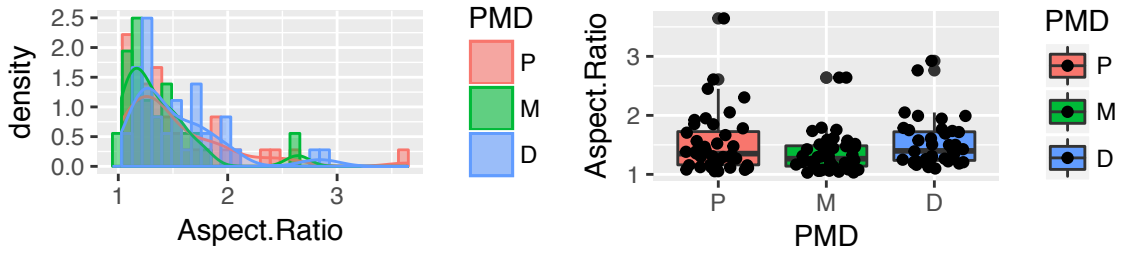

Figure 3.3.4: Histogram and Boxplot for Aspect Ratio by Locations

4 Method

Our goal is to find the best estimator for population mean Area (μ_A) and mean Perimeter (μ_P) of mitochondria. If our case was that every mitochondrion has equal probability of being picked as a sample in every experiment, then the sampling density of mitochondria Area (A) would be equal to its Probability Density Function (PDF), $f(a)$, and things would be simple: Arithmetic Mean ($\frac{\sum_{i=1}^{n} x_i}{n}$, where x_i is the value of the i_{th} observed data, n is the sample size.) is always the best estimator of population mean for its properties of unbiasedness and smaller variability as sample size (n) becomes larger. This is also true for Perimeter (P). However, the observed data we have in this project were sampled with Probability Proportional to Size (PPS). To deal with this situation, Cox (1962) proposed an idea of Weighted Distribution, defined as $f^*(x) = \frac{w(x)f(x)}{E_f(w(x))}$. In this formula, X is a random variable with PDF, $f(x)$; $w(x)$ is a weighted function decided by how the observed probability distribution proportional to x, and $f^*(x)$ is the Weighted Distribution of X, namely the observed probability distribution of X. For example, if the probability distribution of X is proportional to its size x, and then $w(x)$ will equal to x and its Weighted Distribution will follow $f^*(x) = \frac{xf(x)}{E_f(x)}$ which is not only proportional to the original PDF, $f(x)$, but also to its weighted function, $w(x) = x$. The $E_f(X)$ at the denominator is a constant for making the integration of Weighted Distribution be 1 and fitting the Probability Axioms ($\int f^*(a)da = \int \frac{af(a)}{E_f(A)}da = 1$).

Cox (1962) also proposed the Harmonic Mean ($\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$) as an estimator of population mean of X, and proved that it will converge to $\mu = E_f(x)$ as $n \to \infty$. The Harmonic Mean is equal to the Weighted Mean ($\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$, where $w_i = \frac{1}{p_i} = \frac{n\bar{x}}{x_i}$). To show the convergence of

Harmonic Mean, we first rearrange the definition of $f^*(x)$ as follows,

$$\begin{aligned} f^*(x) &= \frac{xf(x)}{E_f(X)} \\ &\Leftrightarrow E_f(X) \cdot \frac{1}{x} \cdot f^*(x) = f(x) \\ &\Leftrightarrow E_f(X) \cdot \int \frac{1}{x} \cdot f^*(x)dx = \int f(x)dx = 1 \quad \text{(By Probability Axioms)} \\ &\Leftrightarrow E_{f^*}(\frac{1}{X}) = \frac{1}{E_f(X)} \end{aligned}$$

then by the Weak Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} \underset{p}{\longrightarrow} E_{f^*}(\frac{1}{X}) = \frac{1}{E_f(X)} \quad \left(\text{Converge in Probability, } \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} - \frac{1}{E_f(X)}\right|\right) = 1\right)$$

$$\Leftrightarrow \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \underset{p}{\longrightarrow} E_f(X) = \mu \quad \text{(By Mapping Theorem)}$$

Then, we get that the Harmonic Mean, $\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$, will converge to $E_f(X) = \mu$ as $n \to \infty$.

In addition to the nonparametric estimators mentioned above, I was also interested in parametric estimators. To obtain parametric estimators, I assumed Area of mitochondria follows Exponential Distribution with mean equal to μ_A, $A \sim Exp(\mu_A)$. Under this distribution assumption, if every mitochondrion had equal probability of being picked as a sample in every experiment, then the sampling density of Area would also be Exponential Distribution, $f(a) = \frac{1}{\mu_A}e^{(-\frac{a}{\mu_A})} = Exp(\mu_A)$. However, as we know, in our case probability of each mitochondrion being picked as a sample is not equal to others but depends on its own Area. In this situation, the true probability distribution of Area would not be Exponential Distribution but Weighted Exponential Distribution, $f^*(a) = \frac{af(a)}{E_f(A)} = \frac{a\frac{1}{\mu_A}e^{(-\frac{a}{\mu_A})}}{\mu_A} = Gamma(2, \mu_A)$ with shape parameter equal to 2 and scale parameter equal to μ_A. By checking the histogram of Area, we found the distribution of Area did close to

$Gamma(2, 1183)$ (Figure 4.0.1). The multiplier, a, on the numerator makes the Weighted Distribution have the properties of being not only proportional to its original density, $f(a)$, but also to its Area, a. Under this distribution assumption of Area, we know the Maximum Likelihood Estimator (MLE) of μ_A will be $\frac{\sum_{i=1}^{n} a_i}{2n} = \frac{\bar{a}}{2}$ and the MLE of $Var(A)$ would be $\frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{2(n-1)} = \frac{\text{Sample Variance}}{2}$.

To obtain parametric estimators of population mean Perimeter (μ_P), I not only needed to assume Area follows $Exp(\mu_A)$ distribution, but also Circularity (C) follows $Beta(\alpha, \beta)$ distribution which is according to the histogram of Circularity (Figure 4.0.1). Then, by utilizing the facts that Area and Circularity are independent to each other and Perimeter has a relationship with Area and Circularity ($\text{P} = \sqrt{4\pi}\sqrt{\frac{\text{A}}{\text{C}}} = f(\text{A}, \text{C})$), we can obtain Delta Method Estimator(DME), $\widehat{\mu_P} = \sqrt{4\pi}\sqrt{\frac{\widehat{\mu_{A,MLE}}}{\widehat{\mu_{C,MLE}}}} = f(\widehat{\mu_{A,MLE}}, \widehat{\mu_{C,MLE}}) = f(\bar{A}/2, \bar{C}) = \sqrt{4\pi}\sqrt{\frac{\bar{A}/2}{\bar{C}}}$, and 2nd Order Taylor's Approximation Estimator of population mean Perimeter(μ_P).

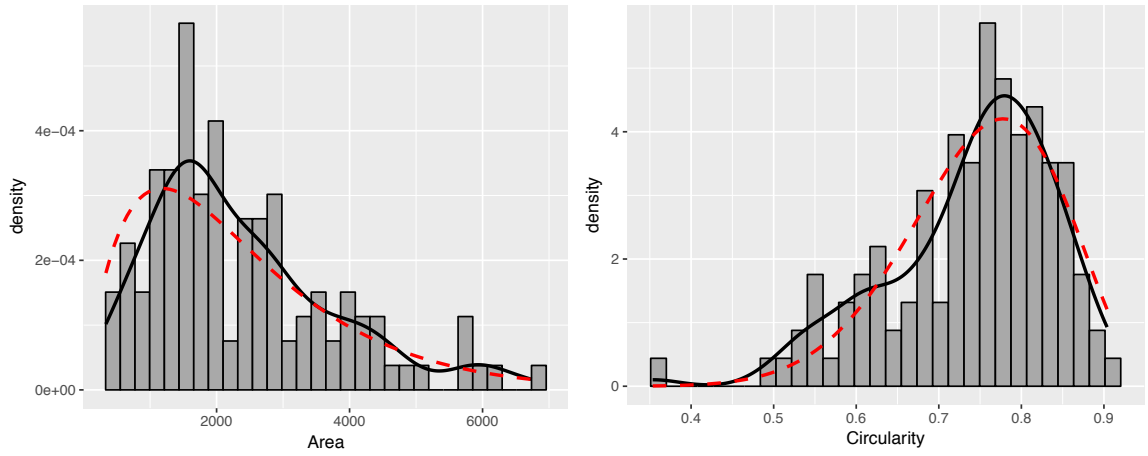

Figure 4.0.1: Histogram of Area and Circularity

The red dash line at the left if is $Gamma(2, 1183)$; the one at the right is $Beta(15, 5)$ distribution.

The following is a brief introduction about how to calculate 2nd Order Taylor's Approximation Estimator of μ_P.

By Taylor's series expanded on μ_A and μ_C for 2nd order,

$$\begin{aligned} E(P) &\doteq E(\sqrt{4\pi}\left[f(\mu_A,\mu_C)+\frac{1}{1!}\big(f_x(\mu_A,\mu_C)(x-\mu_A)+f_y(\mu_A,\mu_C)(y-\mu_C)\big)\right.\\ &\quad\left.+\frac{1}{2!}\big(f_{xx}(\mu_A,\mu_C)(x-\mu_A)^2+f_{yy}(\mu_A,\mu_C)(y-\mu_C)^2+2f_{xy}(\mu_A,\mu_C)(x-\mu_A)(y-\mu_C)\big)\right])\\ &\doteq \sqrt{4\pi}\left[\sqrt{\frac{\mu_A}{\mu_C}}-\frac{1}{8}(\mu_A)^{\frac{-3}{2}}(\mu_C)^{\frac{-1}{2}}Var(A)+\frac{3}{8}(\mu_A)^{\frac{1}{2}}(\mu_C)^{\frac{-5}{2}}Var(C)\right] \end{aligned}$$

Then, plug in the parameters with their MLEs,

$$\widehat{E(P)} \doteq \sqrt{4\pi}\left[\sqrt{\frac{\bar{A}/2}{\bar{C}}}-\frac{1}{8}(\frac{\bar{A}}{2})^{\frac{-3}{2}}(\bar{C})^{\frac{-1}{2}}\frac{s_A^2}{2}+\frac{3}{8}(\frac{\bar{A}}{2})^{\frac{1}{2}}(\bar{C})^{\frac{-5}{2}}s_C^2\right].$$

5 Simulation

5.1 Algorithm

To simulate the reality of how mitochondria were sampled, two-stage sampling was done in this project because each cell has finite mitochondria. For the first stage, I randomly chose N elements as subpopulation from $Exp(\mu)$ which is the assumed distribution of Area of mitochondria. These N elements were like the total mitochondria in a cell with value as their area. Therefore, our interested parameter was the mean Area of the subpopulation, μ_A, rather than μ. In the second stage, we sampled n elements from the subpopulation with sampling probability proportional to the area of elements. Though the observed data we have were from sampling without replacement, sampling with replacement was also done in this project in order to obtain more insights of the estimators' performance.

The Algorithm for simulating the sampling distribution of Area:

1. Set $N = 2000$; *Ratio* between N and n are $(5\%, 10\%, 30\%, 50\%, 70\%, 95\%)$; *Repeated Times* = 1000 and $\mu = 1000$.

2. Generate N samples from $Exp(\mu)$ as subpopulation of Area and calculate subpopulation mean, μ_A, as the known parameter.

3. Sample a set of samples with size n from subpopulation with sampling probability proportional to the value of Area with and without replacement. n is the product of N and a certain *Ratio*.

4. For each set of samples, calculate the candidate estimators: Arithmetic Mean (AM), Weighted Mean (WM) and Maximum Likelihood Estimator (MLE).

5. Repeat 3. 4. for the set *Repeated Times* for each *Ratio*.

6. Calculate the Mean, Standard Deviation and Root MSE for each candidate estimator. Also draw plots of sampling distributions for each candidate estimator.

The process on generating data of Perimeter of mitochondria was similar but based not only on the distribution of Area but also the distribution of Circularity of mitochondria and their independent relationship. In the first stage, randomly chose N elements of Circularity of mitochondria from $Beta(\alpha, \beta)$ and then substitute these N elements Circularity and the N elements of Area sitmulated before in the formula, Perimeter $= \sqrt{4\pi}\sqrt{\frac{\text{Area}}{\text{Circularity}}}$, to obtain N elements of Perimeter as subpopulation. Then, the mean of the N elements of Perimeter were our interested parameter, μ_P. After getting the subpopulation of Perimeter, n elements of Perimeter were chosen with sampling probability proportional to its corresponding Area. Again, sampling with and sampling without replacement both were considered in our simulation.

The Algorithm for simulating the sampling distribution of Perimeter:

1. Set $N = 2000$; *Ratio* between N and n are $(5\%, 10\%, 30\%, 50\%, 70\%, 95\%)$; *Repeated Times* $=$ 1000 and $\mu = 1000$.

2. Generate N samples from $Exp(\mu)$ distribution as subpopulation of Area and N samples from $Beta(\alpha, \beta)$ as subpopulation of Circularity. Assume the observed Circularity data we have are representative enough for the population of Circularity, and α and β are set to be 15 and 5 by observing the data we have.

3. Plug the generated N elements of Area and N elements of Circularity into the formula, Perimeter $= \sqrt{4\pi}\sqrt{\frac{\text{Area}}{\text{Circularity}}}$, and obtain N elements of Perimeter. Calculate the mean of N elements of Perimeter, μ_P, and treat it as the true mean of Perimeter.

4. Sample a set of samples with size n from subpopulation of Perimeter with sampling probability proportional to Area with and without replacement. n is the product of N and a certain *Ratio*.

5. For each set of samples, calculate the candidate estimators: Arithmetic Mean (AM), Weighted Mean (WM), Delta Method Estimator (DME), 2nd Order Taylor's Approximation Estimator (2TAE).

6. Repeat 3. 4. for the set *Repeated Times* for each *Ratio*.

7. Calculate the Mean, Standard Deviation and Root MSE for each candidate estimator. Also draw plots of sampling distributions for each candidate estimator.

5.2 Results for the simulation study

After repeating sampling with and without replacement, sampling distribution of each candidate estimator was constructed and their Biasedness, Standard Deviation and Root MSE were calculated. Figure 5.2.1, Table 5.2.1, Figure 5.2.2, and Table 5.2.2 shows that when sampling with replacement, MLE and 2TAE are always the best estimators for μ_A and μ_P and when sample size becomes larger, their performance on standard deviation (Std.) would also become better. As for the performance of AM, like our expectation, severely overestimates the true mean no matter how sample size changes. Conversly, when it is sampling without replacement, the performance of MLE and 2TAE estimators are not worse only when sample size compared to population size is small since when sample size is small, sampling without replacement is similar to sampling with replacement. For AM, even though its expected value approximates to the true mean as sample size becomes larger, it is still a biased estimator if the sample size is not equal to the population size.

To sum up, in this simulation, when it is sampling with replacement, MLE and the 2TAE perform best and AM has the worst performance. Nevertheless, MLE and 2TAE can only be obtained under strong assumptions on distributions of Area and Circularity. Hence, we also include WM in our following distribution for its sufficient performance and its nonparametric assumption.

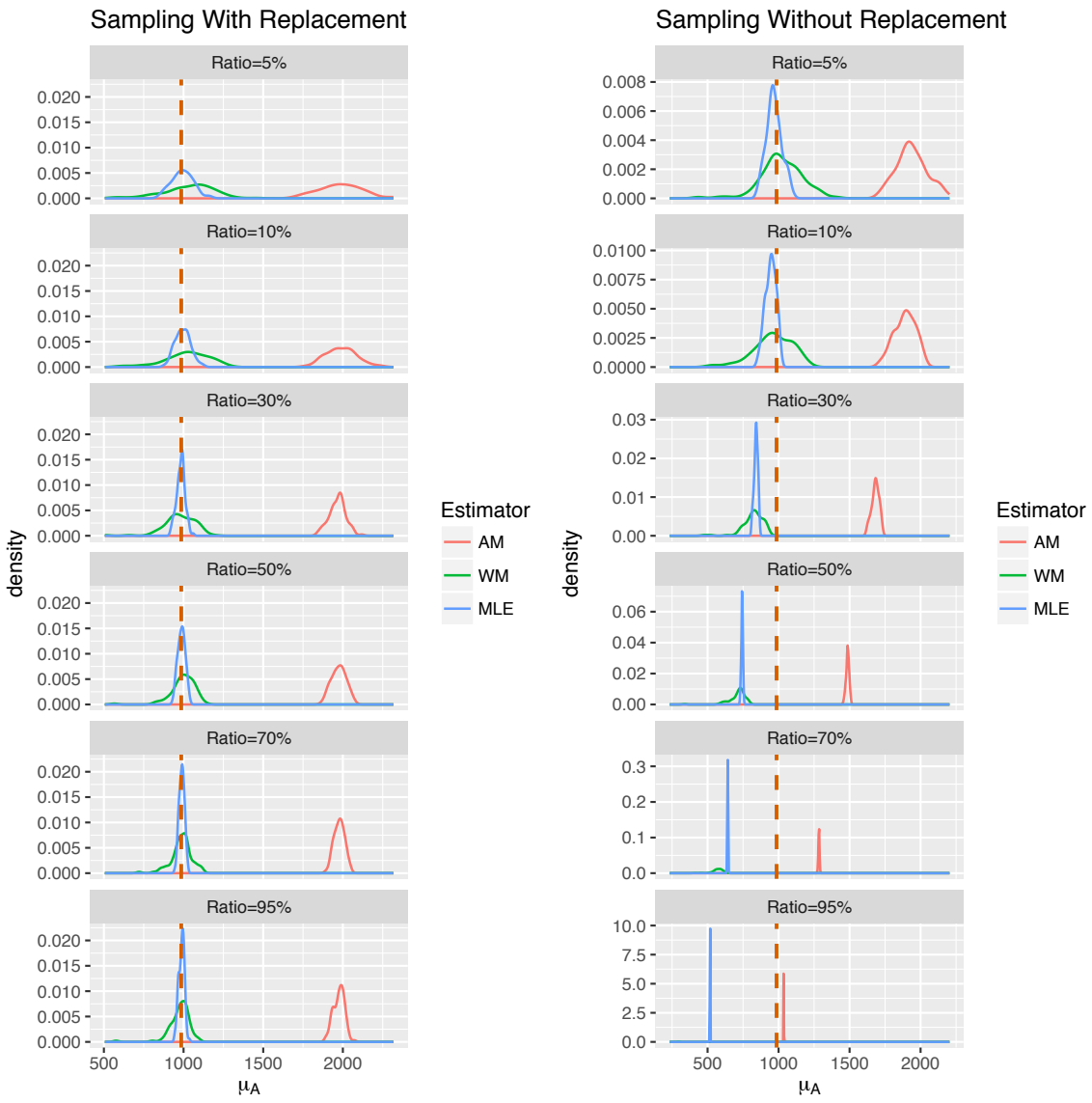

Figure 5.2.1: Sampling Distributions for Area with Different Ratios, N=2000, μ_A=985.75

Sampling WITH Replacement					Sampling WITHOUT Replacement				
Ratio	Estimator	Bias	Std. Dev.	Root MSE	Ratio	Estimator	Bias	Std. Dev.	Root MSE
5%	AM	1000	133.9	1000	5%	AM	951.12	106.95	951.18
	WM	45.2	162	47		WM	33.71	150.39	35.87
	MLE	7.1	67	10.8		MLE	-17.31	53.47	18.79
10%	AM	1003.9	96.3	1003.9	10%	AM	903.12	75.07	903.17
	WM	35	140.4	37		WM	-25.54	137.37	28.10
	MLE	9.1	48.1	11.4		MLE	-41.31	37.53	41.76
30%	AM	980.2	52.6	980.2	30%	AM	696.02	26.21	696.04
	WM	-8.7	100.6	13.2		WM	-163.47	74.31	163.70
	MLE	-2.8	26.3	5.8		MLE	-144.87	13.10	144.91
50%	AM	988.4	46	988.5	50%	AM	500.30	10.95	500.31
	WM	11.2	78	14.3		WM	-274.76	60.98	274.87
	MLE	1.3	23	5		MLE	-242.73	5.47	242.74
70%	AM	991.7	33.2	991.8	70%	AM	298.43	3.52	298.43
	WM	0.9	65.1	8.1		WM	-416.20	41.13	416.25
	MLE	3	16.6	5.1		MLE	-343.66	1.76	343.66
95%	AM	989.2	34.6	989.2	95%	AM	49.93	0.17	49.93
	WM	-8.9	65	12		WM	-697.56	18.23	697.58
	MLE	1.7	17.3	4.5		MLE	-467.91	0.09	467.91

Table 5.2.1: Performance Table for Area

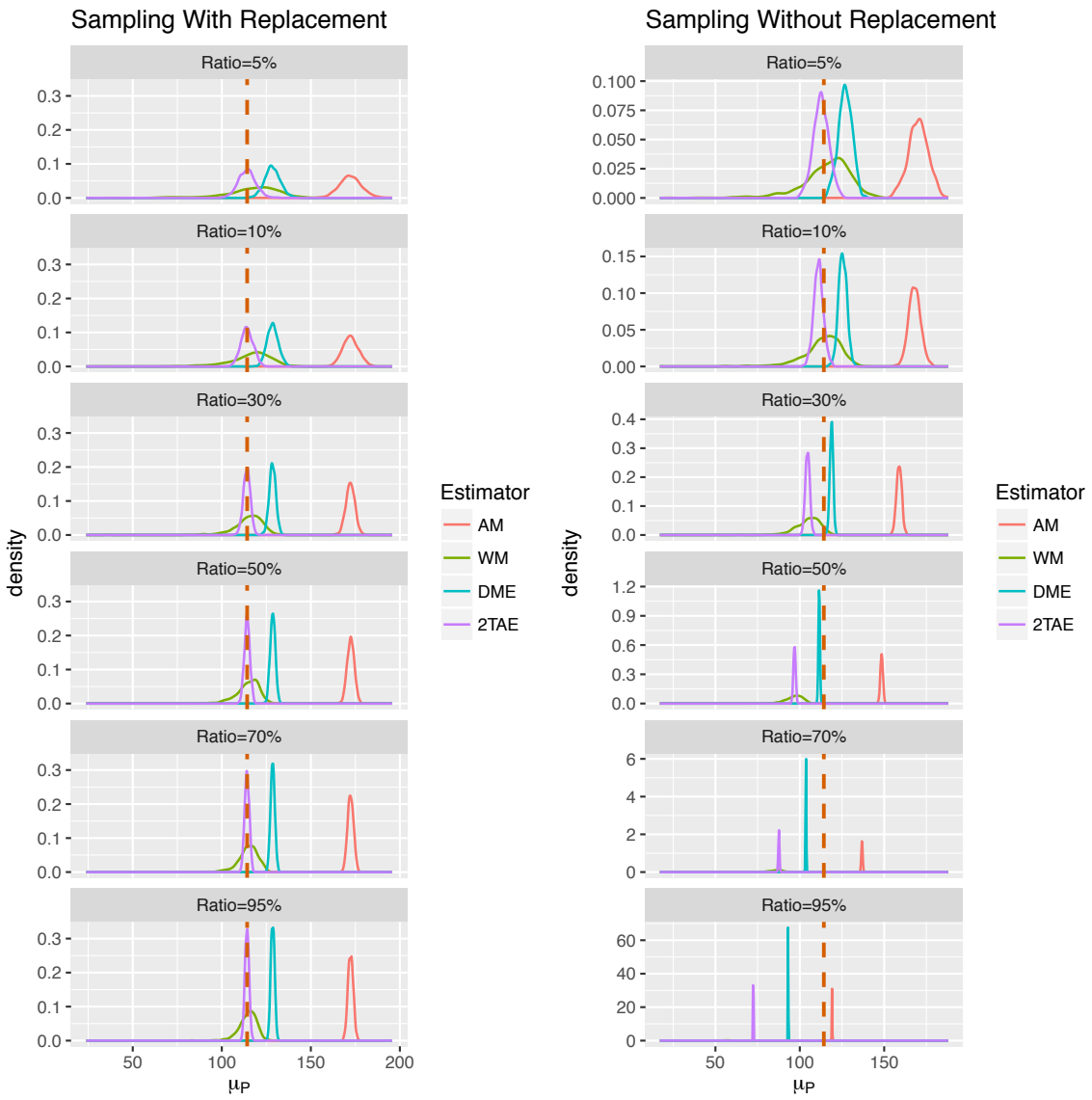

Figure 5.2.2: Sampling Distributions for Perimeter with Different Ratios, N=2000, μ_P=114.24

Sampling WITH Replacement					Sampling WITHOUT Replacement				
Ratio	Estimator	Bias	Std. Dev.	Root MSE	Ratio	Estimator	Bias	Std. Dev.	Root MSE
5%	AM	58.05	6.30	58.11	5%	AM	55.95	5.80	56.01
	WM	2.93	15.81	4.94		WM	1.64	14.76	4.18
	DME	14.23	4.43	14.38		DME	12.70	4.07	12.86
	2TAE	-0.10	4.91	2.22		2TAE	-1.63	4.57	2.69
10%	AM	58.12	4.42	58.16	10%	AM	53.79	3.58	53.82
	WM	1.87	12.25	3.97		WM	-0.87	11.71	3.53
	DME	14.34	3.15	14.45		DME	11.17	2.47	11.28
	2TAE	-0.07	3.44	1.85		2TAE	-3.15	2.85	3.58
30%	AM	57.94	2.59	57.96	30%	AM	44.61	1.56	44.62
	WM	0.72	8.82	3.06		WM	-9.35	7.53	9.75
	DME	14.24	1.86	14.31		DME	4.55	0.98	4.65
	2TAE	-0.22	2.00	1.43		2TAE	-9.66	1.31	9.73
50%	AM	58.17	2.02	58.19	50%	AM	34.24	0.76	34.25
	WM	0.42	6.32	2.55		WM	-17.78	5.69	17.94
	DME	14.39	1.45	14.44		DME	-2.76	0.43	2.84
	2TAE	-0.04	1.56	1.25		2TAE	-17.38	0.66	17.39
70%	AM	58.09	1.65	58.10	70%	AM	22.60	0.35	22.61
	WM	0.42	5.90	2.46		WM	-28.71	4.66	28.79
	DME	14.33	1.18	14.37		DME	-10.61	0.17	10.62
	2TAE	-0.10	1.29	1.14		2TAE	-26.65	0.31	26.65
95%	AM	58.09	1.47	58.11	95%	AM	4.85	0.05	4.85
	WM	0.38	5.18	2.31		WM	-58.41	2.85	58.44
	DME	14.33	1.05	14.36		DME	-21.19	0.02	21.19
	2TAE	-0.09	1.13	1.07		2TAE	-41.83	0.03	41.83

Table 5.2.2: Performance Table for Perimeter

6 Analysis

With the goal to figure out whether Properties (Area, Perimeter, Circularity and Aspect Ratio) of mitochondria are different by Locations (Proximal end, Middle and Distal end) in a single muscle fiber cell from a young mouse, we set our null hypothesis (H_0) to be that the mean Area, Perimeter, Circularity and Aspect Ratio of mitochondria are equal by Locations and alternative hypothesis (H_A) to be that the means are not all equal. Moreover, I also conducted pairwise comparison tests to find out the difference between Locations. In the overall and pairwise comparison hypothesis tests, MLE and Weighted Mean were used as the estimators of mean Area for the whole mitochondria and for each location; 2nd Order Taylor's Approximation Estimator and Weighted Mean were for Perimeter and Arithmetic Mean was for Circularity and Aspect Ratio for their independence to size of Area.

Overall Hypothesis Test:

$$\begin{cases} H_0 : \mu_{i_P} = \mu_{i_M} = \mu_{i_D},\ i = \{\text{Area, Perimeter, Circularity, Aspect Ratio}\} \\ H_A : \text{At least one } \mu_{i_j} \neq \mu_{i_k},\ j, k = \{\text{P, M, D}\} \end{cases}$$

Pairwise Comparison Test:

$$\begin{cases} H_0 : \mu_{i_j} = \mu_{i_k},\ i = \{\text{Area, Perimeter, Circularity, Aspect Ratio}\}\ ;\ j, k = \{\text{P, M, D}\} \\ H_A : \mu_{i_j} \neq \mu_{i_k} \end{cases}$$

Instead of choosing the standard ANOVA (F-test) and T-test as our overall and pairwise comparison test methods, we decided to use a Permutation method. The reasons are that the

estimator of population mean in ANOVA and T-test is sample mean but it is not appropriate to Area and Perimeter because of their size-biased samples. And for Circularity and Aspect Ratio, the data violated the normality assumption of ANOVA and T-test (Figure 3.3.3 and Figure 3.3.4). Hence, the statistics in our case for the overall test was $\sum_{i=\{P,M,D\}} (\widehat{\mu}_i - \widehat{\mu})^2$ and for the pairwise comparison test was $\widehat{\mu}_i - \widehat{\mu}_j$, where $i = \{P, M, D\}$. Their sampling distributions were constructed by assuming the null hypothesis is true, treating the observed data as population, finding out all the possible combinations of elements in groups and then calculating the statistics for every combination. However, it was not efficient for us to calculate all the combinations so we randomly drew large amount of the combinations from the complete set to obtain an approximate sampling distribution. Then, the approximate P-value was the probability for the statistics more extreme than the observed one. After finished all the hypothesis test, we used Bootstrap technique with Bonferroni's correction to obtain confidence interval of true difference of properties means by Locations.

Bonferroni's correction is a method to control Strong Familywise Error Rate in multi-comparison hypothesis tests and to have simultaneous confidence interval for the mean differences. The significance level for each test are defined as α/m, where m is the number of test in this multi-comparison test. So, the simultaneous confidence interval for the mean differences will be $(1 - \frac{\alpha}{m})\%$.

7 Results

7.1 Hypothesis Tests

We found that the difference in Area, Perimeter, Circularity between locations is statistically significant at 0.05 significance level, as seen in Table 7.1.1 regardless of which estimators. For the Aspect Ratio, there is not a statistically significant difference between locations.

Also, we did pairwise comparison tests by using Bonferroni's Correction to correct the significance level of each paired comparison to be $\frac{\text{significance level for Overall Hypothesis Test}}{\text{number of pairwise comparison}} = 0.05/3 \doteq 0.0167$. Hence, as we can be seen in Table 7.1.1, we have enough evidence to say mean Area and Perimeter of mitochondria is significantly different between Middle and Distal end and between Proximal and Distal end. The results from the parametric estimators (MLE and 2TAE) is consistent to the non-parametric estimator (Weighted mean) in this case. For Circularity, we only have enough evidence to reject the null hypothesis that mean Circularity of mitochondria at Middle is equal to at Distal.

Property	Estimator	Overall	P vs. M	M vs. D	P vs. D
Area	WM	<0.0001	0.0974	<0.0001	0.0022
	MLE	0.0001	0.0950	0.0002	0.0140
Perimeter	WM	0.0001	0.2744	<0.0001	0.0018
	2TAE	<0.0001	0.1518	<0.0001	0.0024
Circularity	AM	0.0070	0.2476	0.0022	0.0616
Aspect Ratio	AM	0.1838	0.1046	0.1102	0.9884

Table 7.1.1: Unadjusted p-values from Overall and Pairwise Comparison Hypothesis Tests. The significance level for Overall Hypothesis Test is 0.05 and the significance level for Pairwise Hypothesis Test with the Bonferroni correction 0.0167.

7.2 Confidence Intervals

By using Bootstrapping, we found the 95% confidence interval of the mean by different Locations (Table 7.2.1). Also, the 98.33% simultaneous confidence interval of mean differences for each pair which was shown in Table 7.2.2. The value 98.33% is because of the corrected significant value of 0.0167 from Bonferroni's Correction.

Property	Estimator	Location	Estimate	Lower Bound	Upper Bound
Area	WM	Proximal	1928.23	1680.97	2243.38
		Middle	2381.80	2038.77	2791.91
		Distal	1071.75	873.02	1350.56
	MLE	Proximal	1221.59	1032.41	1426.77
		Middle	1478.75	1276.53	1686.79
		Distal	848.79	669.24	1055.76
Perimeter	WM	Proximal	176.56	164.50	191.47
		Middle	192.90	177.86	209.65
		Distal	133.49	121.05	150.04
	2TAE	Proximal	134.44	123.37	147.17
		Middle	147.95	137.35	158.80
		Distal	107.26	95.12	120.94
Circularity	AM	Proximal	0.744	0.711	0.776
		Middle	0.771	0.747	0.794
		Distal	0.702	0.668	0.732
Aspect Ratio	AM	Proximal	1.529	1.384	1.705
		Middle	1.369	1.269	1.490
		Distal	1.525	1.409	1.657

Table 7.2.1: Estimate for all mean Properties by Locations and their 95% Confidence Interval.

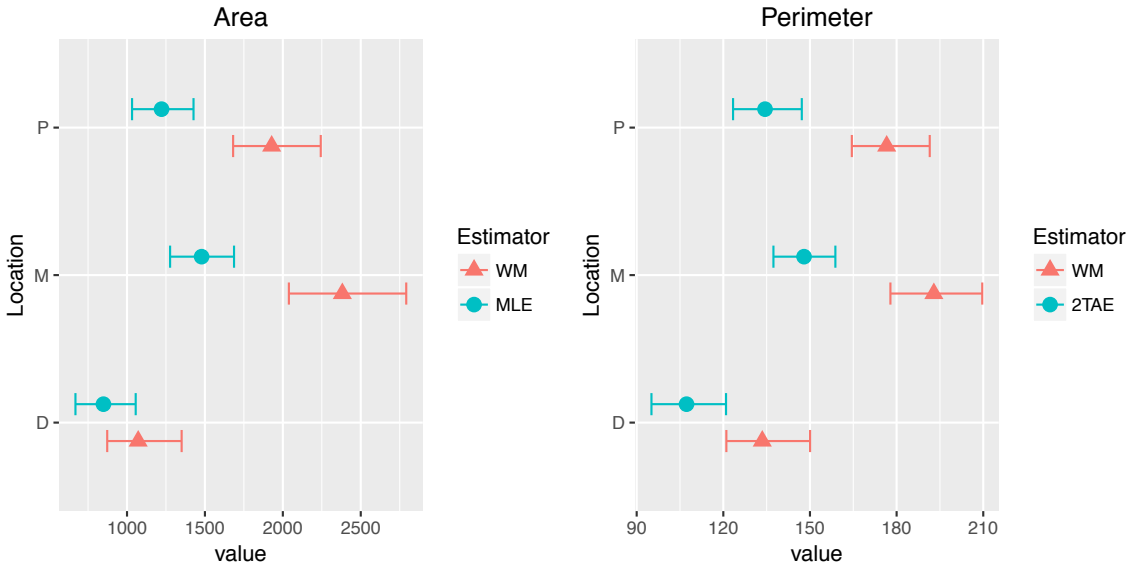

Figure 7.2.1: 95% Confidence Interval for Population Mean Area and Perimeter by Locations.

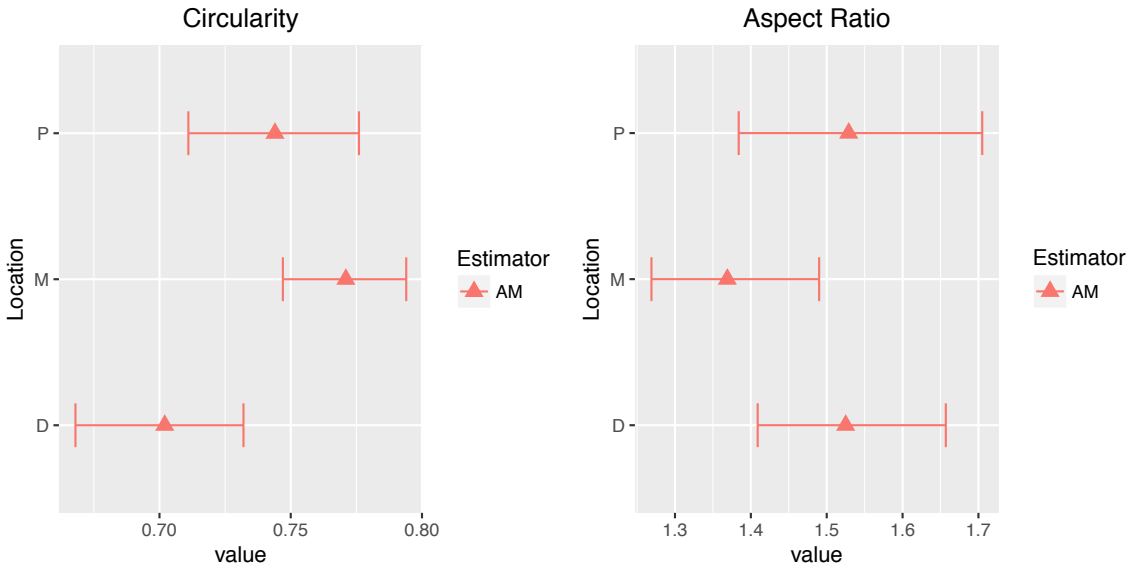

Figure 7.2.2: 95% Confidence Interval for Population Mean Circularity and Aspect Ratio by Locations.

Property	Estimator	Comparison	Difference	Lower Bound	Upper Bound
Area	WM	P vs. M	-453.57	-1036.89	120.93
		M vs. D	1310.06	767.22	1869.39
		P vs. D	856.48	400.39	1313.66
	MLE	P vs. M	-257.16	-602.74	104.31
		M vs. D	629.96	274.51	961.36
		P vs. D	372.80	27.93	711.49
Perimeter	WM	P vs. M	-16.33	-41.63	10.16
		M vs. D	59.41	32.03	84.53
		P vs. D	43.07	17.92	66.72
	2TAE	P vs. M	-13.51	-32.27	6.95
		M vs. D	40.69	19.23	60.70
		P vs. D	27.18	6.09	48.09
Circularity	AM	P vs. M	-0.026	-0.077	0.023
		M vs. D	0.069	0.020	0.119
		P vs. D	0.042	-0.014	0.099
Aspect Ratio	AM	P vs. M	0.160	-0.077	0.408
		M vs. D	-0.156	-0.360	0.044
		P vs. D	0.004	-0.238	0.261

Table 7.2.2: Estimate for mean Difference and their 98.33% Simultaneous Confidence Interval.

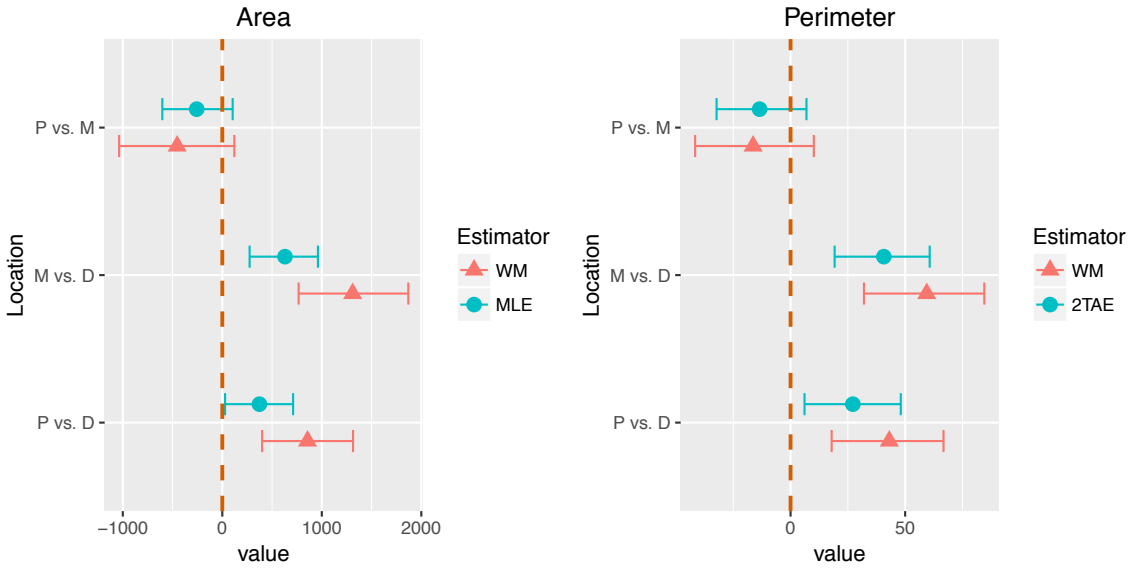

Figure 7.2.3: 98.33% Confidence Interval for mean Difference by Locations for Area and Perimeter.

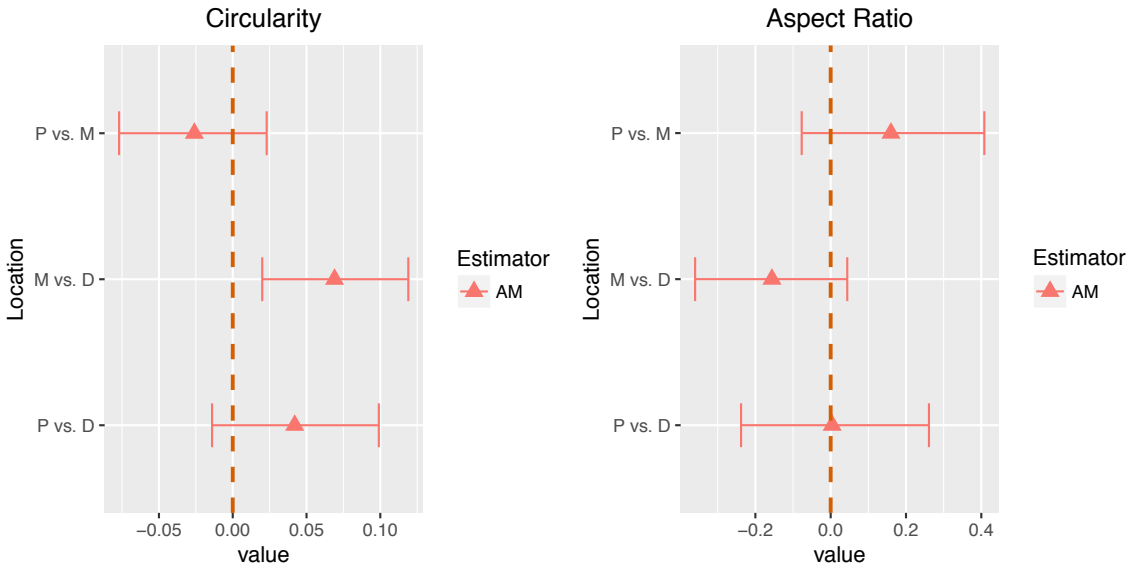

Figure 7.2.4: 98.33% Confidence Interval for mean Difference by Locations for Circularity and Aspect Ratio.

8 Conclusion and Discussion

The results of our analysis showed that generally mitochondria located in Middle of the muscle fiber cell have larger Area, Perimeter and Circularity which means to support muscle contraction more energy is needed in Middle. Conversely, less energy is needed at Distal end for the significantly smallest Area, Perimeter and Circularity.

If this analysis will be applied on more muscle fiber cells in the future, I would recommend them to use Nonparametric Weighted Mean as the best estimator for population mean and do hypothesis test based on this estimator. The reasons are for its none distribution assumptions, simple expression and interpretability. Besides, I would also suggest them to use Sampling With Replacement (SWR) rather than Sampling Without Replacement (SWOR) in their sampling scheme because as we can see in the Simulation section the performance of Weighted Mean is not desirable when the case is SWOR unless they can assure the Ratio between population and samples are around 10% or less. Finding the best estimator for SWOR is a potential area for future work.

Based on the result of simulation study, we expect Nonparametric Weighted Mean should have similar results with the Parametric Estimators (MLE for Area and 2TAE for Perimeter) but wider confidence interval for the Nonparametric Weighted Mean. However, as we see in Figure 7.2.1 and Table 7.2.1, in our data things are not like what we expected. One of the reason might be the improper distribution assumptions on Area and Circularity. Hence, in the future the robustness of the distribution assumptions can be an interesting topic to work on too.

References

[1] Bratic, Ana and Larsson, Nils-Gran. "The Role of Mitochondria in Aging." *Journal of Clinical Investigation* 123, no. 3 (2013): 951-57. doi:10.1172/jci64125.

[2] Cox, D. R. *Renewal Theory.* London: Methuen & Co., 1962.

[3] Patil,G. P. and Ord,J. K. "On Size-Biased Sampling and Related Form-Invariant Weighted Distributions." *Sankhya: The Indian Journal of Statistics, Series B (1960-2002)* 38, no. 1 (1976): 48-61.

[4] Jones, M. C. "Kernel Density Estimation for Length Biased Data." *Biometrika.* 78, no. 3 (1991): 511-19. doi:10.2307/2337020.