Analysis of Size-Biased Mitochondria Data

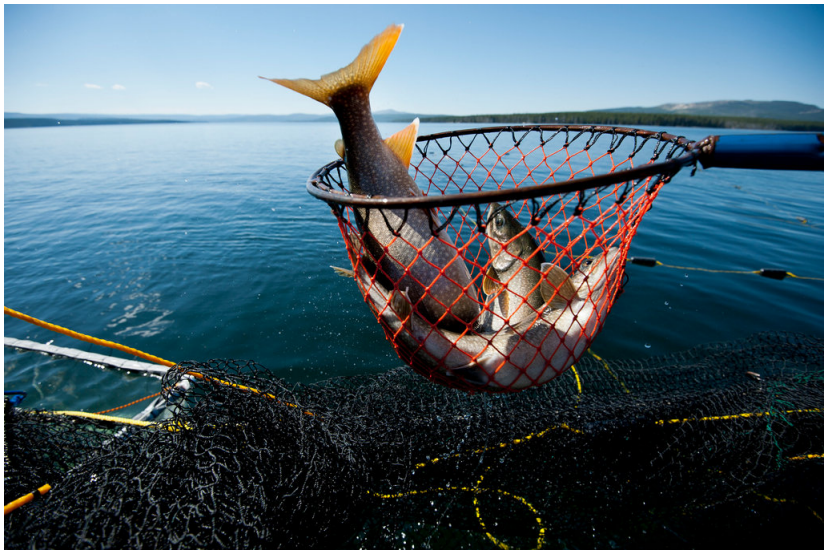
Students: Yin-Ting Chou
Advisor: Aaron Rendahl

5/17/2017

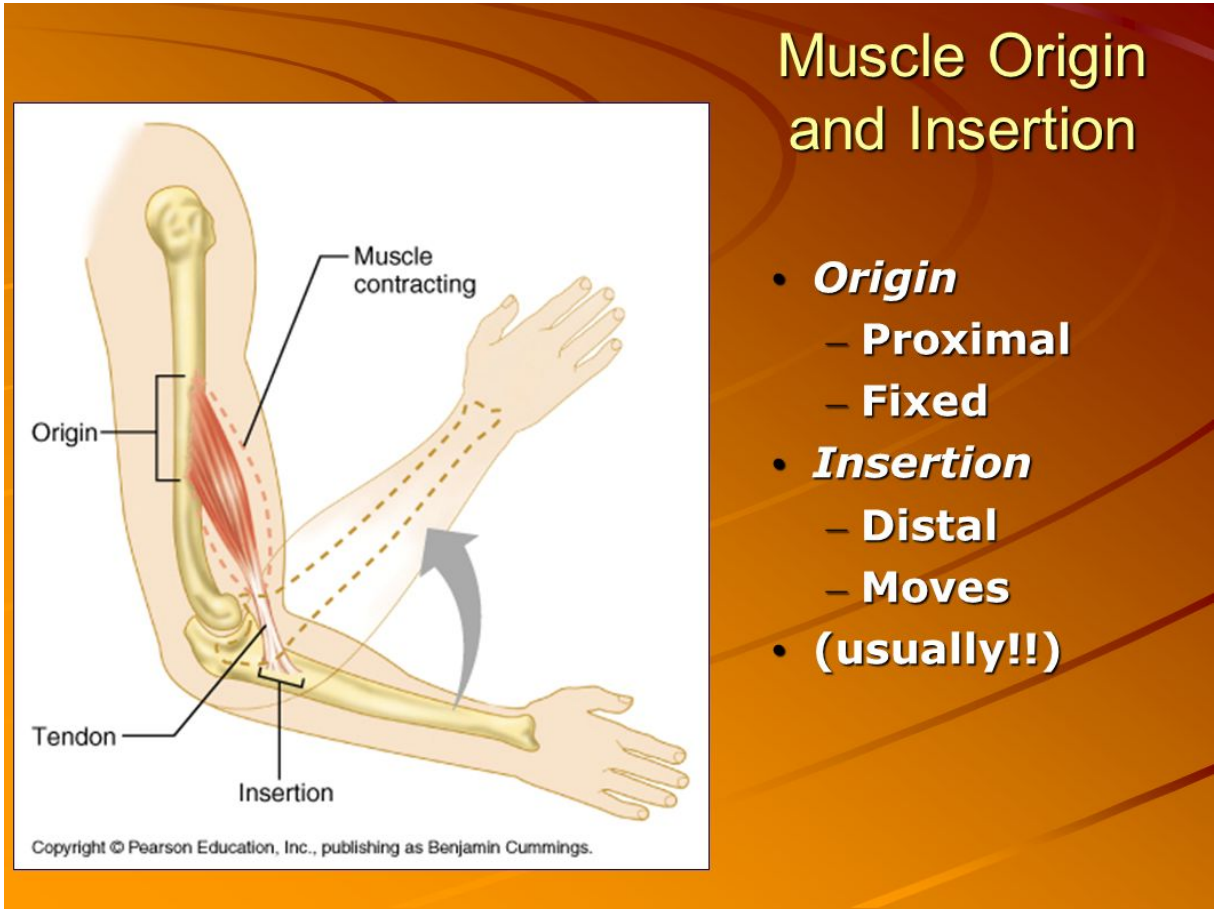
Overview

1. What is Size-Biased Data?
2. Scientific Background for Mitochondria
3. Goals for this project
4. How the sampling process caused size-biased data?
5. Investigate Possible Estimators with simulation study
6. Use the best ones on real data
7. Conclusion
8. Future Works

Examples for Size-Biased Data

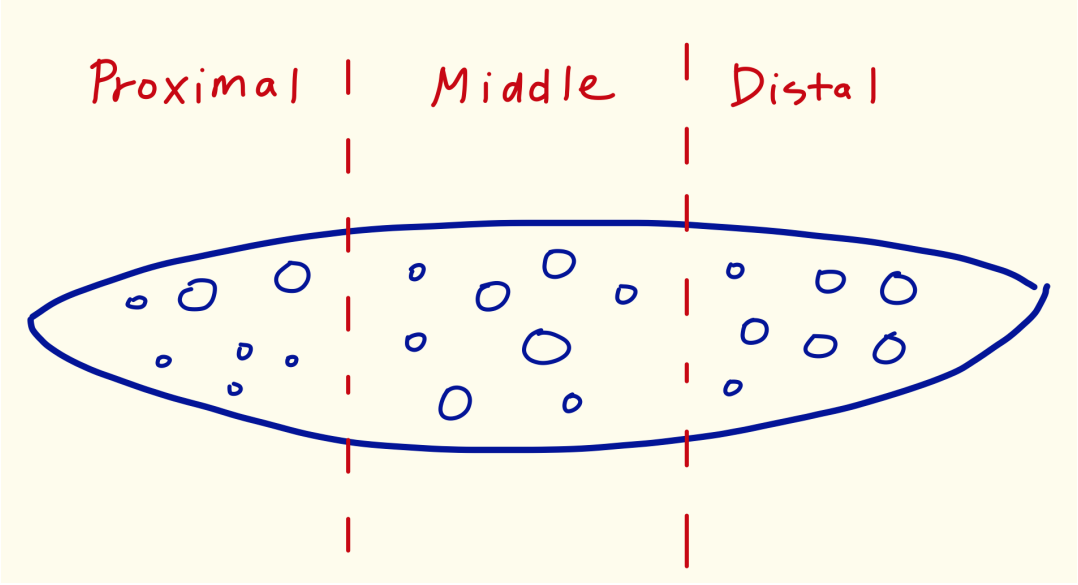


Scientific Background for Mitochondria



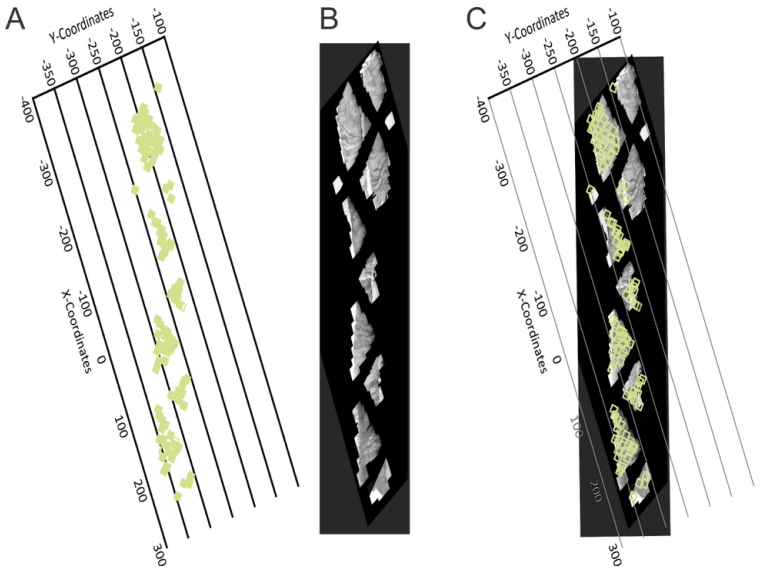
Goals for this project

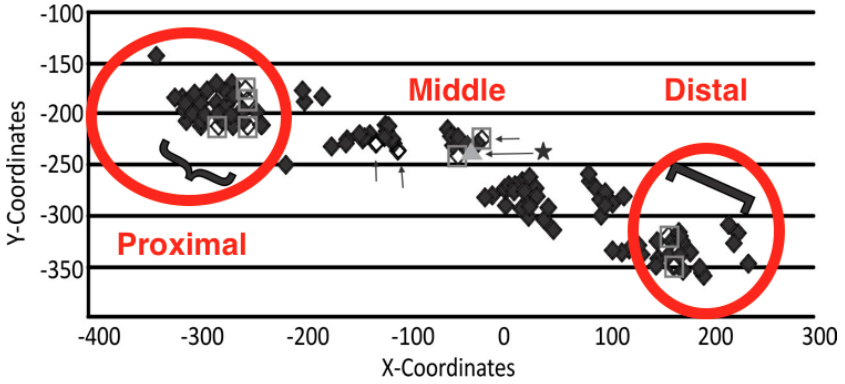
1. Whether **Properties** (area, perimeter, circularity and aspect ratio) of mitochondria are different by **locations** (proximal, middle and distal end).
2. Suggestions on **sampling method** for future research (more cells).



Sampling Process - 1

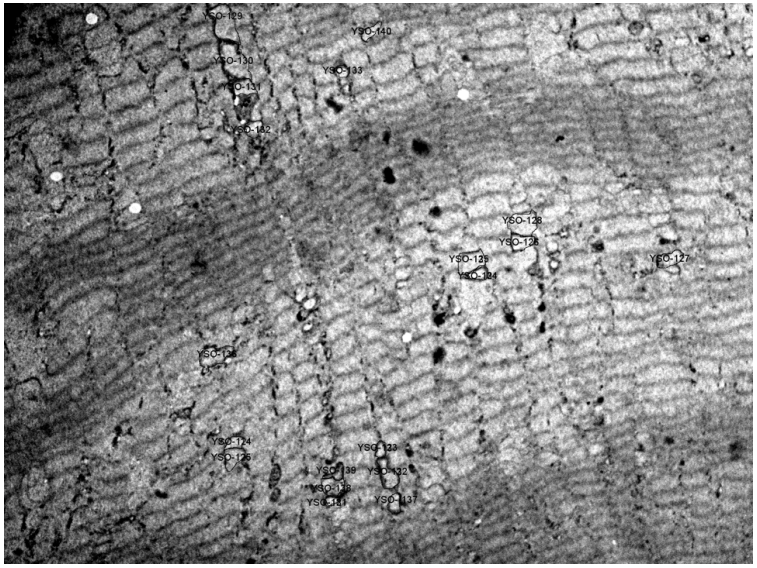
- A young muscle fiber cell was magnifired to 166 different images by using Transmission Electron Microscope (TEM).





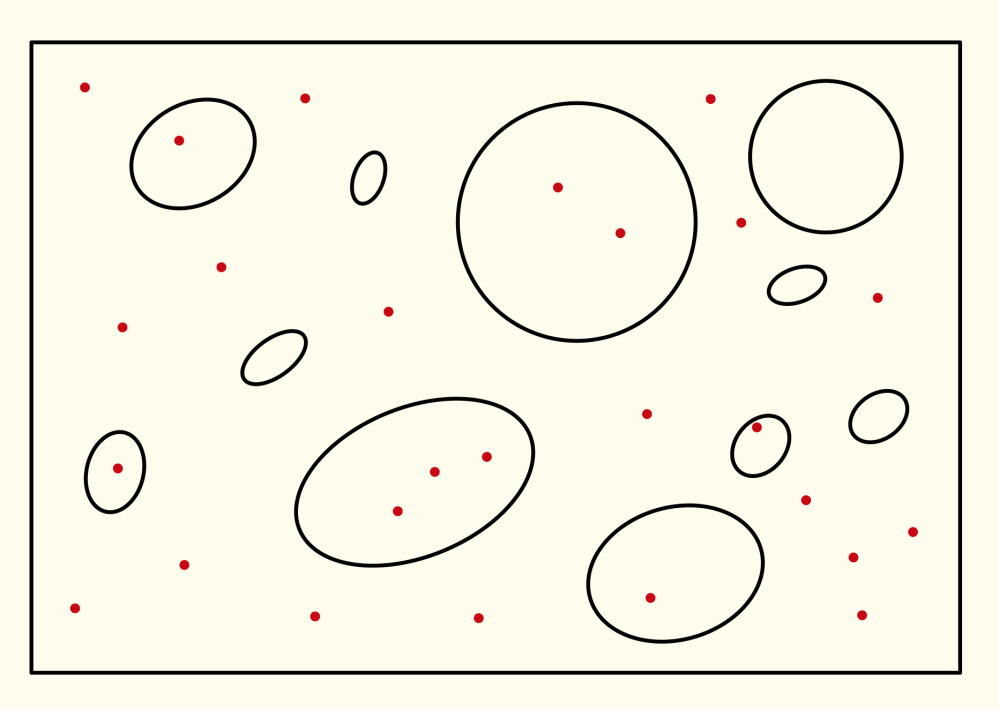
Sampling Process - 2

- For each location, divide images into two groups: Subsarcolemmanl and Interfibrillar group (ignore later).
- In each group, randomly pick one image.
- In each image, sample **20 mitochondria**.



Sampling Process - 3

- Generate a list of random coordinates.
- Pick the mitochondria whose area in the photo includes one or more generated coordinates.



Problems from the Sampled Data

1. It is **NOT** random sample but size-biased!
2. The larger mitochondria are easier to be picked in our sample.
3. If we used sample mean as the estimator of population mean, it will definitely be **overestimated**!

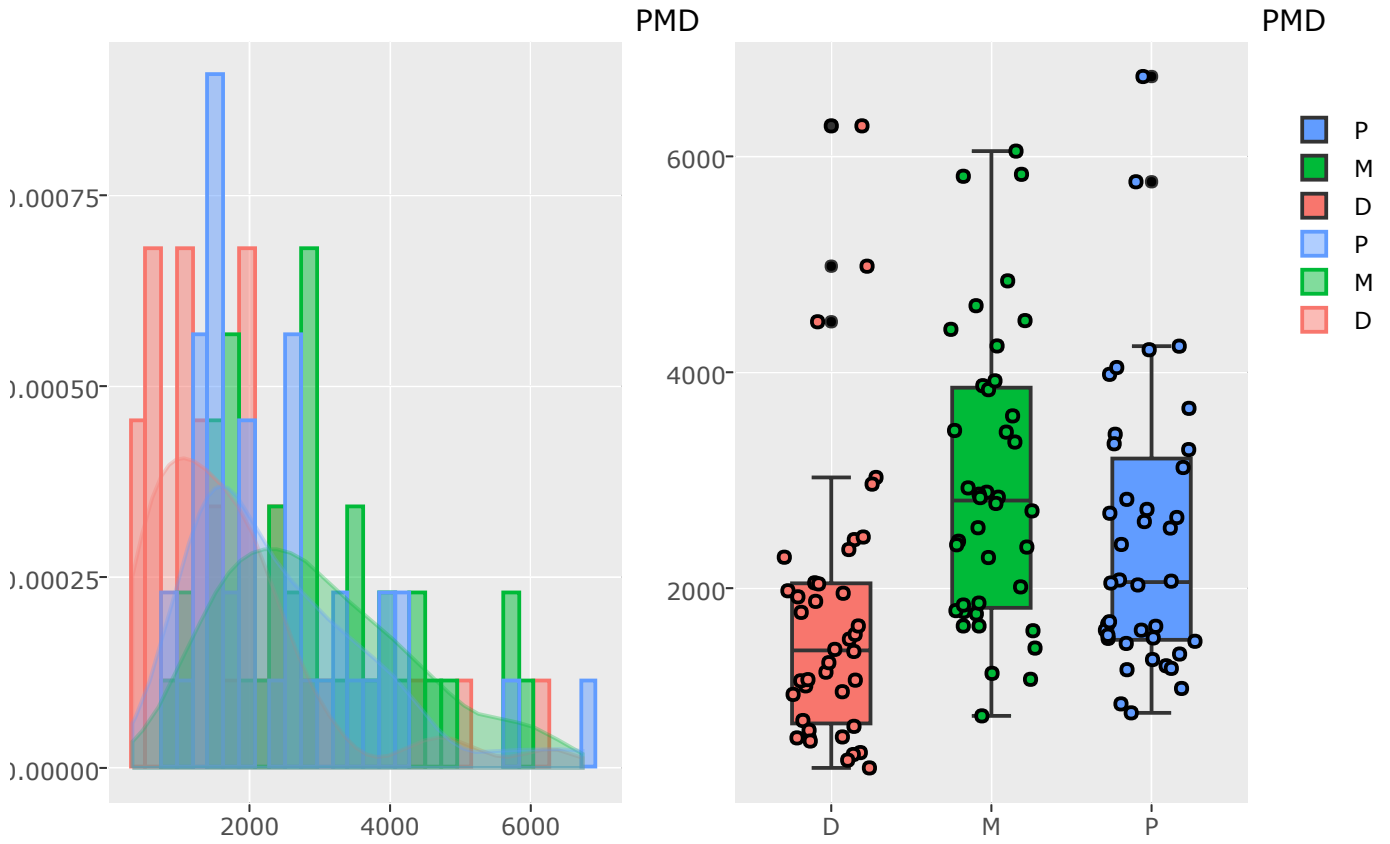
Raw Data

Sample.ID	PMD	SI	Image.No.	Mito.Number.in.Image	Area	Perimeter	Circularity	Aspect.Ratio
YSO-021	M	S	2	1	6049	299.8184	0.8456	1.1821
YSO-022	M	S	2	2	4244	251.818	0.841	1.2079
YSO-023	M	S	2	3	2889	201.1797	0.897	1.1713
YSO-024	M	S	2	4	3461	226.5416	0.8475	1.0624
YSO-025	M	S	2	5	3876	238.18	0.8586	1.3091
YSO-026	M	S	2	6	4847	261.2556	0.8924	1.0598
YSO-027	M	S	2	7	1606	153.9064	0.852	1.0789
YSO-028	M	S	2	8	1445	153.8016	0.7676	1.3125
YSO-029	M	S	2	9	1782	168.7021	0.7868	1.4466
YSO-030	M	S	2	10	3596	227.2638	0.8749	1.1667
YSO-031	M	S	2	11	2843	208.1268	0.8248	1.2926
YSO-032	M	S	2	12	3354	221.1328	0.8619	1.0841
YSO-033	M	S	2	13	4479	266.9746	0.7897	1.0292
YSO-034	M	S	2	14	2435	200.2251	0.7633	1.7356
YSO-035	M	S	2	15	1792	176.2505	0.7249	1.5716
YSO-036	M	S	2	16	2381	209.0433	0.6847	1.328
YSO-037	M	S	2	17	2871	215.4793	0.777	1.5067
YSO-038	M	S	2	18	4617	298.7125	0.6502	1.0454
YSO-039	M	S	2	19	3839	248.3053	0.7824	1.4666
YSO-040	M	S	2	20	4398	276.5955	0.7224	1.4413
YSO-041	M	I	3	1	1156	142.9317	0.7111	1.3029
YSO-042	M	I	3	2	1762	180.2546	0.6815	1.1939
YSO-043	M	I	3	3	817	116.8072	0.7525	1.0776
YSO-044	M	I	3	4	2560	208.4677	0.7402	1.4775
YSO-045	M	I	3	5	2012	174.556	0.8298	1.1338
YSO-046	M	I	3	6	2785	210.8982	0.7868	1.206
YSO-047	M	I	3	7	2839	217.9695	0.7509	1.2122

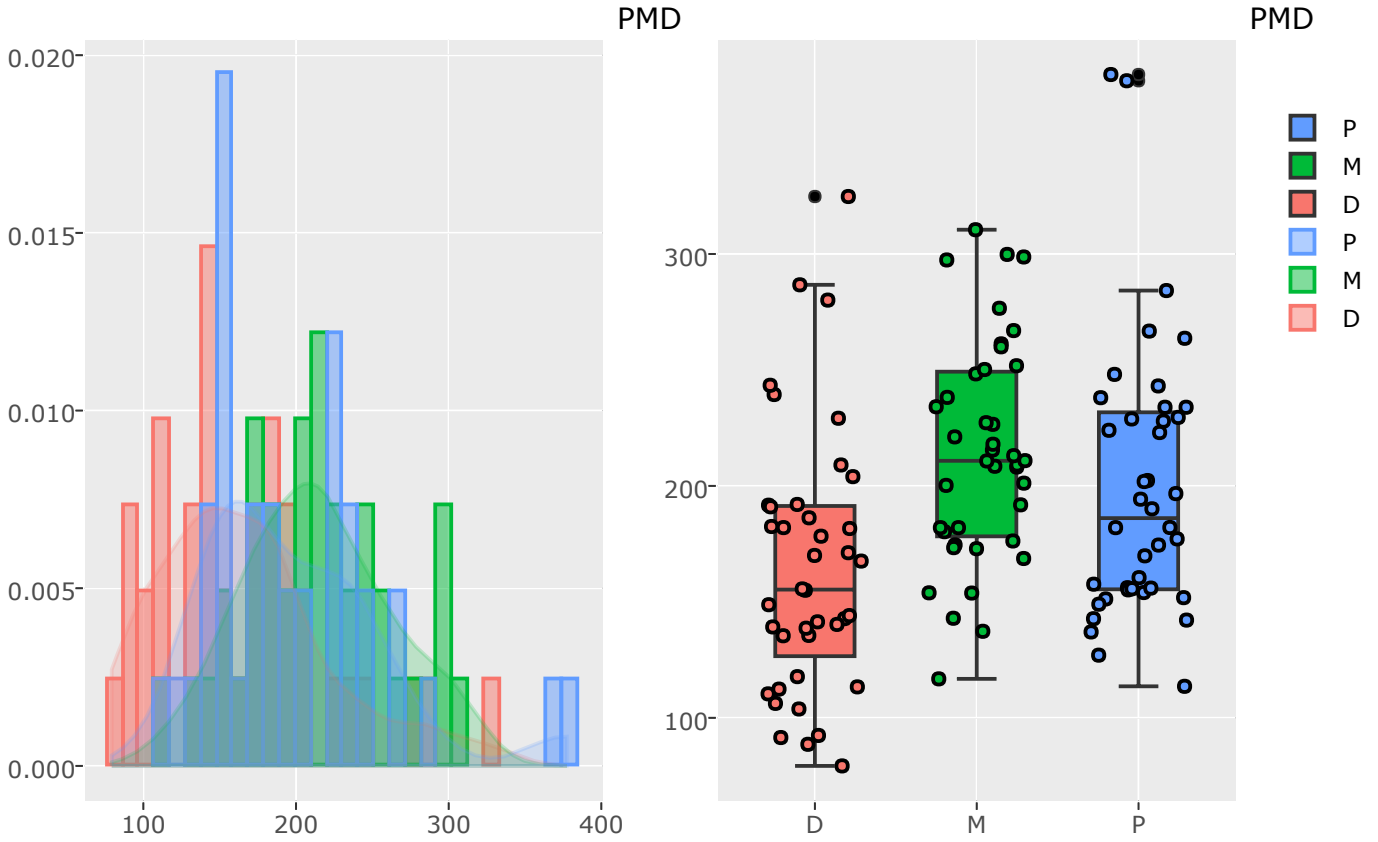
Raw Data

- **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.)

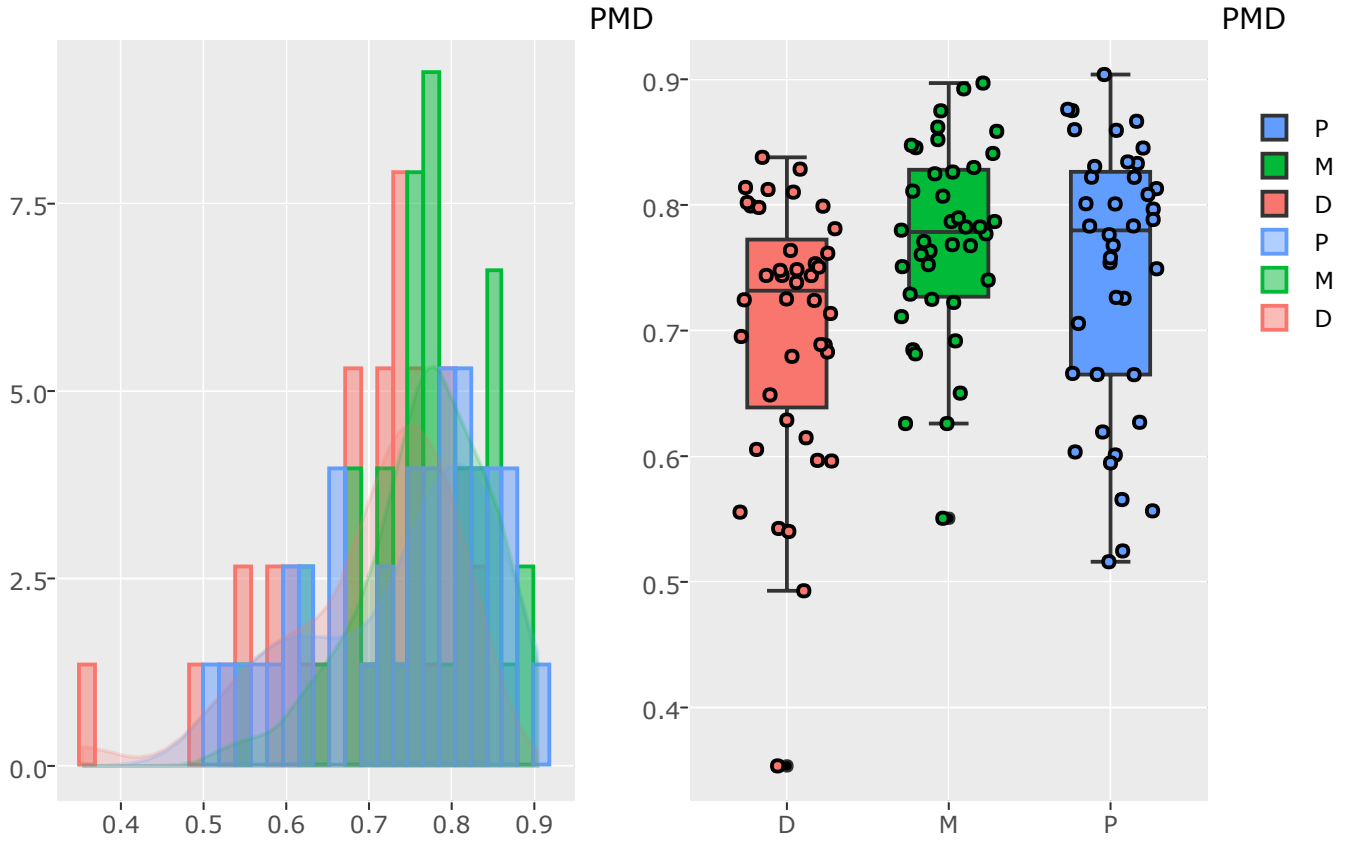
Data Exploration: Area



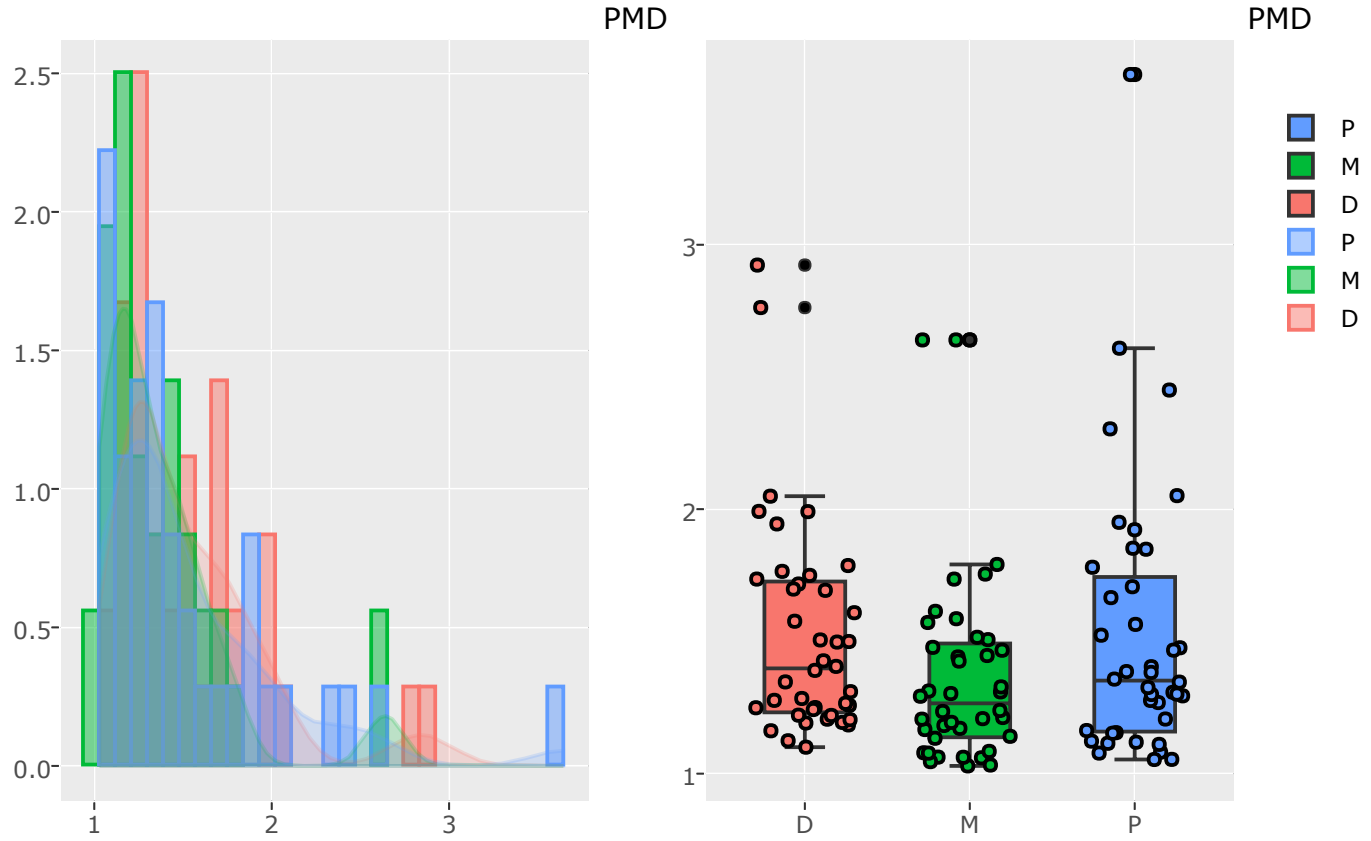
Data Exploration: Perimeter



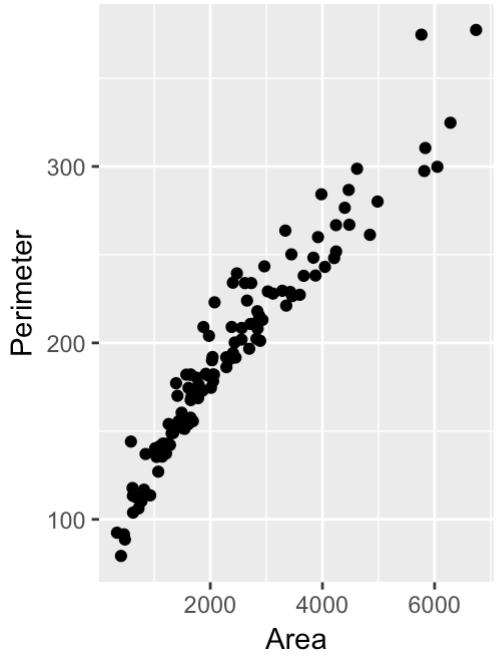
Data Exploration: Circularity

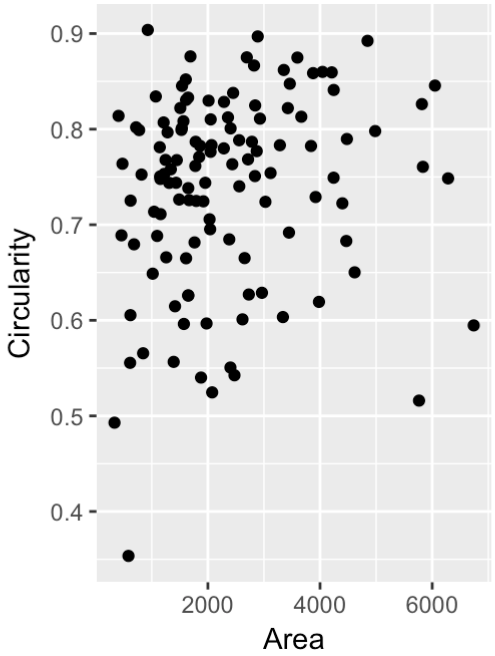


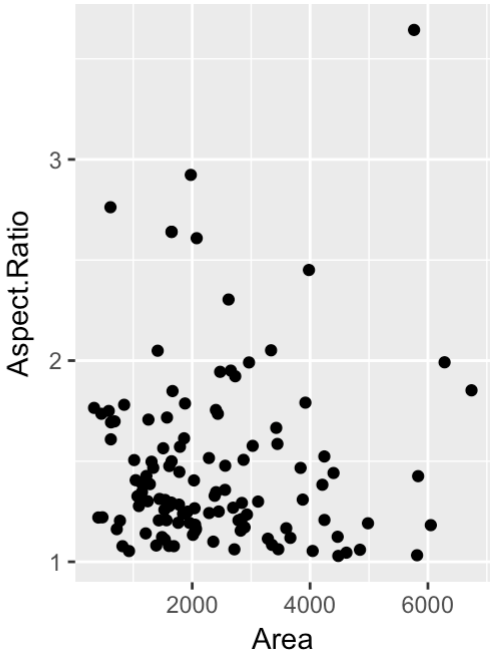
Data Exploration: Aspect Ratio



Data Exploration: Scatter Plots







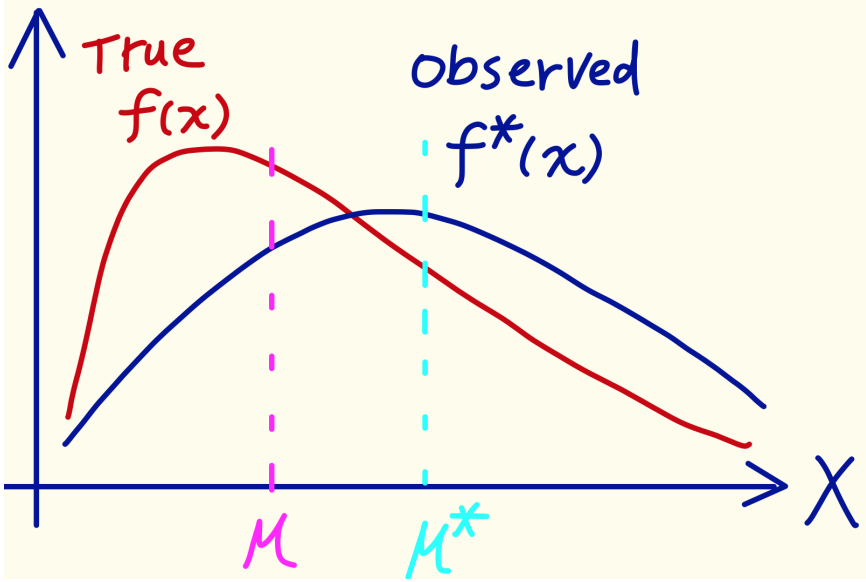
Best Estimators

- **Circularity:**
  Arithmetic Mean
- **Aspect Ratio:**
  Arithmetic Mean

New Goals for this project

1. **What is the appropriate estimator for the size-biased data?**
   A: Simulation Study for finding the best estimator.
2. Whether Properties of mitochondria are different by locations.
   A: Permutation Test and Bootstrapping Confidence Interval
3. Suggestions on sampling scheme for future research.
   A: Based on the Simulation Study.

Weighted Distribution



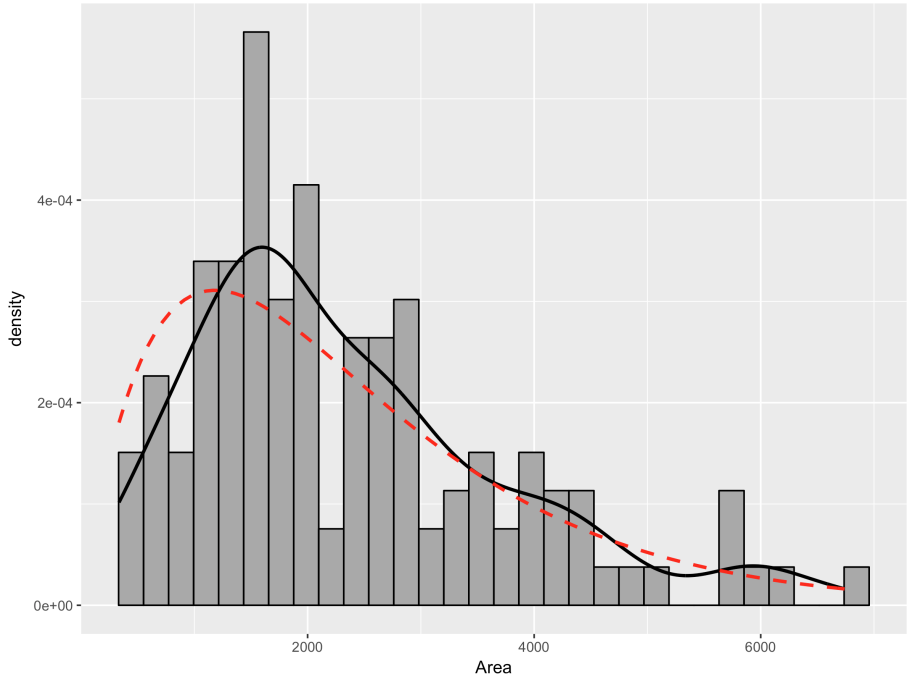
- Cox (1962) proposed an idea of Weighted Distribution,

$$f^*(x) = \frac{w(x)f(x)}{E_f(w(x))}.$$

- Cox (1962) also proposed that 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 $n \to \infty$, when $w(x) = x$.

Simulation Study - Area

- Assume the true distribution, $Area \sim Exp(\theta) = f(A)$.
- Then the observed distribution, $Area \sim Gamma(2, \theta) = f^*(A)$.
- The red dash line is $Gamma(2, \hat{\theta})$, where $\hat{\theta} = \frac{\bar{a}}{2} \doteq 1183$



Candidate Estimators - Area

1. **Arithmetic Mean (AM)**

$$\frac{\sum_{i=1}^{n} a_i}{n}$$

2. **Weighted Mean (WM) or Harmonic Mean**

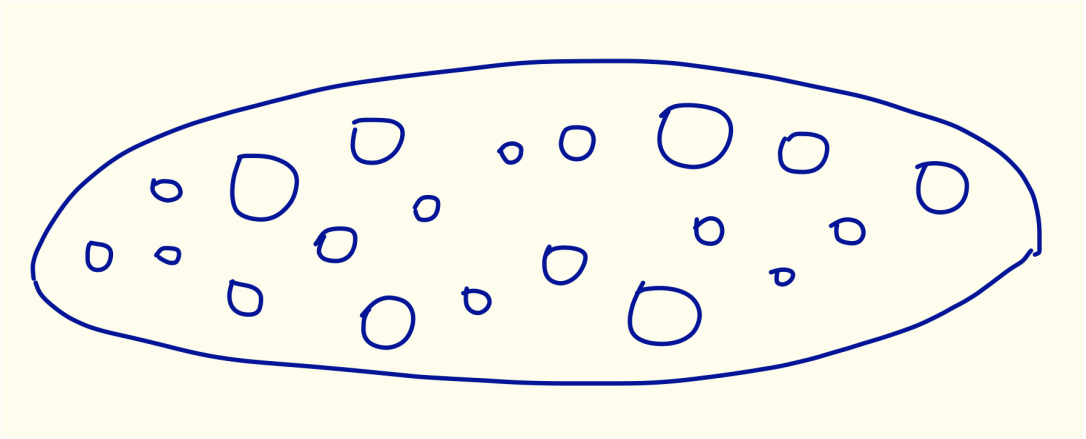
$$\frac{\sum_{i=1}^{n} w_i a_i}{\sum_{i=1}^{n} w_i} = \frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}}, \quad \text{where} \quad w_i = \frac{1}{p_i} = \frac{n\bar{a}}{a_i}$$

3. **Maximum Likelihood Estimator (MLE)**

$$\frac{\sum_{i=1}^{n} a_i}{2n} = \frac{AM}{2}$$

Simulation Study - Area (Overview)

- Simulate mitochondria data in a muscle fiber cell.
- Sample from **finite population (N)** rather than infinite population.
- Do both sampling **with** replacement and **without** replacement.
- **Sample size (n)** is decided by the **Ratio** between **N** and **n**.



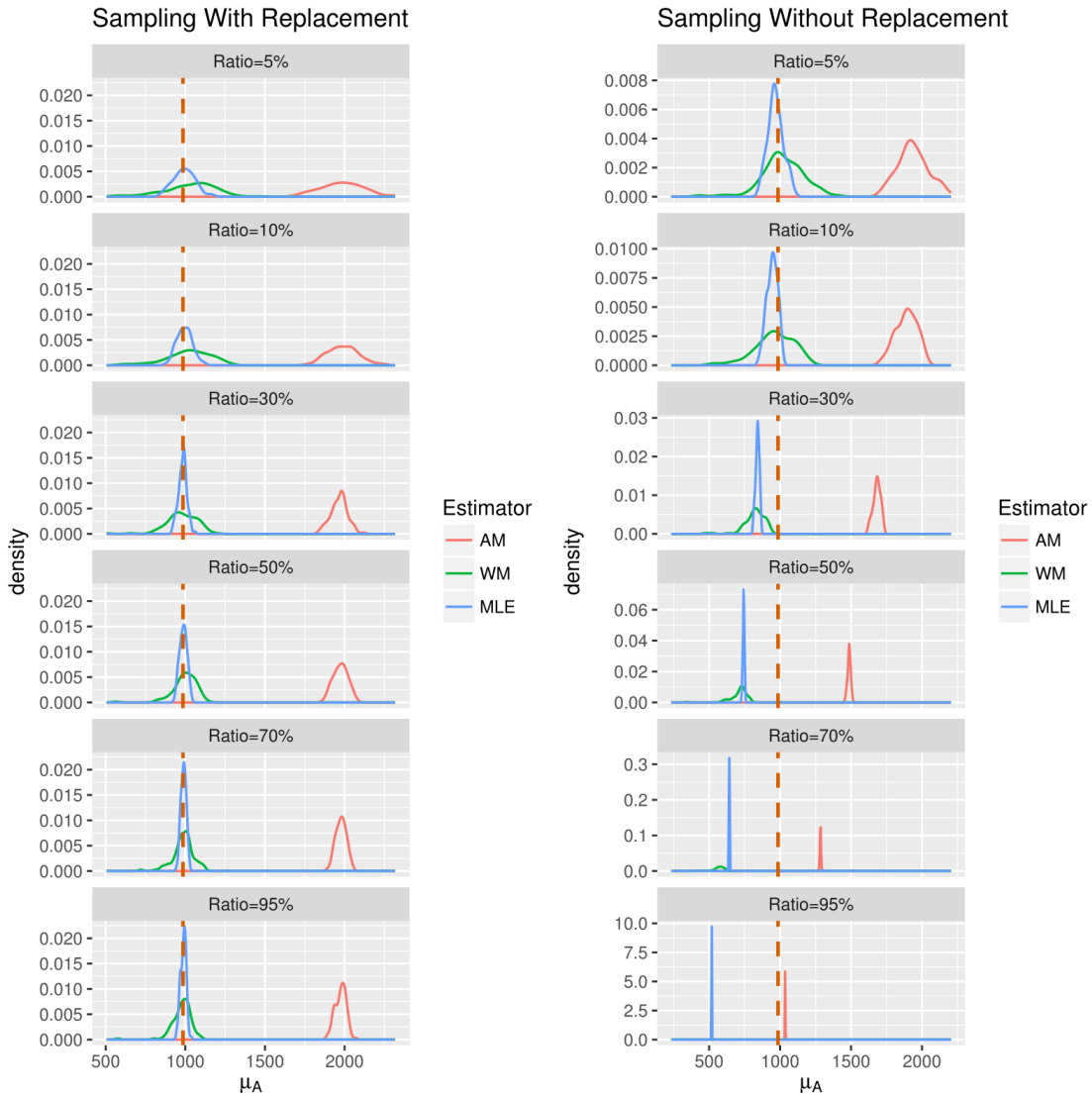
Simulation Study - Area

1. Assume $Area \sim Exp(\mu)$,
   Set $\mu = 1000$
   $N = 2000$,
   $Ratio = (5\%, 10\%, 30\%, 50\%, 70\%, 95\%)$,
   $Repeated\ Times = 1000$.
2. Generate N samples from $Exp(\mu)$ as subpopulation of Area
   Calculate subpopulation mean, μ_A, (This is what we are interested in!).
3. Sample a set of samples (n) from subpopulation (N) with sampling probability proportional to the value of Area with and without replacement ($n = N \times Ratio$).

Simulation Study - Area

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.
   Draw plots of sampling distributions for each candidate estimator.

Results of Simulation Study - Area

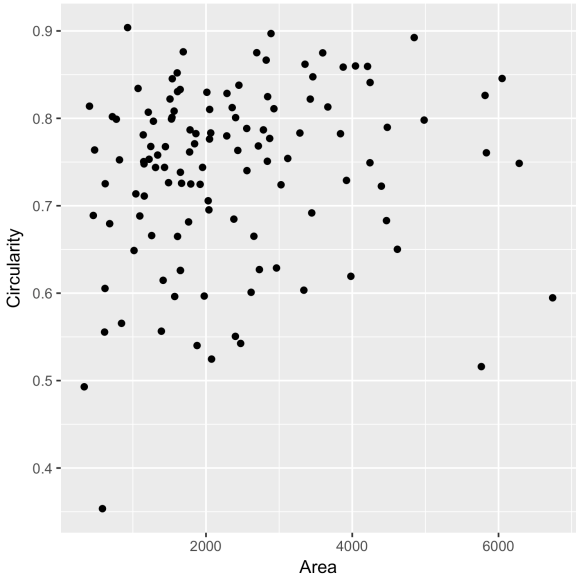


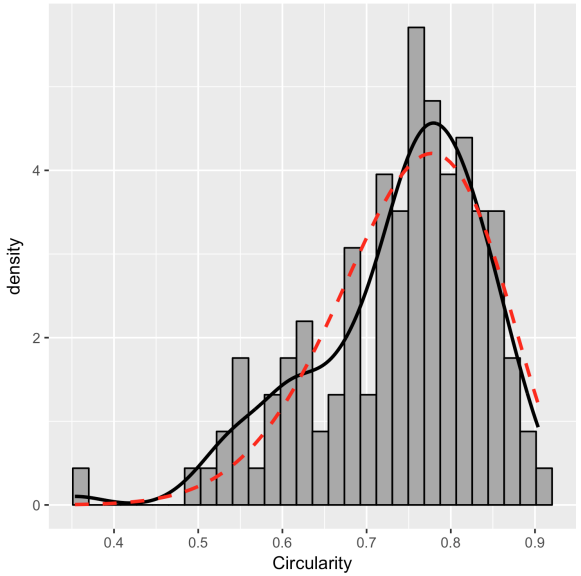
Best Estimators - Area

- **Sampling "WITH" Replacement:**
  Weighted Mean and MLE
- **Sampling "WITHOUT" Replacement:**
  Unfortunately, not clear yet.

Simulation Study - Perimeter

- $Perimeter = \sqrt{4\pi}\sqrt{\frac{Area}{Circularity}}$
- $Area \perp Circularity.$
- The observed distribution of $Circularity \sim Beta(15, 5).$
- Assume that the true distribution of $Circularity \sim Beta(\alpha, \beta).$
- The red dash line is $Beta(15, 5).$





Candidate Estimators - Perimeter

1. **Arithmetic Mean (AM)**

$$\frac{\sum_{i=1}^{n} p_i}{n}$$

2. **Weighted Mean (WM)**

$$\frac{\sum_{i=1}^{n} w_i p_i}{\sum_{i=1}^{n} w_i}, \quad \text{where} \quad w_i = \frac{n\bar{a}}{a_i}$$

3. **Delta Method Esitmator (DME)**

$$\sqrt{4\pi}\sqrt{\frac{\bar{a}/2}{\bar{c}}}$$

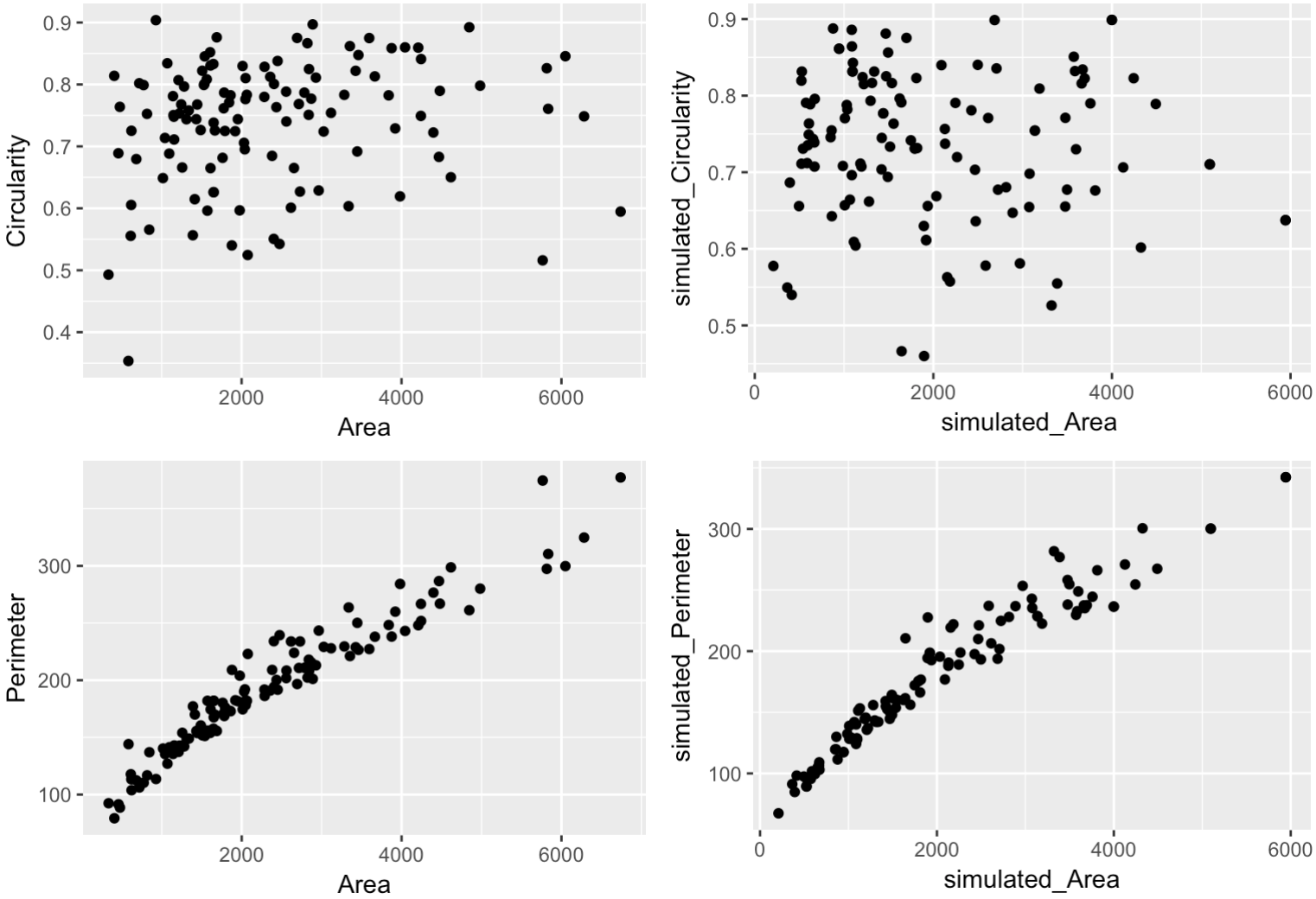
4. **2nd Order Taylor's Approximation Estimator (2TAE)**

$$\sqrt{4\pi}\left[\sqrt{\frac{\bar{a}/2}{\bar{c}}} - \frac{1}{8}\left(\frac{\bar{a}}{2}\right)^{\frac{-3}{2}}(\bar{c})^{\frac{-1}{2}}\frac{s_a^2}{2} + \frac{3}{8}\left(\frac{\bar{c}}{2}\right)^{\frac{1}{2}}(\bar{c})^{\frac{-5}{2}}s_c^2\right]$$

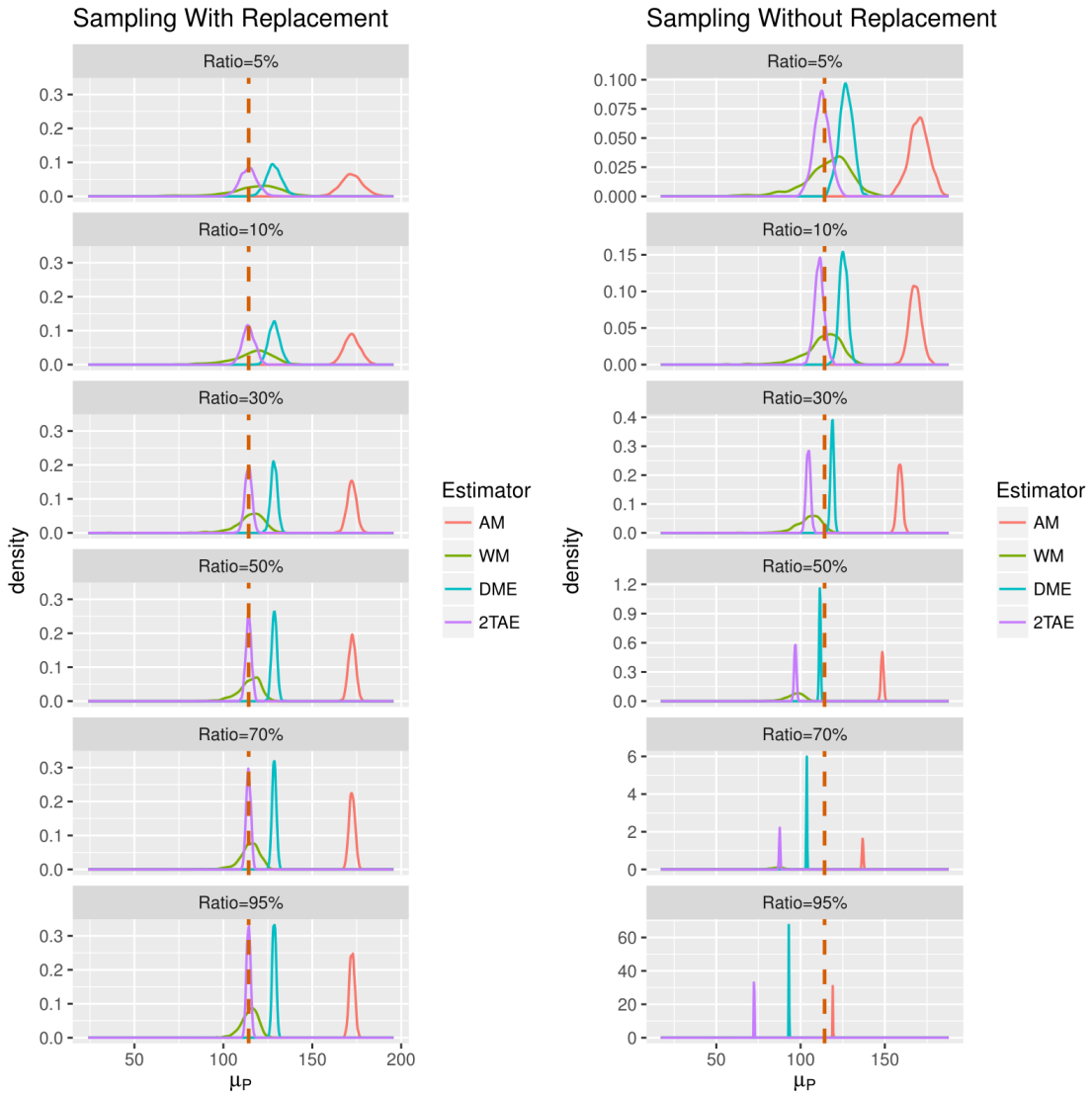
Simulation Study - Perimeter

1. Generate the **finite subpopulation (N)** data from $Circularity \sim Beta(15, 5)$.
2. Plug the generated Area and Circularity data into formula to obtain subpopulation of Perimeter.
3. Sample from the finite subpopulation (N) with sampling probability proportional to Area with and without replacement.
4. See the performance of the candidates estimators: **Arithmetic Mean (AM)**, **Weighted Mean (WM)**, **Delta Method Esitmator (DME)**, **2nd Order Taylor's Approximation Estimator (2TAE)**.

Simulated Data - Perimeter



Results of Simulation Study - Perimeter



Best Estimators - Perimeter

- **Sampling "WITH" Replacement:**
  Weighted Mean and 2TAE
- **Sampling "WITHOUT" Replacement:**
  Unfortunately, not clear yet.

Hypothesis Test

- **Overall Hypothesis Test:**

$$H_0 : \mu_{i_P} = \mu_{i_M} = \mu_{i_D}$$
$$H_A : \text{At least one } \mu_{i_j} \neq \mu_{i_k}$$

- **Pairwise Comparison Test:**

$$H_0 : \mu_{i_j} = \mu_{i_k}$$
$$H_A : \mu_{i_j} \neq \mu_{i_k}$$

$$i = \{\text{Area, Perimeter, Circularity, Aspect Ratio}\}$$
$$j, k = \{\text{P, M, D}\}$$

Hypothesis Test : Permutation Test

- **Reasons:**
  - Area and Perimeter are size-biased.
  - Circularity and Aspect Ratio, the data violated the normality assumption of ANOVA and T-test.

- **Overall Test (Permutation Test of ANOVA):**
  - $\sum_{i=\{P,M,D\}} (\widehat{\mu_i} - \widehat{\mu})^2$
  - significance level = 5%

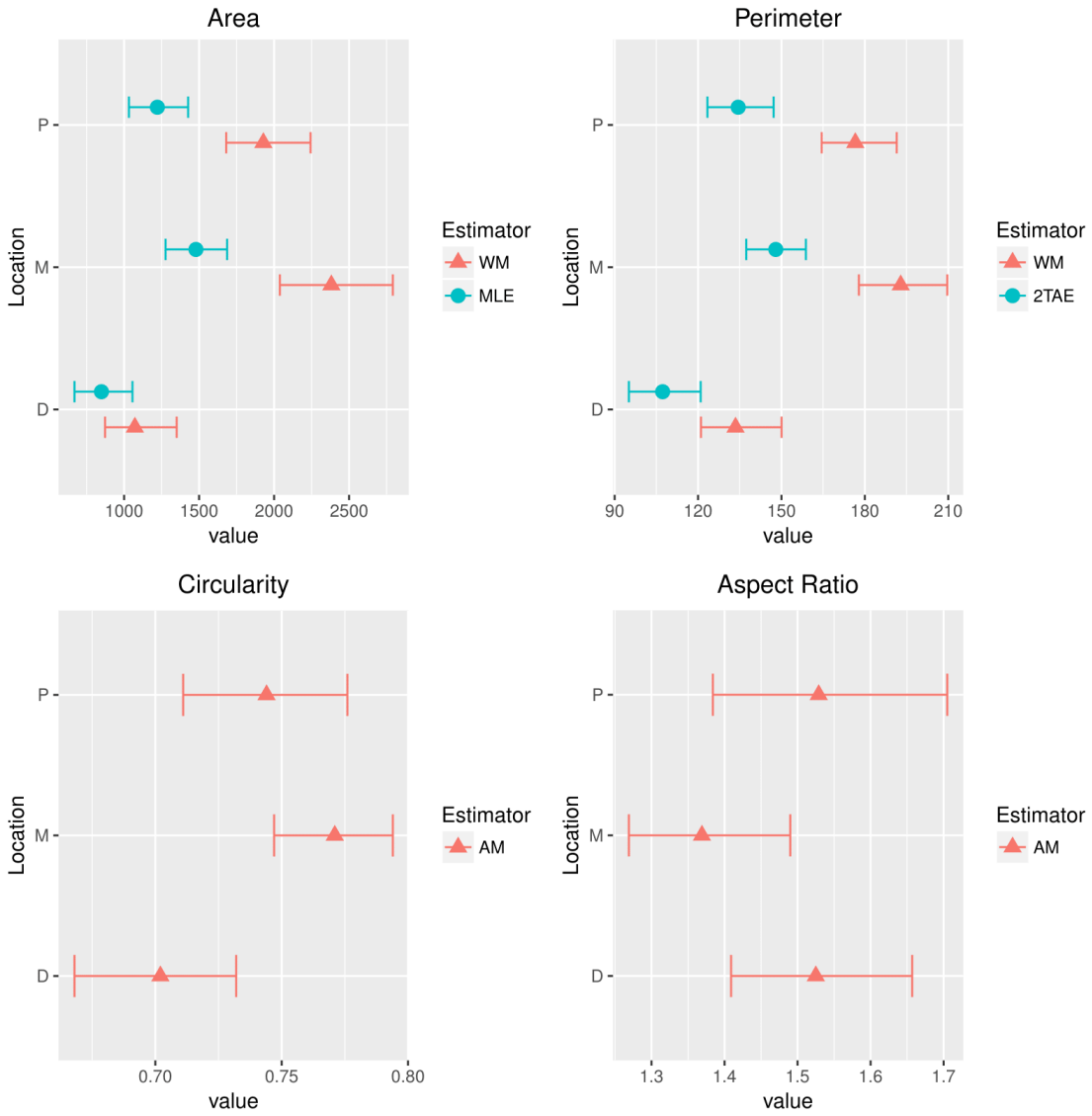
- **Pairwise Comparison Test (Permutation Test of T-test):**
  - $\widehat{\mu}_i - \widehat{\mu}_j$, where $i = \{P, M, D\}$
  - Bonferroni's correction: significance level = $\frac{5\%}{3} = 0.0167$.

Results for the Hypothesis Test

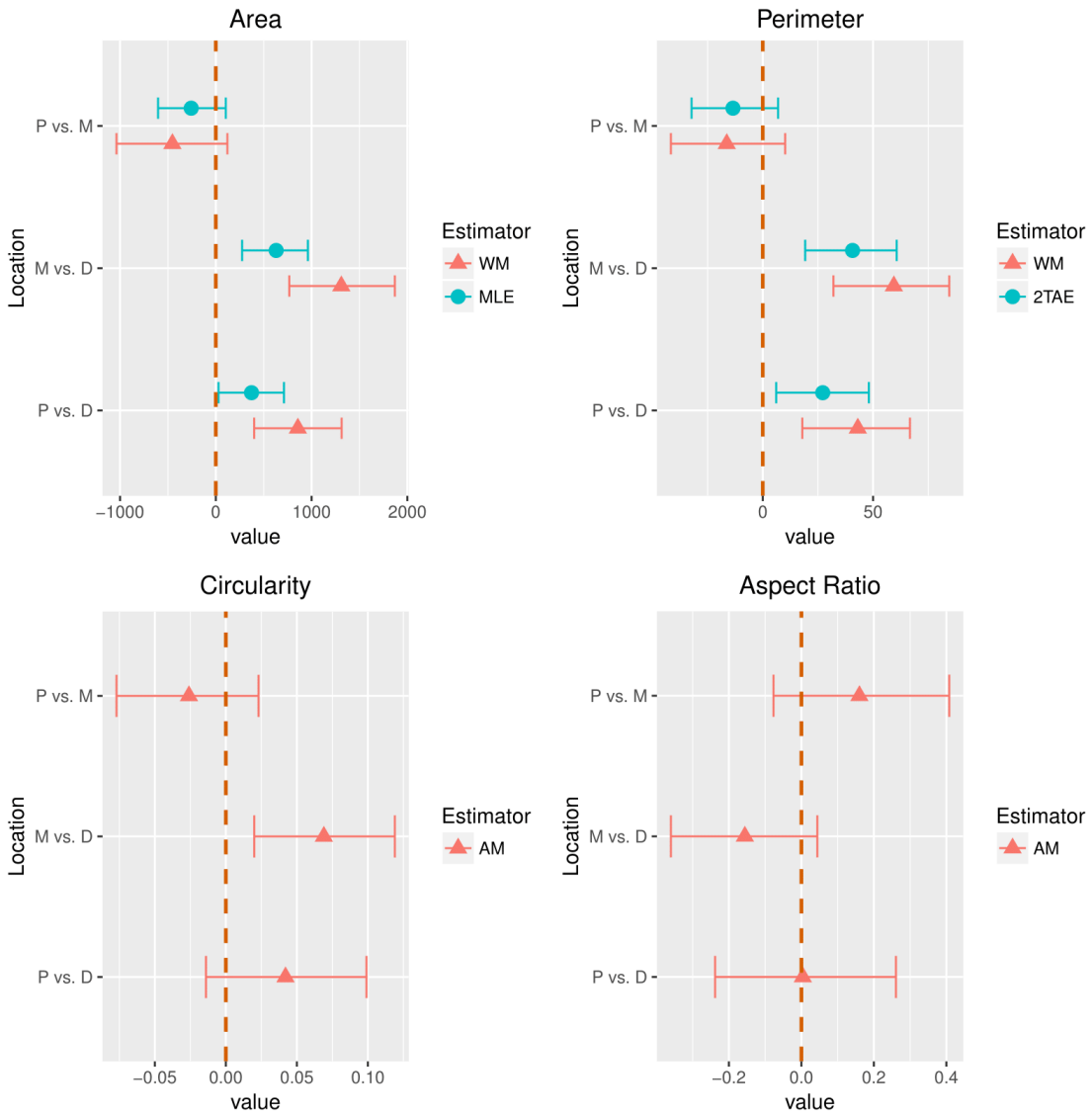
Property	Estimator	Overall	P vs. M	M vs. D	P vs. D
Area	WM	<0.0001	0.0974	<0.0001	0.0022
	MLE	0.0001	0.0950	0.0002	0.0140
Perimeter	WM	0.0001	0.2744	<0.0001	0.0018
	2TAE	<0.0001	0.1518	<0.0001	0.0024
Circularity	AM	0.0070	0.2476	0.0022	0.0616
Aspect Ratio	AM	0.1838	0.1046	0.1102	0.9884

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 to 0.0167.

Bootstrapping CI for Means



Bootstrapping CI for the differences



Conclusions

1. **Middle part and Proximal part** of the muscle fiber cell have significantly large Area, Perimeter and Circularity.
   - Area: $\underline{M > P} > D$
   - Perimeter: $\underline{M > P} > D$
   - Circularity: $\underline{M > P} > D$ & $M > \underline{P > D}$
   - Aspect Ratio: $\underline{M > P > D}$

2. The appropriate estimator for the size-biased data is **Non-parametric Weighted Mean**.

3. Suggest to use **Sampling With Replacement (SWR)** rather than Sampling Without Replacement (SWOR) in their future sampling scheme.

Future Work

- Find the best estimator for SWOR.
- Robustness of the distribution assumptions can be an interesting topic.
  The Nonparametric Weighted Mean had notably different results with the Parametric Estimators (MLE for Area and 2TAE for Perimeter). Maybe it is because of improper distribution assumptions on Area and Circularity.
- Include the effect of Subsarcolemmanl and Interfibrillar group and even possible interaction.

References

- Bratic, Ana and Larsson, Nils-Gran. “The Role of Mitochondria in Aging.” Journal of Clinical Investigation 123, no. 3 (2013): 951-57.
- Cox, D. R. Renewal Theory. London: Methuen, 1962.
- Patil,G. P. and Ord,J. K. “On Size-Biased Sampling and Related Form-Invariant Weighted Dis- tributions.” Sankhya. Series B 38,48-61.
- Jones, M. C. “Kernel Density Estimation for Length Biased Data.” Biometrika. Vol. 78, No. 3 (Sep., 1991), pp. 511-519

Photos

- Fishing Net: https://learning.blogs.nytimes.com/2012/04/19/poetry-pairing-trout/comment-page-1/?_r=0
- Mall: https://www.123rf.com/photo_30920353_people-in-shopping-mall-in-sofia-bulgaria.htmlhttps://www.123rf.com/photo_30920353_people-in-shopping-mall-in-sofia-bulgaria.html
- Muscle: http://slideplayer.com/slide/9024081/

The end

photo from http://www.desicomments.com/desi/thank-you/

Questions?

photo from http://www.bookcovercafe.com/independent-publishing-q-and-a-series-01/