Analysis of Size-Biased Mitochondria Data

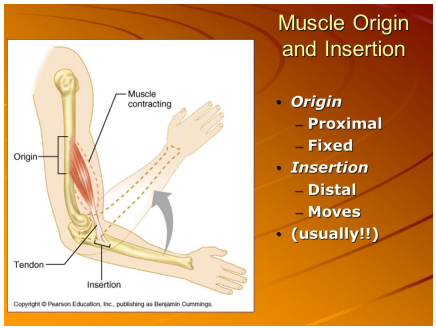
Students: Yin-Ting Chou
Advisor: Aaron Rendahl
5/17/2017

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. Best Estimator
   - Simulation Study
6. Hypothesis Test and Confidence Interval
   - Permutation Hypothesis Test
   - Boostrapping Confidence Interval
7. Conclusion
8. Discussion

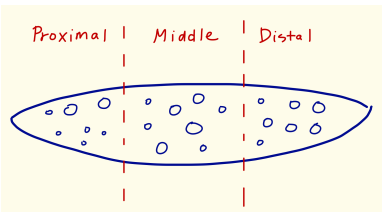
Story about Size-Biased Data

Scientific Background for Mitochondria



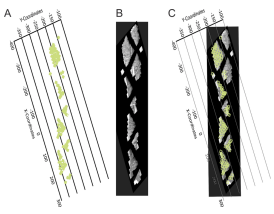
Goals for this project

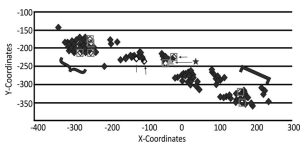
1. Whether **Properties** (area, perimeter, circularity and aspect ratio) of mitochondria are different by **locations** (proximal, middle and distal end).
2. Suggestions on **sampling method** for future research (more cells).



Sampling Process - 1

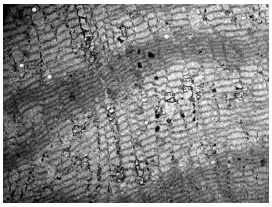
- A young muscle fiber cell was magnifired to 166 different images by using Transmission Electron Microscope (TEM).
- Those falls in " **{** " are defined as beging in **Proximal end**, in " **[** " are being in **Distal end**, and the rest are being in **Middle part**.





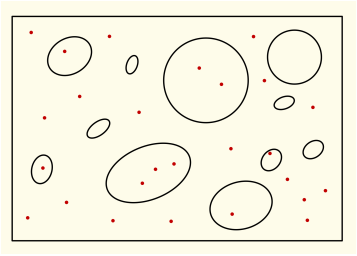
Sampling Process - 2

- For each location, divide images into two groups: Subsarcolemmanl and Interfibrillar group.
- In each group, randomly pick one image.
- In each image, randomly pick **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.



Raw Data

Show 10 entries

	Sample ID	image No.	# in image	image name	location	PMD	SI	Area
1	YSO-021	2	1	CS-120224-VII-Sol-Y-F1-5kx--35_-234_middle_subsarco.tif	MS	M	S	6049
2	YSO-022	2	2	CS-120224-VII-Sol-Y-F1-5kx--35_-234_middle_subsarco.tif	MS	M	S	4244
3	YSO-023	2	3	CS-120224-VII-Sol-Y-F1-5kx--35_-234_middle_subsarco.tif	MS	M	S	2889
4	YSO-024	2	4	CS-120224-VII-Sol-Y-F1-5kx--35_-234_middle_subsarco.tif	MS	M	S	3461
5	YSO-025	2	5	CS-120224-VII-Sol-Y-F1-5kx--35_-234_middle_subsarco.tif	MS	M	S	3876
6	YSO-026	2	6	CS-120224-VII-Sol-Y-F1-5kx--35_-234_middle_subsarco.tif	MS	M	S	4847
7	YSO-027	2	7	CS-120224-VII-Sol-Y-F1-5kx--35_-234_middle_subsarco.tif	MS	M	S	1606

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.)

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 our population mean estimator, it will definitely be overestimated!

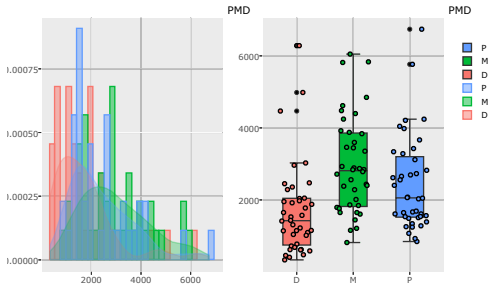
New Goals for this project

1. **What is the appropriate estimator for the size-biased data?**
2. Whether Properties of mitochondria are different by locations.
3. Suggestions on sampling scheme for future research.

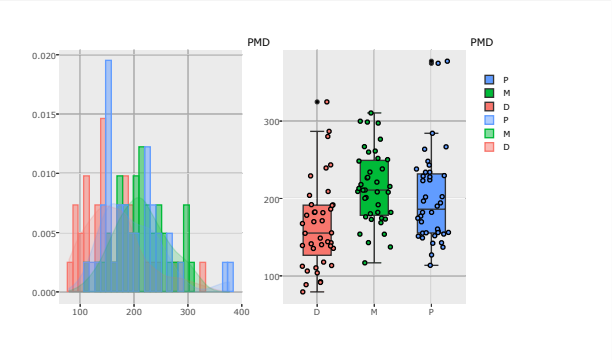
New Goals for this project

1. **What is the appropriate estimator for the size-biased data?**
   A: Simulation Study.
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.

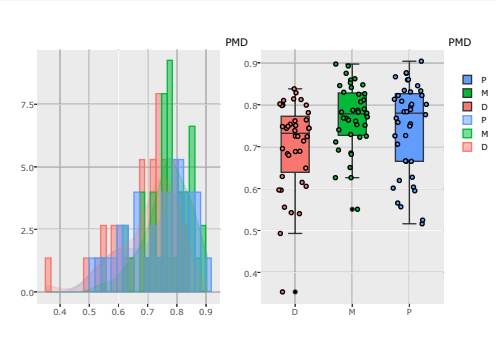
Data Exploration: Area



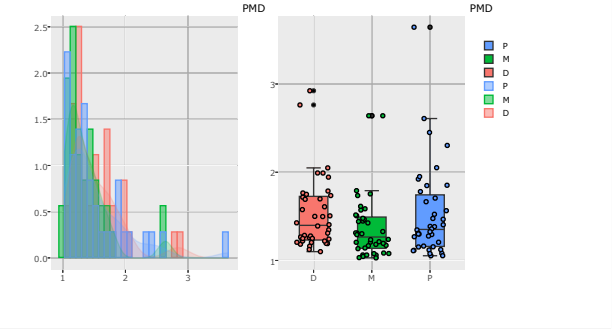
Data Exploration: Perimeter



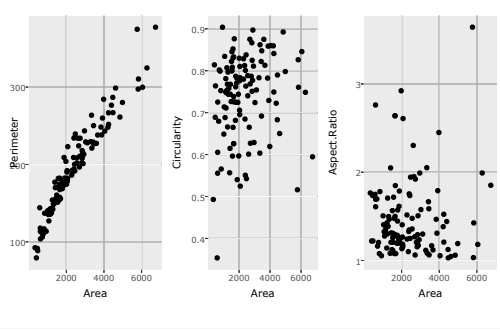
Data Exploration: Circularity



Data Exploration: Aspect Ratio



Data Exploration: Scatter Plots

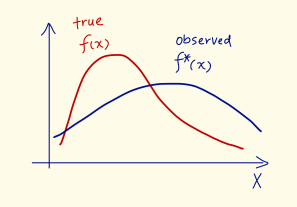


Weighted Distribution

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

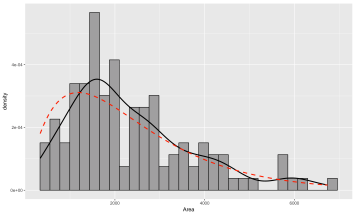
$$f^*(x) = \frac{w(x)f(x)}{E_f(w(x))}$$

- Cox (1962) also proposed 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 $n \to \infty$.



Simulation Study - Area

- Suppose that the true distribution of $Area \sim Exp(\theta)$.
- The observed distribution of $Area \sim Gamma(2, \theta)$.
- The red dash line is $Gamma(2, \hat{\theta})$, where $\hat{\theta} = \bar{a} = 1183$



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. Maxima Likelihood Estimator (MLE)

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

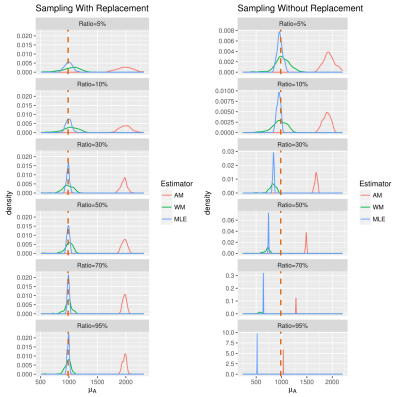
Simulation Study - Area

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 $Ratio$.

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. Also draw plots of sampling distributions for each candidate estimator.

Results of Simulation Study - Area

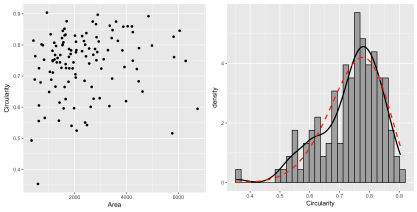


Best Estimators

- **Area:**
  Weighted Mean and MLE.

Simulation Study - Perimeter

- Area is independent to Circularity.
- Perimeter= $\sqrt{4\pi}\sqrt{\frac{\text{Area}}{\text{Circularity}}}$
- Suppose that the true distribution of $Circularity \sim Beta(\alpha, \beta)$.
- The observed distribution of $Circularity \sim Beta(15, 5)$.
- 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)

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

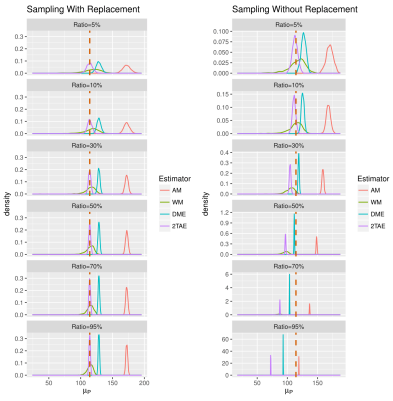
Simulation Study - Perimeter

1. Set $N = 2000$; $Ratio$ between N and n are $(5$; $Repeated\ Times = 1000$ 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. 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, $\text{Perimeter} = \sqrt{4\pi}\sqrt{\frac{\text{Area}}{\text{Circularity}}}$, 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.

Simulation Study - 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.

Results of Simulation Study - Perimeter



Best Estimators

- **Area:**
  Weighted Mean and MLE.
- **Perimeter:**
  Weighted Mean and 2TAE.
- **Circularity:**
  Arithmetic Mean (because it is independent to Area)
- **Aspect Ratio:**
  Arithmetic Mean (because it is independent to Area)

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

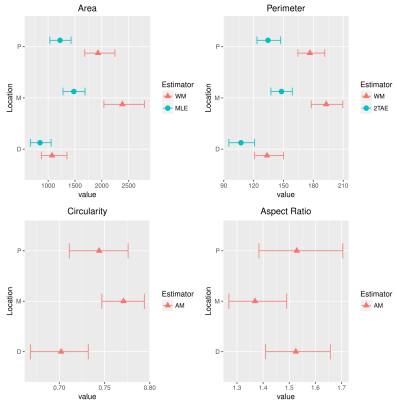
- **Reasons:**
  - Area and Perimeter are size-biased.
  - Circularity and Aspect Ratio, the data violated the normality assumption of ANOVA and T-test.
- **Overall Test (Permutation Test of ANOVA):**
  - significance level = 5%
- **Pairwsie Comparison Test (Permutation Test of T-test):**
  - Bonferroni's correction, for its easy interpretation and its simultaneous confidence interval for the mean differences.
  - significance level = $\frac{5\%}{3} = 0.0167$

Results for the Hypothesis Test

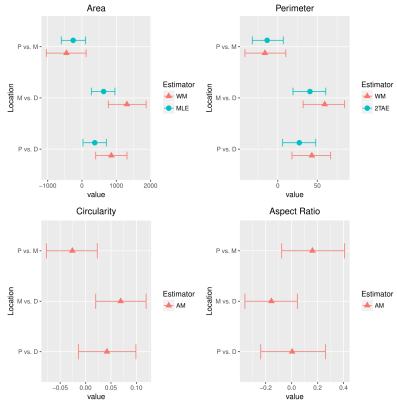
Property	Estimator	Overall	P vs. M	M vs. D	P vs. D
Area	WM	<0.0001	0.0974	<0.0001	0.0022
	MLE	0.0001	0.0950	0.0002	0.0140
Perimeter	WM	0.0001	0.2744	<0.0001	0.0018
	2TAE	<0.0001	0.1518	<0.0001	0.0024
Circularity	AM	0.0070	0.2476	0.0022	0.0616
Aspect Ratio	AM	0.1838	0.1046	0.1102	0.9884

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. What is the appropriate estimator for the size-biased data?
   A: Use Nonparametric Weighted Mean as the best estimator for population mean and do hypothesis test based on this estimator. (because of none distribution assumptions)
2. Whether Properties (Area, Perimeter, Circularity, Aspect Ratio)of mitochondria are different by locations.
   A: Middle part of the muscle fiber cell have larger Area, Perimeter and Circularity which means to support muscle contraction more energy is needed in Middle.
3. Suggestions on sampling scheme for future research.
   A: 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.

Discussion

- Finding the best estimator for SWOR is a potential area for future work.
- 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, in our data, things are not like what we expected.
- Maybe it is because of 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

- 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/