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,

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

$$\Leftrightarrow E_f(X) \cdot \frac{1}{x} \cdot f^*(x) = f(x)$$

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

$$\Leftrightarrow E_{f^*}(\frac{1}{X}) = \frac{1}{E_f(X)}$$

then by the Weak Law of Large Numbers,

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

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

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 ($P = \sqrt{4\pi}\sqrt{\frac{A}{C}} = f(A, C)$), we can obtain Delta Method Estimator(DME), $\widehat{\mu_P} = \sqrt{4\pi}\sqrt{\frac{\mu_{\widehat{A,MLE}}}{\mu_{\widehat{C,MLE}}}} = f(\mu_{\widehat{A,MLE}}, \mu_{\widehat{C,MLE}}) = f(\bar{A}/2, \bar{C}) = \sqrt{4\pi}\sqrt{\frac{\bar{A}/2}{\bar{C}}}$, and 2nd Order Taylor's Approximation Estimator of population mean Perimeter(μ_P).
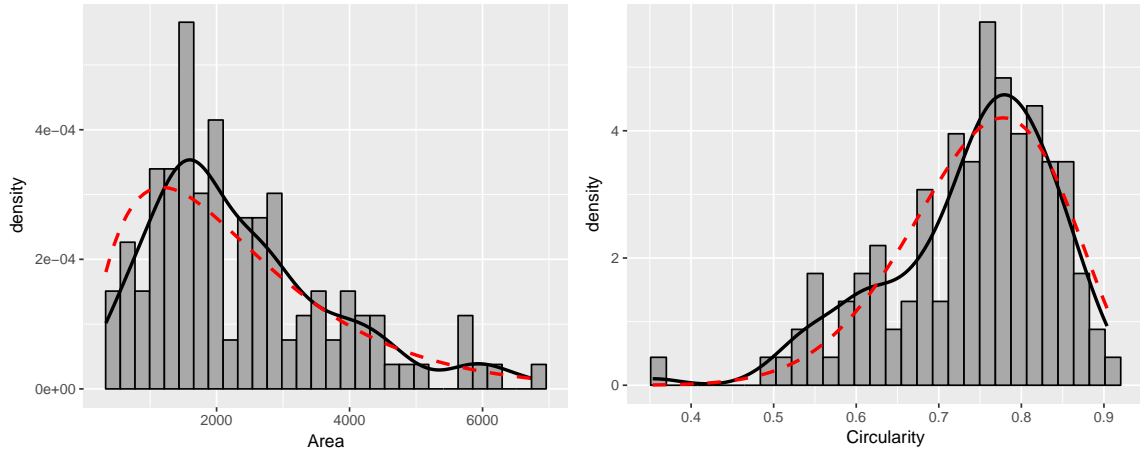

Figure 4.0.1: Histogram of Area and Circularity

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,

$$E(P) \doteq E(\sqrt{4\pi} \left[f(\mu_A, \mu_C) + \frac{1}{1!} \left(f_x(\mu_A, \mu_C)(x - \mu_A) + f_y(\mu_A, \mu_C)(y - \mu_C) \right) \right.$$

$$\left. + \frac{1}{2!} \left(f_{xx}(\mu_A, \mu_C)(x - \mu_A)^2 + f_{yy}(\mu_A, \mu_C)(y - \mu_C)^2 + 2f_{xy}(\mu_A, \mu_C)(x - \mu_A)(y - \mu_C) \right) \right])$$

$$\doteq \sqrt{4\pi} \left[\sqrt{\frac{\mu_A}{\mu_C}} - \frac{1}{8}(\mu_A)^{\frac{-3}{2}}(\mu_C)^{\frac{-1}{2}} Var(A) + \frac{3}{8}(\mu_A)^{\frac{1}{2}}(\mu_C)^{\frac{-5}{2}} Var(C) \right]$$

Then, plug in the parameters with their MLEs,

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