NUMT Risk Confidence Intervals

Problem Setup

This note documents the method used to calculate 95% confidence intervals for the risk of mitochondrial diseases caused by nuclear gene mutations, as inferred from nuclear mitochondrial (NUMT) gene sequence data.

The problem setup is as follows:

- You have k loci.
- For each locus i, you have:
  - n_i trials,
  - s_i observed non-wild-type alleles,
  - An estimated probability $\hat{p}_i = \frac{s_i}{n_i}$.
- The probability P of having double non-wild-type alleles at any of the k loci is (one minus the probability of being unaffected):

$$P = 1 - \prod_{i=1}^{k} (1 - p_i^2)$$

We compute the confidence interval on the risk of having double non-wild-type alleles at any of k nuclear genome loci using the delta method.

Step 1: Confidence Interval for p_i

The Clopper-Pearson method can be used to construct an exact confidence interval for the binomial proportion p_i. For each locus i, the estimate $\hat{p}_i$ is calculated as $\hat{p}_i = \frac{s_i}{n_i}$. The confidence interval for p_i is given by:

$$\text{CI}_{95\%}(p_i) = [L_i, U_i]$$

where L_i and U_i are the lower and upper bounds of the Clopper-Pearson confidence interval for p_i based on the observed successes s_i and trials n_i.

Step 2: Confidence Interval for p_i^2

Given that the probability p_i lies within the interval $[L_i, U_i]$, the confidence interval for p_i^2 can be derived by squaring the endpoints:

$$\mathrm{CI}_{95\%}(p_i^2) = \left[L_i^2, U_i^2\right]$$

This interval is valid because the squaring function is a continuous and monotonic bijection on the interval $[0, 1]$, preserving the confidence level.

Step 3: Compute the Probability P

First compute the probability P using the estimated probabilities $\hat{p}_i$:

$$P = 1 - \prod_{i=1}^{k}(1 - \hat{p}_i^2)$$

Step 4: Apply the Delta Method

For each locus i, compute the partial derivative of P with respect to $\hat{p}_i$:

$$\frac{\partial P}{\partial \hat{p}_i} = 2\hat{p}_i \cdot \prod_{j \neq i}(1 - \hat{p}_j^2)$$

This derivative reflects how P changes with respect to each $\hat{p}_i$.

Approximate the variance of P by summing the contributions from each $\hat{p}_i$:

$$\mathrm{Var}(P) \approx \sum_{i=1}^{k} \left(\frac{\partial P}{\partial \hat{p}_i}\right)^2 \cdot \mathrm{Var}(\hat{p}_i)$$

where

$$\mathrm{Var}(\hat{p}_i) = \frac{\hat{p}_i(1 - \hat{p}_i)}{n_i - 1}$$

The standard deviation of P is the square root of the variance:

$$\sigma_P = \sqrt{\mathrm{Var}(P)}$$

Step 5: Construct the Confidence Interval for P

Assuming P follows a normal distribution (reasonable under the Central Limit Theorem for large n_i), a 95% confidence interval for P can be constructed as:

$$\mathrm{CI}_{95\%} = [P - 1.96 \cdot \sigma_P, P + 1.96 \cdot \sigma_P]$$