Confidence Intervals for Risk of Mitochondrial Diseases Inferred from NUMT Sequence Data

Problem Setup

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

1. Definitions:
   - Gene copy: One of the two copies of a gene present at a locus in a diploid organism.
   - Non-wild-type variant: Any variant of a gene that differs from the wild-type (normal) sequence.
   - Biallelic non-wild-type variants: Both gene copies at a locus carry non-wild-type variants, which can be identical (homozygous) or different (compound heterozygous).
2. Data and assumptions:
   - We have k loci in the nuclear genome linked to mitochondrial disease.
   - For each locus i, we have:
     - n_i observations, i.e. gene copies examined. These have been obtained from a population of $\frac{n_i}{2}$ individual healthy older adults, each contributing two gene copies. In practice, in the dataset in question, n_i is identical for all loci i.
     - s_i observed occurrences of any non-wild-type variant,
     - An estimated probability $\hat{p}_i = \frac{s_i}{n_i}$, representing the probability that a gene copy at locus i is a non-wild-type variant.
   - Assumptions:
     - The loci are independent of each other.
     - Allele frequencies are in Hardy-Weinberg equilibrium.
3. Calculating the probability P:
   - Individuals with biallelic non-wild-type variants are likely to be affected by mitochondrial disease.

- The probability of being unaffected at locus i is $1 - p_i^2$.
- Therefore, the probability P of being affected due to biallelic non-wild-type variants at any of the k loci is:

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

- We compute an approximate confidence interval on the risk of having biallelic non-wild-type variants at any of k nuclear genome loci using the delta method and test this using bootstrap resampling.

Delta Method

1. Confidence Interval for $\hat{p}_i$:
   - Use the Clopper-Pearson method to construct an exact confidence interval for each $\hat{p}_i$.
2. Confidence Interval for $\hat{p}_i^2$
   - Since squaring is a monotonic function on $[0, 1]$, square the endpoints of the confidence interval for $\hat{p}_i$ to obtain the interval for $\hat{p}_i^2$.
3. Compute the Probability P
   - Compute the probability P using the estimated probabilities $\hat{p}_i$:

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

4. Apply the Delta Method:
   - Compute the partial derivative of P with respect to each $\hat{p}_i$:

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

   This derivative reflects how P changes with respect to each $\hat{p}_i$.
   - Approximate the variance of P by summing estimates of 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

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

5. Construct the Confidence Interval for P:

   - Calculate the standard deviation of P:

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

   - Assuming P follows a normal distribution—reasonable for large n_i or k—a 95% confidence interval for P can be constructed as:

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

     where z is the z-score for the desired confidence level (1.96 for 95%).

Limitations of the approach

While the Delta method is useful for approximating variances, its application in our analysis faces challenges, due to the nature of our data:

1. **Violation of Normality Assumptions**

   Many loci have very low counts of non-wild-type variants (1 or 2 out of ∼5,000 gene copies), resulting in extremely small estimated probabilities ($\hat{p}_i$). The sampling distribution of $\hat{p}_i$ is skewed, violating the normality assumption required by the Delta method. This can lead to unreliable variance estimates and confidence intervals.

2. **Nonlinearity and Boundedness of P**

   The function $P = 1 - \prod_{i=1}^{k}(1 - \hat{p}_i^2)$ is nonlinear, and P is bounded between 0 and 1. The Delta method's reliance on linear approximation and normal distribution may not adequately capture the behavior of P, especially when P is near 0. This could in principle result in confidence intervals that extend beyond [0,1] or do not reflect the true variability of P.

Mitigating these issues is the fact that loci with higher variant counts contribute most to the overall risk P. Nevertheless, inaccuracies at loci with low variant counts could cumulatively affect the confidence interval. It is hard to assess the extent to which this might impact the derived confidence interval *a priori*. The most direct way to check this is to use a bootstrap analysis.

Bootstrap Approach Rationale

The bootstrap method is nonparametric and does not rely on normality assumptions or large sample sizes. It can accommodate the nonlinear nature of P and effectively handle small counts and skewed distributions by empirically estimating the sampling distribution of P. By resampling from the observed data, the bootstrap approach incorporates variability from all loci, ensuring that the confidence interval reflects the combined uncertainty in our risk estimate.

Implementation of the Bootstrap Approach

We implement the bootstrap method as follows:

- For each locus i, generate bootstrap samples by resampling n_i observations with replacement from the original data.
- For each bootstrap sample, calculate the estimated probabilities $\hat{p}_i^*$ and compute $P^* = 1 - \prod_{i=1}^{k}(1 - (\hat{p}_i^*)^2)$.
- Repeat this process many times (we use 10,000 iterations) to build an empirical distribution of P^*.
- Determine the 95% confidence interval for P using the 2.5th and 97.5th percentiles of the bootstrap distribution.

Results

We calculate the probability P of individuals having biallelic non-wild-type variants at any of the k loci, representing the estimated risk of mitochondrial diseases caused by nuclear gene mutations inferred from NUMT sequence data.

- Using the Delta method, the estimated probability corresponds to approximately **99.5 affected individuals per 100,000 people**. The 95% confidence interval for P obtained via the Delta method ranges from **84.8 to 114.2 affected individuals per 100,000 people**.
- Applying the bootstrap method, the estimated probability corresponds to approximately **104.0 affected individuals per 100,000 people**. The 95% confidence interval for P from the bootstrap method ranges from **89.7 to 119.9 affected individuals per 100,000 people**.

Interpretation

The results from the Delta method and the bootstrap method are similar, with overlapping confidence intervals. This similarity suggests that, despite the limitations associated with the Delta method in the context of low variant counts, it provides an acceptable approximation for the overall risk estimation.