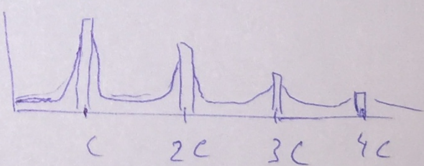
Andrey Bzikadze

② We need a test for selection based on statistics $D = \hat{\Theta}_1 - \hat{\Theta}_2$ where $\hat{\Theta}_1, \hat{\Theta}_2$ are two different estimates of Θ. We'd like to have a z-score $Z = D/\text{std}(D) \underset{n\to\infty}{\xrightarrow{\sim}} N(0;1)$ under null hypothesis that there is no selection. $\Rightarrow$ pvalue $= 2\min(\Phi^{-1}(z), 1-\Phi^{-1}(z))$.

1) I propose $\hat{\Theta}_1$ - a Tajima's $\underline{\underline{k}}$. From assignment 2.5 we know that it looks like $k = \sum_{j=1}^{m} f_j(1-f_j)$, where f_j is the frequency of mutation allele at site j. I'm not sure that $Ek = \Theta$ (exactly), but I think it is possible to estimate the possible bias with simulations and account for it.

2) For $\hat{\Theta}_2$ I'd propose a simple method to estimate coverage of a distinct indiv. First, suppose that $n_i = n_j$ $\forall i,j$ and reads are error-free. We as usual denote m_i as # mutations with exactly i ones (reads)
$\Rightarrow$ $\overline{m}_i$ (from data) : $\overline{m}_i = 0$ $\forall i < C$ where C is coverage of a distinct indiv.
If the reads have errors and the coverage of distinct individuals is not exactly identical (but close) then $\overline{m}_C \gg \overline{m}_i$ $\forall i < C$ and AFS could look like this (under no selection)



The same when $n_i \neq n_j$ exactly, but close.
$\Rightarrow$ We can define estimates: $\hat{\Theta}_i^{(2)} = i\cdot C \cdot \overline{m}_{iC}$ // $iC \le n-1 \Rightarrow$ $i \le \left\lfloor \frac{n-1}{C} \right\rfloor$
It is known that $E\hat{\Theta}_i^{(2)} = \Theta$.

Finally, Define $\hat{\Theta}_2 := \frac{1}{\underbrace{\left\lfloor \frac{n-1}{C} \right\rfloor}_{=: n_C}} \cdot \sum_{i=1}^{\left\lfloor \frac{n-1}{C} \right\rfloor} \hat{\Theta}_i^{(2)} = \frac{1}{n_C} \sum_{i=1}^{n_C} \hat{\Theta}_i^{(2)}$.

$E\hat{\Theta}_2 = \Theta$. Motivation for $\hat{\Theta}_2$ is $Var(\hat{\Theta}_2) \approx \frac{1}{n_C^2} \sum_{i=1}^{n_C} Var(\hat{\Theta}_i^{(2)})$. (if we ignore correlations)
$\Rightarrow$ $Var(\hat{\Theta}_2)$ could be potentially lower than that of any $\hat{\Theta}_i^{(2)}$ $\forall i$.
$\Rightarrow$ $D := \hat{\Theta}_1 - \hat{\Theta}_2$.

---

3) When the selection is positive $\Rightarrow$ $\hat{\Theta}_1 \searrow$ (was explained during lections) but $\hat{\Theta}_2 \nearrow$ as $\overline{m}_{iC} \uparrow$ for large i, and $\overline{m}_{iC}$ has a weight iC. $\Rightarrow D < 0$

4) Analog. for negative selection $\Rightarrow$ $\hat{\Theta}_1 \uparrow$ and $\hat{\Theta}_2 \downarrow$ (as $\overline{m}_{iC} \uparrow$ for small i, but the coefficient is low) $\Rightarrow$ $D > 0$.