Background

Peppered moths are a species of moth that have been widely studied as an example of natural selection in evolution. Each moth will inherit one allele from each parent. The wing color of these moths is determined by a single gene with three possible alleles, denoted C, I, and T. C is dominant to I and T is recessive to I. The combination of the two alleles that a moth has is called its genotype.

Because C is dominant, the genotypes CC, CI, and CT result in the same *carbonaria* phenotype where the wings exhibit solid black coloring. The genotype TT results in the *typica* phenotype with light-colored wings. And because T is recessive to I, both of the II and IT genotypes produce a middle-ground *insularia* phenotype with mottled/patchy wings. **Note that there are six possible genotypes ($CC, CI, \ldots, TT$) but only three phenotypes (*carbonaria*, *insularia*, and *typica*) are visibly measurable in field work.** That is, by sight, the genotypes CC, CI, CT all result in wings that have the same *carbonaria* coloring.

We are interested in estimating the allele frequencies from the phenotype counts under the assumption of Hardy-Weinberg. In particular, let p_C, p_I and $p_T = 1 - p_C - p_I$ be the frequencies/proportions of the three alleles. Then according to Hardy-Weinberg, we should have the following genotype frequencies (brackets) that sum to 1:

		Parent 2		
		$C \quad [p_c]$	$I \quad [p_I]$	$T \quad [p_T]$
Parent 1	$C \quad [p_c]$	$CC \quad [p_c^2]$	$CI \quad [p_C p_I]$	$CT \quad [p_C p_T]$
	$I \quad [p_I]$	$CI \quad [p_C p_I]$	$II \quad [p_I^2]$	$IT \quad [p_I p_T]$
	$T \quad [p_T]$	$CT \quad [p_C p_T]$	$IT \quad [p_I p_T]$	$TT \quad [p_T^2]$

Problem set-up

We want to estimate p_C, p_I, and p_T, the true frequencies of the *carbonaria*, *insularia*, and *typica* alleles, respectively. Let the unknown parameter vector be $\boldsymbol{\theta} = (p_C, p_I)$, since $p_T = 1 - p_C - p_I$ (this will make our lives easier later on). Suppose we observe n peppered moths. The observable data/random variables are $\mathbf{X} = (N_C, N_I, N_T)$, the counts for the three phenotypes (clearly $n = N_C + N_I + N_T$). Let $\mathbf{Y} = (N_{CC}, N_{CI}, N_{CT}, N_{II}, N_{IT}, N_{TT})$ be the missing data of the counts of each genotype. The capital N's are used to emphasize that the counts are random variables.

Note that the observed data $\mathbf{X}$ follow a Multinomial$(n, (p_C, p_I, p_T))$ distribution. We observe a single vector of counts (so instead of $X_1, \ldots, X_n$, we have a single $\mathbf{X}$). If we didn't assume Hardy-Weinberg, we could easily obtain MLE estimates for $\boldsymbol{\theta}$ using results from problem 5 on Problem Set 3. However, convince yourself that finding the MLEs in this three-allele case under Hardy-Weinberg is harder than the two-allele case in problem 3 on Problem Set 3.

Derivations

You will be guided through deriving the calculations for obtaining MLEs for $\boldsymbol{\theta}$ via EM.

E-step

1. First, briefly explain why the complete data $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ are equivalent to just $\mathbf{Y}$.

2. The complete data $\mathbf{Z}$ are also multinomial. Find the complete log-likelihood for $\boldsymbol{\theta} = (p_C, p_I)$ using the complete data $\mathbf{Z}$.

3. Given the observed data, the conditional distribution of the subset of $\mathbf{Z}$ that relate to the black *carbonaria* moths is also a multinomial. In order words,

$$(N_{CC}, N_{CI}, N_{CT})|\mathbf{X} = (n_C, n_I, n_T), \boldsymbol{\theta} \sim \text{Multinom}(< blank >, < blank >)$$

   Give the exact multinomial distribution (i.e. specify the parameters). Be careful that your probabilities sum to 1! (Remember/look up the property of multinomial conditional distributions.)

4. For a multinomial random variable, the expectation of the j-th element X_j is $\mathbb{E}[X_j] = np_j$. Use this fact to find $\mathbb{E}[N_{CC}|\mathbf{X} = (n_C, n_I, n_T), \boldsymbol{\theta}^{old}]$ in terms of n_C, p_C^{old}, and p_I^{old}.

5. Using similar logic as the previous two parts, find

   - $\mathbb{E}[N_{CI}|\mathbf{X} = (n_C, n_I, n_T), \boldsymbol{\theta}^{old}]$
   - $\mathbb{E}[N_{CT}|\mathbf{X} = (n_C, n_I, n_T), \boldsymbol{\theta}^{old}]$
   - $\mathbb{E}[N_{II}|\mathbf{X} = (n_C, n_I, n_T), \boldsymbol{\theta}^{old}]$
   - $\mathbb{E}[N_{IT}|\mathbf{X} = (n_C, n_I, n_T), \boldsymbol{\theta}^{old}]$

6. Why is $\mathbb{E}[N_{TT}|\mathbf{X} = (n_C, n_I, n_T), \boldsymbol{\theta}^{old}] = n_T$?

7. Finally! Find $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$. (Don't be scared! You've done all of the hard work already). Remember that the missing data are viewed as random variables in the expectation for $Q()$. For ease of notation, use the following shorthand:

$$n_{CC}^{old} \equiv \mathbb{E}[N_{CC}|\mathbf{X} = (n_C, n_I, n_T), \boldsymbol{\theta}^{old}]$$
$$n_{CI}^{old} \equiv \mathbb{E}[N_{CI}|\mathbf{X} = (n_C, n_I, n_T), \boldsymbol{\theta}^{old}]$$
$$\text{etc.}$$

M-step

Find the partial derivatives of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$ with respect to p_C and p_I, set them equal to 0, and solve to obtain your $\boldsymbol{\theta}^{new}$. No need to verify maximum here!

Implement

Navigate over to the provided `.Rmd` template for some starter code for implementing your EM algorithm!

Submission

Please submit your solutions to these derivations above alongside the knitted coding implementation.