

BIODATA ANALYSIS CA534

ASSIGNMENT 1

Deadline: End-Week 9; due 4.00p.m. November 25th

Instructions:

Students should work within their own small groups and produce a **single report** for that group, with the names of all contributors clearly indicated on the top sheet. A statement should also be included on the top sheet to indicate that members of the group have contributed equally (or reasonably so) to the work. (It is recognised that different students will bring different skill sets to the task to some extent). The mark given for the assignment will apply equally to all members of the group, unless an individual student fails to make a contribution (as claimed by his/her co-members) or feels that he/she has done the majority of the work. In such a case, each member of the group will be invited separately to defend his/her contribution by presentation to a panel of staff and students. Plagiarism across groups will be penalised.

Report Length:

The aim in producing the report is to provide answers to the specific points raised, describe methods and solutions and to draw conclusions in a clear and concise way. The report is not expected to exceed 10 A4 pages in length, inclusive of any analyses, background references or (short) code. Detailed code or listing of software commands should not be necessary, but if students feel that this is required to make a specific point, they can include a disk. The report and inclusions (e.g. analyses) should be typed. Large amounts of printout should not be appended. If further details are necessary to make a point, this should be indicated clearly in the text and the relevant appendix cited, which should then appear in a readily accessible and annotated format on the accompanying disk. Any printout which is included without annotation (i.e. clear description of what it purports to show), either in hard or electronic copy, will not be marked.

Background

Given a single locus, the number of different detectable forms in nature are the alleles. For controlled cross data, the number of alleles and their frequencies are typically known before experimentation or are easy to determine. Consequently while not contributing to likelihood information, such information is important for genomic analysis in natural populations. For example, in order to design efficient experiments, we need to discover how many alleles there are in nature. It is also valuable information in gene conservation, since the number of alleles and their distributions are indicators of genetic diversity. The measures of interest, several of which we have met with, include e.g. determining the number of alleles and their "frequencies" or probabilities, estimating the probability of heterozygosity (proportion of heterozygotes), the Hardy-Weinberg equilibrium concept and how to determine disequilibrium, as well as other features.

Following the notes, we use ℓ to denote the No. alleles for a single locus. In general, therefore, choosing a suitable (matrix) notation, we have

$$Alleles = \begin{bmatrix} A_1 \\ A_2 \\ \cdot \\ A_\ell \end{bmatrix} \quad P\{Alleles\} = \begin{bmatrix} p_1 \\ p_2 \\ \cdot \\ p_\ell \end{bmatrix}$$

We have seen that , for a diploid system, there exist $\ell(\ell+1)/2$ possible genotypes for a locus with ℓ alleles in nature, where

$$Genotypes = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1\ell} \\ A_{21} & A_{22} & \dots & A_{2\ell} \\ \cdot & & & \\ A_{\ell 1} & A_{\ell 2} & \dots & A_{\ell\ell} \end{bmatrix}$$

i.e. symmetric about diagonal. Then, the observed counts for corresponding genotypes in a sample from the population are given by:

$$n [Genotypes] = \begin{bmatrix} n_{11} & n_{12} & \dots & n_{1\ell} \\ n_{21} & n_{22} & \dots & n_{2\ell} \\ \cdot & & & \\ n_{\ell 1} & n_{\ell 2} & \dots & n_{\ell\ell} \end{bmatrix}$$

and the corresponding estimated frequencies are:

$$\hat{p} [\text{Genotypes}] = \begin{bmatrix} \hat{p}_{11} & \hat{p}_{12} & \dots & \hat{p}_{1\ell} \\ \hat{p}_{21} & \hat{p}_{22} & \dots & \hat{p}_{2\ell} \\ \dots & \dots & \dots & \dots \\ \hat{p}_{\ell 1} & \hat{p}_{\ell 2} & \dots & \hat{p}_{\ell\ell} \end{bmatrix} = \frac{1}{N} \begin{bmatrix} n_{11} & n_{12} & \dots & n_{1\ell} \\ n_{21} & n_{22} & \dots & n_{2\ell} \\ \dots & \dots & \dots & \dots \\ n_{\ell 1} & n_{\ell 2} & \dots & n_{\ell\ell} \end{bmatrix}$$

where Total sample size is

$$N = \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} n_{ij}$$

Clearly, for N individuals in a sample, there are $2N$ alleles in a diploid system. Again, frequency matrices is symmetric.

The observed genotypic count and estimated genotypic frequency distribution are starting points for most of the analyses to characterise a single locus in the population. The number of alleles and their frequencies are estimated from the distribution and the relationship between the genotypic distribution and allelic frequencies is commonly used to infer equilibrium status. For a single population, the genotypic frequency is the frequency of a given genotype in the population. The variance of the estimated genotypic frequency is the variance for a multinomial proportion. Clearly, genotypic frequency is the simple count in a sample from the population. Allelic frequencies can be estimated as:

$$\hat{p} [\text{Alleles}] = \begin{bmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \dots \\ \hat{p}_\ell \end{bmatrix} = \text{Diagonal} \left\{ \begin{bmatrix} \hat{p}_{11} & \hat{p}_{12} & \dots & \hat{p}_{1\ell} \\ \hat{p}_{21} & \hat{p}_{22} & \dots & \hat{p}_{2\ell} \\ \dots & \dots & \dots & \dots \\ \hat{p}_{\ell 1} & \hat{p}_{\ell 2} & \dots & \hat{p}_{\ell\ell} \end{bmatrix} \begin{bmatrix} 1 & 0.5 & \dots & 0.5 \\ 0.5 & 1 & \dots & 0.5 \\ \dots & \dots & \dots & \dots \\ 0.5 & 0.5 & \dots & 1 \end{bmatrix} \right\}$$

So, by element

$$\begin{aligned} \hat{p}_i &= \hat{p}_{ii} + 0.5 \sum_{j \neq i} \hat{p}_{ij} \\ &= \frac{1}{N} \left[n_{i\ell} + 0.5 \sum_{j \neq i} n_{ij} \right] \end{aligned}$$

Thus, the variance of an estimated frequency and the covariance between two estimated allelic frequencies follow.

When dominant markers are used, the heterozygote and one of the homozygous genotypes cannot be identified in one generation. It is usually necessary to use further testcross (or crosses) in advanced generations in order to distinguish the two genotypic classes). An algorithm, such as the Expectation-Maximisation algorithm can be used to estimate allelic frequencies under the assumption of Hardy-Weinberg equilibrium.

The expected frequencies of the three possible genotypes are as given in the notes - see below. What can be **observed** in the dominant case, is clearly a combination of these frequencies,

$$p_i^0 = p_{ii} + p_{ij}$$

where the superscript denotes either dominant or recessive allele . Then the probability that an allele is A_i **given** genotype is A_i^0 is:

$$P\{A_i | p_i^0\}$$

(For the two allele case, the frequency might also be obtained by getting an estimate for the recessive allele in terms of the counts of course).

In terms of **single allele detection**, the No. of alleles at a locus is usually estimated by initial screening, so,

$$P\{at\ least\ one\ individual\ with\ a\ certain\ allele\ in\ a\ sample\ of\ size\ N\} = \gamma_i = 1 - (1 - p_i)^{2N}$$

under H-W equilibrium, where second term is just the $P\{No\ genotype\ contains\ allele\ A_i\}$ Sample size for detecting an allele with frequency p_i and power of γ_i can obviously be obtained by manipulating this equation.

Clearly, the usual key question is not whether just one, but how many alleles can be detected and we can define a whole series of probabilities $\gamma_1, \gamma_2, \dots, \gamma_\ell$.

The average detection probability γ_m for detecting at least m alleles therefore is

$$\gamma_m = \frac{1}{C} \sum_{j=1}^{\ell} \prod_{i=1}^{k_j} \gamma_i \quad \ell \leq k_j \leq m \quad C = \sum_{k_j}^{\ell} \binom{\ell}{k_j} = No. \ of \ combinations$$

For a large number of alleles and unequal frequencies, this represents a lot of work. It is possible, however, to use simple Monte Carlo simulation with 1 indicating detection of allele and 0 non-detection. For allele i if a random uniform (0,1) is less than or equal to γ_i , the allele is detected and set $I_i=1$, -incremented by one each time. Clearly $I_i=0$ if not detected, (random uniform). Therefore, overall, No. of alleles detected = $\sum I_i$ and for the process repeated a large number of times γ_m is determined by counting frequencies, s.t.

$$\sum I_i \geq m = \hat{\gamma}_m = P\left\{\sum I_i \geq m\right\}$$

Data

Multiple allelic frequencies can be modelled using a geometric series, with parameter λ , s.t. $0 \leq \lambda \leq 1$ and the table below lists some examples.

Table: Allelic frequency distributed as geometric series, different λ

ALLELE $\lambda=1/3$ $\lambda=2/3$ $\lambda=9/11$ $\lambda=19/21$ $\lambda=39/41$

ALLELE ↓	$\lambda=1/3$	$\lambda=2/3$	$\lambda=9/11$	$\lambda=19/21$	$\lambda=39/41$
1	0.6667	0.3333	0.1818	0.0952	0.0488
2	0.2222	0.2222	0.1488	0.0862	0.0464
3	0.0741	0.1481	0.1217	0.0780	0.0441
4	0.0247	0.0988	0.0996	0.0705	0.0420
5	0.0082	0.0658	0.0815	0.0638	0.0399
6	0.0027	0.0439	0.0667	0.0577	0.0380
7	0.0009	0.0293	0.0545	0.0522	0.0361
8	0.0003	0.0195	0.0446	0.0473	0.0344
9	0.0001	0.0130	0.0365	0.0428	0.0327
10		0.0087	0.0299	0.0387	0.0311
11		0.0058	0.0244	0.0350	0.0296
12		0.0039	0.0200	0.0317	0.0281
13		0.0026	0.0164	0.0287	0.0268
14		0.0017	0.0134	0.0259	0.0255
15		0.0011	0.0110	0.0235	0.0242
16		0.0008	0.0090	0.0212	0.0230
17		0.0005	0.0073	0.0192	0.0219
18		0.0003	0.0060	0.0174	0.0208
19		0.0002	0.0049	0.0157	0.0198
20		0.0002	0.0040	0.0142	0.0189
21		0.0001	0.0033	0.0129	0.0179
22		0.0001	0.0027	0.0116	0.0171
23			0.0022	0.0105	0.0162
24			0.0018	0.0095	0.0154
25			0.0015	0.0086	0.0147
26			0.0012	0.0078	0.0140
27			0.0010	0.0071	0.0133
28			0.0008	0.0064	0.0126
29			0.0007	0.0058	0.0120
30			0.0005	0.0052	0.0114
31			0.0004	0.0047	0.0109
32			0.0004	0.0043	0.0104
33			0.0003	0.0039	0.0098
34			0.0002	0.0035	0.0094
35			0.0002	0.0032	0.0089
36			0.0002	0.0029	0.0085
37			0.0001	0.0026	0.0081
38			0.0001	0.0023	0.0077
39			0.0001	0.0021	0.0073
40			0.0001	0.0019	0.0069
41			0.0001	0.0017	0.0066
42				0.0016	0.0063
43				0.0014	0.0060
44				0.0013	0.0057
45				0.0012	0.0054
46				0.0011	0.0051
47				0.0010	0.0049
48				0.0009	0.0046
49				0.0008	0.0044
50				0.0007	0.0042

Questions and Analysis

1. From the background described, give expressions for the variance of an estimated allelic frequency and the covariance between two estimated allelic frequencies. Comment for a population in equilibrium.
2. Give expressions for the frequencies of alleles A_i and A_i **conditional** on the **observable** genotypic frequencies. These will be of the general form, $P\{A_i / p_i^0\}$, $P\{A_j / p_i^0\}$, $P\{A_i / p_{jj}\}$, $P\{A_j / p_{jj}\}$ as noted above.
3. Research and discuss (briefly) what is involved in the Expectation-Maximisation algorithm and indicate how you might use this to estimate the allelic frequencies as suggested. What information would you need?
4. For the data in the table of geometric series given (and ascribed to No. of alleles), obtain the most frequent allele for each value of the geometric parameter from the theoretical expression. What does this imply?
 - For the distributions of allelic frequencies given in the table and different screening sample sizes, and using a simple Monte Carlo simulation, such as that described or a suitable alternative method, sketch- from the results of a large number of such experiments, -the empirical statistical power (of detecting numbers of alleles) vs sample size. Comment on the type of curves and the effect of sample size. What do you conclude?
 - Give in each case the No. of alleles corresponding approximately to 50%, 70%, 80%, 90%, 95% of alleles distributed as in the table. For example 70% of alleles having frequency higher than 0.001 is $m=19$ for the distribution corresponding to $\lambda=9/11$. The number of loci with frequency higher than 0.01 for the same λ is 26. What is the detection probability in each case from the simulation experiments? Hence, give the detection probabilities for 70% and 95% of alleles (with similar frequencies) from the different allelic frequency distribution patterns. What do you notice?
 - For a fixed detection probability of 0.8 (80%), what guidelines can you give in terms of sample size needed and No. alleles detected for the different λ . How is this likely to change if lower or higher power is required and a large number of alleles have equal frequency, e.g. 20 or 50 or 100 alleles say?
 - Given that one of the most important characteristics of a locus is its heterozygosity, obtain the heterozygosity and Polymorphic Information Content PIC estimates for the allelic frequency distributions in the table, where

$$PIC = 1 - \sum_{i=1}^{\ell} p_i^2 - 2 \sum_{i=2}^{\ell} \sum_{j=1}^{i-1} (p_i^2 p_j^2)$$
