4.9 Problems

1. Test for Hardy-Weinberg equilibrium in the MN Syrian data presented in Chapter 2.

2. Table 4.4 lists frequencies of coat colors among cats in Singapore [32]. Assuming an X-linked locus with two alleles, estimate the two allele frequencies by gene counting. Test for Hardy-Weinberg equilibrium using a likelihood ratio test.

<table>
<thead>
<tr>
<th>Females</th>
<th></th>
<th>Males</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dark t/t</td>
<td>Calf t/y</td>
<td>Yellow y/y</td>
</tr>
<tr>
<td></td>
<td>63</td>
<td>55</td>
<td>12</td>
</tr>
</tbody>
</table>

3. Let \((N_1, \ldots, N_m)\) be the outcome vector for a multinomial experiment with \(n\) trials and \(m\) categories. Prove that

\[
\Pr(N_1 \leq t_1, \ldots, N_m \leq t_m) \leq \prod_{i=1}^{m} \Pr(N_i \leq t_i) \tag{4.7}
\]

\[
\Pr(N_1 \geq t_1, \ldots, N_m \geq t_m) \leq \prod_{i=1}^{m} \Pr(N_i \geq t_i) \tag{4.8}
\]

for all integers \(t_1, \ldots, t_m\). If all \(t_k = 0\) in (4.8) except for \(t_i\) and \(t_j\), conclude that

\[
\Pr(N_i \geq t_i, N_j \geq t_j) \leq \Pr(N_i \geq t_i) \Pr(N_j \geq t_j)
\]
as stated in the text. (Hints: It suffices to show that (4.7) holds when \( n = 1 \) and that the set of random vectors satisfying (4.7) is closed under the formation of sums of independent random vectors. For (4.8) consider the vectors \(-N_1, \ldots, -N_m\).)

4. Using the Chen-Stein method and probabilistic coupling, Barboza et al. [4] show that the statistic \( W_d \) satisfies the inequality

\[
\sup_{d \in \mathcal{N}} |\Pr(W_d \in A) - \Pr(Z \in A)| \leq \frac{1 - e^{-\lambda}}{\lambda} |\lambda - \Var(W_d)|,
\]  

(4.9)

where \( Z \) is a Poisson random variable having the same expectation \( \lambda = \sum_{i \in \mathcal{N}} \mu_i \) as \( W_d \) and where \( \mathcal{N} \) denotes the set \( \{0, 1, \ldots\} \) of nonnegative integers. Prove that

\[
\lambda - \Var(W_d) = \sum_i \mu_i^2 - 2 \sum_{i < j} \text{COV}(1_{\{N_i > d\}}, 1_{\{N_j > d\}}).
\]

In view of Problem 3, the random variables \( 1_{\{N_i > d\}} \) and \( 1_{\{N_j > d\}} \) are negatively correlated. It follows that the bound (4.9) is only useful when the number \( \lambda^{-1}(1 - e^{-\lambda}) \sum \mu_i^2 \) is small. What is the value of \( \lambda^{-1}(1 - e^{-\lambda}) \sum \mu_i^2 \) for the hemoglobin data when \( d = 27 \)? Careful estimates of the difference \( \lambda - \Var(W_d) \) are provided in [4].

5. Consider a multinomial model with \( m \) categories, \( n \) trials, and probability \( p \) attached to category \( i \). Express the distribution function of the maximum number of counts \( \max_i N_i \) observed in any category in terms of the distribution functions of the \( W_d \). How can the algorithm for computing the distribution function of \( W_d \) be simplified to give an algorithm for computing a \( p \)-value of \( \max_i N_i \)?

6. Continuing Problem 5, define the statistic \( U_d \) to be the number of categories \( i \) with \( N_i < d \). Express the right-tail probability \( \Pr(U_d > j) \) in terms of the distribution functions of the \( W_d \). This gives a method for computing \( p \)-values of the statistic \( U_d \). In some circumstances \( U_d \) has an approximate Poisson distribution. What do you conjecture about these circumstances?

7. The nonparametric linkage test of de Vries et al. [10] uses affected sibbing data. Consider a nuclear family with \( s \) affected sibs and a heterozygous parent with genotype \( a/b \) at some marker locus. Let \( n_a \) and \( n_b \) count the number of affected sibs receiving the \( a \) and \( b \) alleles, respectively, from the parent. If the other parent is typed, then this determination is always possible unless both parents and the child are simultaneously of genotype \( a/b \). de Vries et al. [10] suggest the statistic \( T = |n_a - n_b| \). Under the null hypothesis of independent transmission of the disease and marker genes, Barth et al. [8] show that \( T \) has

\[ 
\text{mean and var} \]

should be appropriate because between the markers.

8. To compute nonparametric linkage for \( s \) even and small, use \( u \) to be a falling factorial, integers indexed by \( j \), and \( u_j = \sum \lambda^{u_j} \).

In particular, verify:

9. Verify the mean and item 8. Alternatively, and calculate the \( n \) methods give the same two components. Set one equal to 1 or 2. The 4th indicator case is a case and has gen.
mean and variance

$$E(T) = \begin{cases} s \left( \begin{array}{c} \frac{1}{2} \end{array} \right) + \left( \frac{1}{2} \right) & \text{if } s \text{ is even} \\ s \left( \begin{array}{c} \frac{3}{2} \end{array} \right) - \left( \frac{3}{2} \right) & \text{if } s \text{ is odd} \end{cases}$$

$$Var(T) = s - E(T)^2.$$

Prove these formulas. If there are $n$ such parents (usually two per family), and the $i$th parent has statistic $T_i$, then the overall statistic

$$\frac{\sum_{i=1}^{n} T_i - E(T_i)}{\sqrt{\sum_{i=1}^{n} Var(T_i)}}$$

should be approximately standard normal. A one-sided test is appropriate because the $T_i$ tend to increase in the presence of linkage between the marker locus and a disease predisposing locus. (Hint: The identities

$$\sum_{i=0}^{s-1} \left( \begin{array}{c} s \\ i \end{array} \right) = 2^{s-1} - \left( \frac{s}{2} \right)$$

$$\sum_{i=0}^{s-1} \left( \begin{array}{c} s \\ i \end{array} \right) = s \left( 2^{s-2} - \left( \frac{s-1}{2} \right) \right)$$

for $s$ even and similar identities for $s$ odd are helpful.)

8. To compute moments under the Fisher-Yates distribution (4.4), let

$$\nu^r = \left\{ \begin{array}{ll} u(u-1) \cdots (u-r+1) & r > 0 \\ 1 & r = 0 \end{array} \right.$$ 

be a falling factorial power, and let $T_i$ be a collection of nonnegative integers indexed by the haplotypes $i = (i_1, \ldots, i_m)$. Setting $l = \sum_i l_i$ and $l_{ij} = \sum_i 1(i_j = k)$, show that

$$E\left( \prod_{j=1}^{m} n_j^{l_{ij}} \right) = \frac{\prod_{j=1}^{m} l_j^{n_j} l_{ij}^{l_{ij}}}{\nu^{l_{ij}}}.$$ 

In particular, verify that $E(n_j) = n_i$.

9. Verify the mean and variance expressions in equation (4.6) using Problem 8. Alternatively, write $c_j$ as a sum of indicator random variables and calculate the mean and variance directly. Check that the two methods give the same answer. (Hints: In applying Problem 8, I has two components. Set all but one of the $l_i$ equal to 0. Set the remaining one equal to 1 or 2 to get either a first or second factorial moment. The $l$th indicator random variable indicates whether the $k$th person in the case and has genotype $j$.)
4. Hypothesis Testing and Categorical Data

A genotypic phenotypes \( x \) unrelated people at each of \( n_i \) loci with codominant alleles and records a vector \( i = (i_1, i_2, \ldots, i_{n_i}) \) of genotypes for each person. Because phase is unknown, it cannot be rescored into haplotypes. The data gathered can be summarized by the number of people \( n_i \) counted for each genotype vector \( i \). Let \( n_{jk} \) be the number of alleles of type \( k \) at locus \( j \) observed in the sample, and let \( n_i \) be the total number of heterozygotes observed over all loci. Assuming genetic equilibrium, prove that the distribution of the counts \( n_i \) conditional on the allele totals \( n_{jk} \) is

\[
Pr((n_1) \mid (n_{jk})) = \left( \frac{n_{jk}}{n_i} \right)^{n_{jk}} \prod_{j=1}^{n_i} \left( \frac{n_{jk}}{n_{jk}} \right).
\]

(4.10)

The moments of the distribution (4.10) are computed in [23]; just as with haplotype count data, all allele frequencies cancel.

11. Describe and program an efficient algorithm for generating random permutations of the set \( \{1, \ldots, n\} \). How many calls of a random number generator are involved? How many interchanges of two numbers? You might wish to compare your results to the algorithm in [27].

12. Describe and program a permutation version of the two-sample t-test. Compare it on actual data to the standard two-sample t-test.

4.10 REFERENCES

[10] de Vries RP, HLA-linked g. Lancet 2:1230-1231
[12] Evans WJ, G, Statistiaal anal. the x2 test, C.
[16] Hanash SM, Bo random dis. cultured human 169