Biostatistics 666
Problem Set 1

1. **Consider a population where allele frequencies differ between the sexes. Assume that there are equal numbers of males and females and that genotypes occur in Hardy-Weinberg proportions within each sex. Focus on a single di-allelic marker in this population. The marker has allele frequency p_M = $p + \Delta$ in males and $p_F = p - \Delta$ in females, where $p=(p_F + p_M)/2$.**

   a) **Calculate offspring genotype frequencies after one generation of random mating.**

   After one generation of random mating, the genotype frequencies will be $p^2 - \Delta^2$ and $(1 - p)^2 - \Delta^2$ for homozygotes and $2p(1 - p) + 2\Delta^2$ for heterozygotes.

   There are several different ways to arrive at the solution. One possibility is to list all mating type frequencies and then, for each one, calculate the probability of each type of offspring being generated.

   Another simpler possibility is to note that a male and a female gamete must combine to generate each offspring. Then, we can calculate the probability of each offspring genotype simply as the probability of sampling its constituent alleles, one from a female and another from a male. For example, to obtain a homozygous genotype for the allele of interest, the probability is simply $p_F p_M = (p - \Delta)(p + \Delta) = p^2 - \Delta^2$.

   b) **How do genotype frequencies differ from those expected under Hardy-Weinberg equilibrium?**

   Compared to HWE expectations, there is an excess of heterozygotes ($2\Delta^2$) and a deficit of homozygotes ($-\Delta^2$ for each class)

   c) **How many additional generations are required before Hardy-Weinberg equilibrium is reached?**

   Since there is no longer a sex difference in allele frequencies, a single additional generation of random mating should ensure Hardy-Weinberg equilibrium is reached.

2. **In a sample of 100 individuals, 97 homozygotes for allele A, 2 homozygotes for allele B and 1 heterozygote were observed. Conditional on the number of observed A and B alleles, answer the following questions:**

a) **What is the probability of this particular sample configuration?**

The probability of any particular configuration can be calculated using the formula below:

$$P(n_{AB} = 1 | n_A = 195, n_B = 5, n = 100) = \frac{2^{n_{AB}} n!}{n_{AA}! n_{AB}! n_{BB}!} \cdot \frac{n_A! n_B!}{(2n)!}$$

The following table gives the probability of various possible configurations, each calculated using the formula above.

n_{AA}	n_{AB}	n_{BB}	Probability
97	1	2	.0003826
95	3	1	.0494860
93	5	0	.9501314

Thus, the probability of the observed configuration is .0003826.

b) **What is the probability of observing an equal or greater number of heterozygotes?**

From the table in Part A, we get:

$$P(n_{AB} \geq 1 | n_A = 195, n_B = 5, n = 100) = 1 - P(n_{AB} = 0 | n_A = 195, n_B = 5, n = 100) = 1$$

c) **What is the probability of observing a smaller number of heterozygotes?**

$$P(n_{AB} < 1 | n_A = 195, n_B = 5, n = 100) = P(n_{AB} = 0 | n_A = 195, n_B = 5, n = 100) = 0$$

d) **What is the chi-squared statistic for Hardy-Weinberg equilibrium?**

Here, we first estimate sample allele frequencies

$$n = n_{AA} + n_{AB} + n_{BB}$$

$$p_A = \frac{2n_{AA} + n_{AB}}{n} = 0.975$$

$$p_B = \frac{2n_{BB} + n_{AB}}{n} = 0.025$$

Then, we estimate expected counts asssuming Hardy-Weinberg equilibrium:

$$E(n_{AA}) = np_A^2 = 95.0625$$

$$E(n_{BB}) = np_B^2 = 0.0625$$

$$E(n_{AB}) = 2np_A p_B = 4.875$$

Finally, we compare observed and expected counts to obtain a chi-squared statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 63.18$$

3. **Consider two loci in disequilibrium in a large population. Assume that the recombination fraction between the two loci is 0.0001. In how many generations do you expect the disequilibrium coefficient D to be halved?**

Here, we can use the very simple expectation for the decay of D, which is:

$$D_{AB}^t = (1 - \theta)^t D_{AB}^0$$

We have to solve:

$$(1 - \theta)^t = 0.5$$

Which gives:

$$t = \frac{log(0.5)}{log(1 - 0.0001)}$$

And finally:

$$t = 6931.13 \text{ or } \sim 6932 \text{ generations}$$

4. **Consider the following set of haplotype frequencies:**

$p_{AB} = 0.4$, $p_{Ab} = 0.2$; $p_{aB} = 0.2$; $p_{ab} = 0.2$

a) **Calculate the frequency of alleles A, a, B, and b.**

The estimated allele frequencies are $p_A = 0.6$, $p_a = 0.4$, $p_B = 0.6$, $p_b = 0.4$.

b) **Calculate D, D' and Δ^2 between the two markers.**

$D_{AB} = p_{AB} - p_A p_B = 0.40 - 0.36 = 0.04$

Since $D_{AB} > 0$, we have

$$D'_{AB} = \frac{D_{AB}}{min(p_A p_b, p_a p_B)} = \frac{.04}{.24} = .167$$

And finally:

$$\Delta^2 = \frac{D_{AB}^2}{p_A p_a p_B p_b} = 0.0278$$

c) **What is the probability that allele A will be present in a chromosome that carries allele b?**

To solve this, we need to consider the frequency of chromosomes carrying the b allele, P_b, and the frequency of chromosomes carrying alleles A and b, P_{Ab}. Thus:

$$P(A|b) = \frac{P(Ab)}{P(b)} = \frac{.2}{.4} = 0.5$$

d) **What is the maximum possible value of r^2 for this marker pair?**

This will occur when D_{AB} takes its maximum or minimum value. So, we first consider the bounds of D_{AB}, which are:

$$max(-P_A P_B, -p_a p_b) < D_{AB} < min(p_A p_b, P_a P_B)$$

This gives

$$-0.36 < D_{AB} < 0.24$$

So that the maximum value of Δ^2 will occur when D is 0.36 (this is the larger absolute value and maximizes D^2 and, therefore Δ^2).

$$\Delta_{max}^2 = \frac{|D_{AB}^2|max}{p_A p_a p_B p_b} = 1$$

5. **The BRAVO browser (https://bravo.sph.umich.edu) lists variants and allele frequency information for many genes. Pick a gene whose name starts with the same initial as your last name. Download frequency information for missense variants in the gene and plot a histogram to summarize the data.**

> There should be a large excess of very rare variants. For the typical gene, nearly 50% of variants are seen in only one individual and have frequency of ~1 x 10^{-5}