STAT3306 A1

Hugo Burton

August 2024

Student Number: s4698512
Due: Friday 30th August 2024 @ 5pm

Question 1 [1 Mark]

What does the acronym SNP stand for?

SNP in the context of genetics stands for single nucleotide polymorphism.

Question 2 [2 Marks]

Describe two important aspects of meiosis that generate variation amongst offspring.

Note the distinction between mitosis and meiosis: the production of gametes. During meiosis, the aligned chromosomes from the mother and father can cross over at a random point along the arms of said chromosomes in a process known as recombination. Note this doesn't occur all the time, but when it does, a new combination of alleles is the result. Genetic code can be thought of as binary (base 2), and for example in humans, there are are 23 pairs of chromosomes, leading to 2^{23} theoretical combinations of alleles. It is worth noting the number of combinations may be lower than this due to genetic linkage, but nevertheless, this is one reason for variation amongst offspring.

Secondly, independent assortment (also part of meiosis), contributes to genetic variation in offspring. This is the process by which chromosome pairs are randomly distributed among cells to form haploids. That is, each cell will contain a random subset from both the mother and the father, leading to variation in offspring.

Question 3 [3 Marks]

Which of the people in the table below could be Luke's father? Explain your choice.

	SNP 1	**SNP 2**	**SNP 3**	**SNP 4**	**SNP 5**
Luke	A/A	C/T	T/T	A/C	G/G
Anakin	A/A	C/T	T/C	A/C	C/C
Yoda	A/A	T/T	T/T	A/A	G/C
Obi-Wan	A/T	C/C	C/C	A/A	G/C
Han	A/A	C/C	C/C	A/A	G/G

Yoda.

Observe Luke's father is highly unlikely to be Yoda, Obi-Wan or Han because in SNP 3, both loci are C, however, Luke has both T. Hence even if the mother had C, Luke's outcome would theoretically not be possible.

For each SNP, Yoda shares at least one locus with Luke, and hence it is possible he could his father.

Question 4 [7 Marks]

You are studying a genetic locus in a lizard that appears to have an effect on their reproductive success. You have genotyped 500 lizards at the locus and obtained the following counts.

Genotype	AA	AB	BB
Count	15	230	255

Table 1: Allele Counts for Lizards

Part A [2 Marks]

What is the frequency of the A and B alleles?

We can calculate the frequency by simply dividing the count by the total for each locus.

$$p_A = \frac{2 \cdot 15 + 230}{2 \cdot 500} = 0.26 \tag{1}$$
$$p_B = 1 - p_A \tag{2}$$
$$= 1 - 0.26 \tag{3}$$
$$= 0.74 \tag{4}$$

	A	B	N_{total}
Freq. Obs	0.26	0.74	500

Table 2: Allele Frequencies for Lizards

Part B [4 Marks]

Test the genotype frequencies for deviation from Hardy-Weinberg equilibrium
The Hardy-Weinberg equilibrium states

$$p_{AA} = p_A^2;$$
$$p_{AB} = 2p_A\left(1 - p_A\right) = 2p_A p_B;$$
$$p_{BB} = p_B^2$$

We can test the genotype frequencies for deviation by performing a χ^2 test with 1 degree of freedom on the data in table 1.

$$p_A = \frac{2 \cdot 15 + 230}{2 \cdot 500} = 0.26 \tag{5}$$
$$p_B = 1 - p_A \tag{6}$$
$$= 1 - 0.26 \tag{7}$$
$$= 0.74 \tag{8}$$
$$p_{AA} = \frac{15}{500} \tag{9}$$
$$= 0.03 \tag{10}$$
$$p_{AB} = \frac{230}{500} \tag{11}$$
$$= 0.46 \tag{12}$$
$$p_{BB} = \frac{255}{500} \tag{13}$$
$$= 0.51 \tag{14}$$

Now compute the expected value

$$\begin{aligned}
\mathbb{E}[n_{AA}] &= p_A^2 \cdot N_{\text{total}} && (15)\\
&= 0.26^2 \cdot 500 && (16)\\
&= 0.0676 \cdot 500 && (17)\\
&= 33.8 && (18)\\
\mathbb{E}[n_{AB}] &= 2p_A p_B \cdot N_{\text{total}} && (19)\\
&= 2 \cdot 0.26 \cdot 0.74 \cdot 500 && (20)\\
&= 2 \cdot 0.1924 \cdot 500 && (21)\\
&= 192.4 && (22)\\
\mathbb{E}[n_{BB}] &= p_B^2 \cdot N_{\text{total}} && (23)\\
&= 0.74^2 \cdot 500 && (24)\\
&= 0.5476 \cdot 500 && (25)\\
&= 273.8 && (26)
\end{aligned}$$

and compute the test statistic

$$\begin{aligned}
\chi_1^2 &= \frac{(15-33.8)^2}{33.8} + \frac{(230-192.4)^2}{192.4} + \frac{(255-273.8)^2}{273.8} && (27)\\
&= 19.096 && (28)
\end{aligned}$$

Taking pchisq we obtain a p-value of 0.9999876, meaning a significant deviation from the HWE.

Part C [1 Mark]

What Hardy-Weinberg assumption is being violated in this locus?

HW assumes an infinite population size, and random mating. However, we are limited to $n = 500$ lizards in this case. Hence we cannot assume mating is random, as there are a finite number of mates for each lizard.

This small population exacerbates the fact that the locus has in question has an impact on reproductive success. In other words, our selection is not random since the genotype which affects reproduction has far less change of survival in the evolution of the lizards.

Question 5 [1 Mark]

Define the broad and narrow sense heritability of a trait.

We know from the simplistic model of $P = G + E$ that a phenotype is comprised of both genotypic value (from genes) as well as environmental factors. The heritability of a trait refers to the influence genetics has on differences between individuals in a given population.

With this in mind, broad sense heritability refers to the proportion of phenotypic variation that is explained by **all** forms of genetic variance. In lectures we covered the following forms

1. **Additive Variance**. We know that additive **value** is the sum of individual alleles which combined have influence over a trait. E.g. in the lectures, we explored the example of height where each allele might add 1 or 2 cm to a base height. Hence we say this is additive since the sum of the individual effects from each allele has influence over an individual's height. Hence, the additive genetic variance quantifies the variance of these additive genetic values.
2. **Dominance Variance**. Dominance works differently from additive value in genetics. Some traits are not simply the result of the sum of individual allele effects. Rather, sometimes a particular allele can block or dominate the effect of another allele. In this way, the genotype is different than if the alleles were to be modelled additively. Hence the dominance variance is the deviation due to these non-linear interactions.

Mathematically, we express broad sense heritability as

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2}$$

where (as from lectures),

$$\sigma_G^2 = \sum_{\ell=1}^{L} \text{Var}\left(\alpha\left[A_{i\female}^{(\ell)}\right] + \alpha\left[A_{i\male}^{(\ell)}\right]\right) + \sum_{\ell=1}^{L} \text{Var}\left(\delta\left[A_{i\female}^{(\ell)} A_{i\male}^{(\ell)}\right]\right).$$

On the other hand, narrow sense heritability refers only to the proportion of phenotypic (total) variance caused by additive **genetic** variance. See above for the what additive genetic variance is. Mathematically, we express narrow sense heritability as

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}$$

where

$$\sigma_A^2 = \sum_{\ell=1}^{L} \text{Var}\left(\alpha\left[A_{i\female}^{(\ell)}\right] + \alpha\left[A_{i\male}^{(\ell)}\right]\right).$$

Question 6 [4 Marks]

You are given a large number of full-subs pairs, all measured for height. Explain one method for estimating the heritability of height. What are the limitations of this approach to estimating heritability?

Observe we only know the heights of the offspring, so we cannot use the parent's height in our estimation. With this in mind, we can estimate the heritability of height using the variance of height between the sibling pairs and the total variance of the sample.

Begin with the covariance between two (full) siblings in the following derivation where we substitute for the coefficients of ancestry and fraternity. Here $\mathcal{S}$ refers to the set of siblings.

$$\begin{aligned}
\text{Cov}\left(G_i, G_j\right) &= 2\Theta_{ij}\sigma_A^2 + \Delta_{ij}\sigma_D^2 && (29)\\
&= 2\Theta_{ij}\sigma_A^2 + \Theta_{ij}^2\sigma_D^2 && (30)\\
&= \frac{1}{4}\cdot 2\sigma_A^2 + \frac{1}{4}\cdot\sigma_D^2 && (31)\\
&= \frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4} && (32)\\
&\approx \frac{1}{N_{\text{pairs}}}\sum_{k=1}^{N_{\text{pairs}}}\left((x_{ik}-\bar{x})(x_{jk}-\bar{x})\right), \qquad \text{for each } (i,j)\in\mathcal{S} && (33)
\end{aligned}$$

Now, derive the correlation of height between **each** sibling pair $(i, j) \in \mathcal{S}$ as

$$\begin{aligned}
\text{Corr}\left(G_i, G_j\right) &= \frac{\frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4}}{\sigma_P^2} && (34)\\
&= \frac{h^2}{2} && (35)\\
\Rightarrow h^2 &= 2\cdot\text{Corr}\left(G_i, G_j\right) && (36)\\
&= 2\cdot\frac{\frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4}}{\sigma_P^2} && (37)\\
&= \frac{\sigma_A^2 + \frac{\sigma_D^2}{2}}{\sigma_P^2} && (38)
\end{aligned}$$

where σ_P^2 is the overall or **total** phenotypic variance. As we have a large sample, it is sufficient to use the following estimation.

$$\hat{\sigma}_P^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2, \qquad N \text{ being the total population, not sibling pairs}$$

Limitations are

- we assume the full siblings are sharing a common environment; and
- we assume dominance variance for height is negligible - this is a reasonable assumption given we can observe height amongst humans follows a normal distribution and therefore is more likely to be a result of additive genetic value. Nonetheless, I've left σ_D^2 in the formula above even though we assume it's contribution is far outweighed by σ_A^2.

Question 7 [2 Marks]

Your friend is generating SNP array data on a human cohort. He tells you that it is a small dataset, so he is going to perform the analysis on his laptop. To keep the data anonymous, he has used IDs based on the individuals initials and date of birth. What two ethical issues to you observe with their? (his) analysis plan.

While names aren't included in the data, it still contains PII from the initials and date of birth. For example, if the list of names were to ever be leaked, they could be matched with this dataset. There could be two participants with the same DOB and initials, however, it is unlikely in a small dataset. Hence such a leak would effectively make the data non-anonymous once again.

Even if the data were to be anonymised properly, this process still needs to be done somewhere. If he's performing the anonymisation on his laptop, this means the original dataset is still present. Hence it remains unethical to store such a dataset on a personal laptop.

Question 8 [3 Marks]

Part A

Define linkage disequilibrium
Linkage disequilibrium is the non-random association between alleles at different loci. In simple terms, two loci which are close together on the chromosome are more likely to originate from one parent because there is less chance the recombination split will occur between two close together points. Conversely, two loci which are separated by a greater distance are more likely to be separated by recombination because there is more or a chance for a split to occur over said greater distance.

Part B

The linkage disequilibrium (LD) between loci A and B is 0.74, while the LD between A and C is 0.35. Does this inform us of the relative distance of loci B and C to A? Explain your answer.

Yes it does. While we cannot know the absolute distance of either, this information does tell us about the relative distance. That is, a lower LD of 0.35 indicates more recombination, and in turn indicates loci A and C are further apart on the chromosome than loci A and B.

Question 9 [7 Marks]

In Lecture 3 "Inheritance of Genetic Information", we saw that it takes the combination of only a few loci (in addition to environmental variance) to generate a quantitative trait distribution. Demonstrate this by simulating a trait using R with the following conditions:

1. Generate 200 diploid individuals
2. The trait is controlled by 15 independent loci.
3. Allele frequency at a locus is drawn from a Uniform$(0, 1)$ distribution
4. The effect size at each locus is drawn from a Normal distribution.
5. The heritability of the trait is 0.5.

Provide a histogram of the simulated trait distribution, and the R code used to generate it.

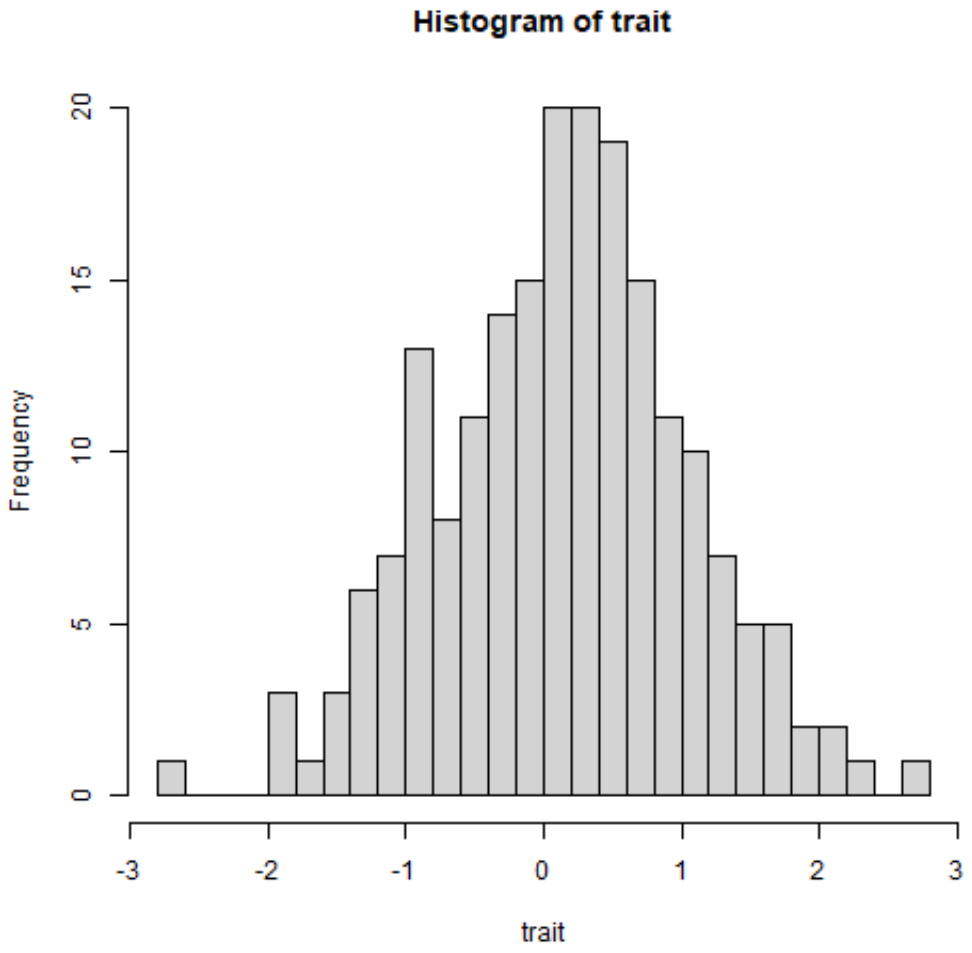

Figure 1: Simulated trait distribution with bin width 20

R code provided on the following page

```r
loci <- 15
indiv <- 200
h2 <- 0.5

af <- runif(loci)

x <- t(replicate(indiv, rbinom(loci, 2, af)))

# Check polymorphism
polymorphic <- apply(x, 2, var) == 0
num_polymorphic <- sum(polymorphic)
if (num_polymorphic != 0) {
  stop("There are ", num_polymorphic, " monomorphic loci")
}

x <- scale(x)

beta <- rnorm(loci, 0, sqrt(h2/loci))
genetic <- x %*% beta

# Environmental is per person
environmental <- rnorm(indiv, 0, sqrt(1 - h2))

trait <- genetic + environmental

# Save histogram to file
filename <- "histogram.png"
png(filename)
hist(trait, breaks = 20)
dev.off()
```