Medical Data Science
University of Tuebingen
WS 2023/2024
Karahan Sarıtaş - 6661689
Atahan Özer - 6317973

Assignment I

Problem 1

(a) GWAS (Genome-wide association studies) is a method in the field of medical data science that investigates the connections between genetic variations and uncovers links between these variations and medical conditions. GWAS can be used to detect if adult height is causally associated with the risk of lung cancer. (Apparently, it is [1])

(b) SNP (Single Nucleotide Polymorphism) refers to a singular alteration in a particular nucleotide within DNA. In theory, we can have one of A/C/G/T at any position but in practice, a specific position in the genome presents only two available options in the human population.

(c) An allele is one of the two variants in a particular nucleotide within DNA. *Major* allele occurs in a majority of the population (∼80%), whereas the *minor* allele occurs in a minority of the population (∼20%). The presence of different alleles leads to the genetic diversity of a population.

(d) SNPs within close proximity tend to be correlated with each other. Therefore we can predict the other SNPs with a very high accuracy if we know the SNP at a certain position. Therefore we can visualize blocks of SNP linkages by just examining a single SNP in the block - which reduces the required amount of data for the operation.

(e) In *classification* tasks, we aim to map features to predefined labels, while in *regression* tasks, we aim to map features to real values. For instance, classifying whether a patient will be admitted to the hospital is a classification task, and predicting the duration of their stay is a regression task. Predicting the price of a house is typically framed as a regression problem, where we seek to forecast a real-number value based on the house's relevant features, but it can be treated as a classification task if predefined price categories (or specific intervals) exist in the dataset.

(f) For every $\alpha \in (0,1)$, there is a size α test with rejection region R_a. Then,

$$\text{p-value } = \inf\left\{\alpha : T\left(X^n\right) \in R_\alpha\right\}.$$

That is, the p-value is the smallest level at which we can reject H_0 [3]. Informally, It quantifies the probability of observing a test statistic as extreme as, or more extreme than, the one obtained from a sample, assuming that the null hypothesis is true. We reject the null hypothesis when the p-value is less than or equal to the significance level, which is a predetermined threshold used to define the critical region.

(g) *Multiple testing* involves the simultaneous analysis of multiple statistical tests. As the number of tests conducted increases, the expected number of false positives (rejecting the null hypothesis incorrectly) increases. Therefore we have to find a way to compensate for the increasing probability of false rejections. *Bonferroni correction* and *False Discovery Rate* are two methods for that. In the Bonferroni correction, the p-value is adjusted by dividing it by the number of tests, effectively reducing the likelihood of rejections. On the other hand, in the context of the False Discovery Rate, the focus is on estimating the expected number of false rejections within the set of all rejections made.

(h) *Confounding factors* are variables that are not the main focus of our analysis but can influence the relationship between the variables of interest and the outcome. For example, if we want to measure the relationship between genetics and the probability of having twins at birth, there are possible confounding factors such as age, or previous pregnancies. If confounding factors exist, the p-values might be inflated or deflated. P-values can be calibrated with the use of genomic control. It has the potential to decrease the occurrence of false positives while possibly leading to an increase in false negatives.

(i) In our context, *homozygous* refers to having same type of alleles such as aa or AA - either minor or major, whereas *heterozygous* refers to having different types of alleles such as Aa. *Cochran-Armitage test* is used to test the association between the risk allele and the disease assuming that the risk is additive. It examines whether there is a linear trend in the disease risk as genotypes move from 0 to 2, where the y axis represents the disease risk $= \frac{\text{\# cases}}{\text{\# cases+\# controls}}$. The main problem with this test method is that it cannot detect the cases where the higher risk of getting the disease is achieved with *heterozygous* alleles (non-additive effect).

(j) A *Linear Mixed Model*, often referred to as a mixed-effects model, extends the simple linear model ($y = X\beta + \epsilon$) by incorporating additional random effects into the equation ($y = X\beta + u + \epsilon$). Variance of these additional random effects is proportional to the pairwise genotypic similarity of individuals. This way, in contrast to simple linear models, we can include the genetic relationships in our analysis.

(k) $O(MN^2 + N^3)$ was achieved by approximating the variance components beforehand instead of computing them every time for each individual. The original FaST-LMM, which is one of the methods to approximate the variance component, uses a subset of equally spaced SNPs given the fact that SNPs within close proximity are correlated with each other. Later on, it was shown that the same speed-up can be achieved with the exact solution as well.

Problem 2

(a) (i) Frequency of allele A:

$$\frac{299 * 2 + 490}{2000} = 0.544$$

Frequency of allele G:

$$\frac{211 * 2 + 490}{2000} = 0.456$$

(ii) The expected number of individuals of the genotype AA:

$$p_1^2 n = (0.544)^2 = 296$$

The expected number of individuals of the genotype AG:

$$2p_1 p_2 n = 2 \cdot (0.544) \cdot (0.456) = 496$$

The expected number of individuals of the genotype GG:

$$p_2^2 n = (0.456)^2 = 208$$

(iii)
1. Observed number of individuals with genotype AA = 299
2. Observed number of individuals with genotype AG = 490
3. Observed number of individuals with genotype GG = 211
4. Expected number of individuals with genotype AA = 296
5. Expected number of individuals with genotype AG = 496
6. Expected number of individuals with genotype GG = 208

$$\begin{aligned}\tilde{\chi}^2 &= \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} \\ &= \left(\frac{(299-296)^2}{296} + \frac{(490-496)^2}{496} + \frac{(211-208)^2}{208} \right) \\ &= (0.0304 + 0.0726 + 0.0433) \\ &= 0.1462\end{aligned}$$

There is one degree of freedom for the test of Hardy-Weinberg proportions (degrees of freedom equals the number of genotypes minus the number of alleles). The critical value at a 5% significance level for one degree of freedom is 3.84 [2]. Since the $\tilde{\chi}^2$ value is less than this critical value, we do not reject the null hypothesis that the population follows Hardy-Weinberg frequencies.

(b) (i) *Linkage Equilibrium* means that the alleles of different genes at different loci are inherited independently from each other, whereas *disequilibrium* indicates a connection between the alleles.

(ii) In linkage equilibrium, a_1 and b_1 are independent from each other. Given that A and B are independent events, the probability of the occurrence of both A and B being equal to the product of the probabilities of A and B. Therefore $p_{1,1} = a_1 \cdot b_1$.

(iii) If all haplotypes are in equilibrium, then the following equations hold:

$$\begin{aligned}p_{1,1} &= a_1 \cdot b_1 \\ p_{1,2} &= a_1 \cdot b_2 \\ p_{2,1} &= a_2 \cdot b_1 \\ p_{2,2} &= a_2 \cdot b_2 \\ \sum_{i,j} p_{i,j} &= a_1(b_1 + b_2) + a_2(b_1 + b_2) = 1\end{aligned}$$

This shows that, deviance D is equal to zero in equilibrium:

$$\begin{aligned}
D &= p_{1,1}p_{2,2} - p_{1,2}p_{2,1} \\
D &= a_1b_1a_2b_2 - a_1b_2a_2b_1 \\
D &= 0
\end{aligned}$$

Now, we know that the deviance from the linkage equilibrium for each haplotype is as follows:

$$D = p_{i,j} - a_ib_j$$

$$\begin{aligned}
D_{1,1} = p_{1,1} - a_1b_1 &= p_{1,1} - (1 - a_2)b_1 \\
&= p_{1,1} - b_1 + a_2b_1 \\
&= p_{1,1} - (p_{1,1} + p_{2,1}) + a_2b_1 \quad \text{(marginalization of } b_1\text{)} \\
&= -p_{2,1} + a_2b_1 \\
&= -(p_{2,1} - a_2b_1) \\
&= -D_{2,1}
\end{aligned}$$

Applying the same logic, it can be shown that $D_{1,1} = -D_{2,1} = D_{2,2} = -D_{1,2}$. It follows:

$$\begin{aligned}
p_{1,1}p_{2,2} - p_{1,2}p_{2,1} &= (D_{1,1} + a_1b_1)(D_{2,2} + a_2b_2) - (D_{1,2} + a_1b_2)(D_{2,1} + a_2b_1) \\
&= (D + a_1b_1)(D + a_2b_2) - (-D + a_1b_2)(-D + a_2b_1) \\
&= D^2 + Da_2b_2 + Da_1b_1 + a_1b_1a_2b_2 - (D^2 - Da_1b_2 - Da_2b_1 + a_1b_2a_2b_1) \\
&= D(a_2b_2 + a_1b_1 + a_1b_2 + a_2b_1) \\
&= D(a_2(b_2 + b_1) + a_1(b_2 + b_1)) \\
&= D(a_2 + a_1)(b_2 + b_1) \\
&= D
\end{aligned}$$

As a result, we show that the Linkage Disequilibrium D is equal to $p_{1,1}p_{2,2} - p_{1,2}p_{2,1}$.

References

[1] Wang, L., Huang, M., Ding, H., Jin, G., Chen, L., Chen, F., & Shen, H. (2018). Genetically determined height was associated with lung cancer risk in East Asian population. Cancer Medicine, 7(7), 3445-3452. doi: 10.1002/cam4.1557. PMID: 29790669; PMCID: PMC6051217.

[2] Walpole R. E. Myers R. H. Myers S. L. & Ye K. (2012). Probability & Statistics for Engineers & Scientists (9th ed.). Prentice Hall.

[3] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* Springer, New York, 2010. ISBN: 9781441923226.