Medical Data Science, WS 2021/2022
Prof. Dr. Nico Pfeifer
Chair for Methods in Medical Informatics
University of Tuebingen

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

2021-10-26

Assignment 1

Deadline: Tuesday, November 9, 8:00 p.m.

This problem set is worth 50 points. You can submit in groups of two people or alone. Submit your solutions digitally by uploading to the ILIAS webpage (none of the other students can see the files you upload). Just upload a zipped folder containing all necessary files and name the folder by your last names. The folder should be named according to the following scheme:

```
[MDS][Assignment1]_lastname1_lastname2
```

Problem 1 (T, 20 Points)

Basics of statistical learning and population association studies (Use your own words and answer the quesions briefly!).

(a) (1P) Explain the main purpose of GWAS and name a question that could potentially be answered using GWAS.

(b) (1P) What are SNPs? How many possible combinations are there for SNPs?

(c) (2P) How is *allele* defined and what ist the difference between *major* and **minor** allele? How does this effect the genome encoding?

(d) (1P) What is meant by *linkage* in the context of *SNPs* and how can it be used to decrease the amount of required data and therefore computational power?

(e) (2P) Explain the terms *classification* and *regression*. Explain how the price of a house can be interpreted as both, a classification and a regression task.

(f) (3P) What is the (mathematical) definition of a *p-value*? What is the difference between the *p-value* and the *significance level* α of a hypothesis test? When do we accept a hypothesis? You can use a numerical example to support your statements.

(g) (3P) What problems occur with *multiple testing*? Name and briefly describe two approaches that tackle these problems.

(h) (3P) What are *confounding factors* and how do they influence the results of association studies? What are examples for confounding factors? Describe one method that corrects for this bias. Which advantages/disadvantages does this method have?

(i) (1P) Explain the *Cochran-Armitage test*. What is its main problem? Use the words *homozygous* and *heterozygous* in this context and explain them shortly.

(j) (2P) What are Linear Mixed Models? How do they differ from Simple Linear Models and why is this beneficial in the field of Genomic Control?

(k) (1P) Explain in your own words how the FAST-LMM algorithm achieve a computation time of O(MN^2+N^3).

Problem 2 (T, 10 Points)

Read about the Hardy-Weinberg Equilibrium.
Let A be a gene with two alleles A_1 and A_2. The Hardy-Weinberg principle denotes that

- the probability of observing the genotype A_1A_1 is p_1^2
- the probability of observing the genotype A_1A_2 is $2p_1p_2$
- the probability of observing the genotype A_2A_2 is p_2^2

where p_i is the probability of observing the corresponding allele A_i with $i \in \{1, 2\}$.

(a) Assume that we have a study consisting of 1000 people and observe the following genotypes of a SNP in the coding region of protein X on a certain chromosome: 299 AA, 490 AG and 211 GG.

(i) (2P) Calculate the allele frequencies of A and G.

(ii) (1P) Calculate the expected numbers of individuals of each genotype (assuming Hardy-Weinberg Equilibrium).

(iii) (2P) Using χ^2-test and a significance level $\alpha = 0.05$, determine whether or not this population is in Hardy-Weinberg Equilibrium for this SNP.

(b) The Hardy-Weinberg Equilibrium is defined for one single locus. Let us extend it to two loci.
Assume we have two genes (A and B) on the same chromosome with two alleles each (A_1 and A_2, and B_1 and B_2). Let $p_{i,j}$ denote the probability to observe the haplotype A_iB_j with $i, j \in \{1, 2\}$. Let $a_i\, b_i$ denote the probability to observe the allele A_i, B_i with $i \in \{1, 2\}$, respectively.

(i) (1P) What is meant by Linkage Equilibrium and Disequilibrium in general?

(ii) (1P) What has to hold if the haplotype A_1B_1 is in Linkage Equilibrium. Give an equation using the above defined probabilities.

(iii) (3P) Linkage Disequilibrium is the deviance (D) from the Equilibrium. Prove that $D = p_{1,1}p_{2,2} - p_{1,2}p_{2,1}$.

Problem 3 (P, 20 Points)

In this exercise, you will perform an association analysis on synthetic SNP data using the toolset plink (https://www.cog-genomics.org/plink2/) and BOLT-LMM (https://data.broadinstitute.org/alkesgroup/BOLT-LMM/) or FAST-LMM (https://fastlmm.github.io). Additional parts that cannot be completed using *plink*, can be performed in python. Provide the commands you used for the *plink* and *BOLT-LMM|FAST-LMM* tool in your submission. Download *plink* and *BOLT-LMM* or *FAST-LMM* from the official websites and the data (data.zip) from the password protected area of the course website.
Hint: Mostly one line of code is sufficient to obtain the result. Take your time to study the documentation and find the right commands for the tasks! Afterwards interpret the results to answer the questions.

(a) Calculate the p-values for the different SNPs using the Cochran-Armitage test and generate a Q-Q plot. This can be done using *plink*. Remove all SNPs/hypotheses for which you do not get a p-value. Show that the data is not calibrated (according to the measure of λ, discussed in lecture 2).

(b) Recalibrate the test statistic using λ (genomic control) and perform the Cochran-Armitage test, both can be done by expanding the command of (a). What changed compared to (a)?

(c) Apply the Cochran-Armitage test and Bonferroni correction for multiple testing in combination with genomic control. Give a possible explanation for the obtained result.

(d) Correct for population structure using *BOLT-LMM* or *FAST-LMM* and correct the p-values for multiple testing. Compare the corrected p-values to those of exercise (c). Which p-value correction led to better results?