Medical Data Science, SS 2019
Prof. Dr. Nico Pfeifer
Chair for Methods in Medical Informatics
University of Tuebingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

2019-04-24

Assignment 1

Deadline: Thursday, May 9, 9:59 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 name(s). The folder should be named according to the following scheme:

`[MDS][Assignment 1]_lastname`
or
`[MDS][Assignment 1]_lastname1_lastname2`

Problem 1 (T, 20 Points)

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

(a) (3P) Define *genotype* and *phenotype* by using and explaining the terms (*major* and *minor*) *allele*, *dominant* and *recessive*.

(b) (3P) Describe *classification*, *regression*, and the *kernel trick*. In which scenarios is the latter used?

(c) (4P) What is the (mathematical) definition of a *p-value*? What is the *significance level* α of a hypothesis test (mathematically)? What problems occur with *multiple testing*? Name and describe two approaches that tackle these problems.

(d) (5P) What is the general aim of a *population association study*? How can *confounding factors* influence the results of association studies and what are examples for confounding factors? Describe methods to correct for this bias. Which advantages and disadvantages do these methods have?

(e) (1P) Explain the main idea behind *genomic control*.

(f) (2P) How are *linear mixed models* an extension of linear models and what are their advantages?

(g) (1P) What is meant by *linkage* in the context of population association studies and how can it be exploited in the design of GWAS?

(h) (1P) The lecture focused on single SNP analyses. Provide a sketch for a possible method that takes several SNPs into consideration.

Problem 2 (T, 10 Points)

Read about the Hardy-Weinberg Equilibrium (e.g., in the book 'Principles of Population Genetics' by Daniel Hartl and Andrew Clark, which is available in our library). 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: 298 AA, 489 AG and 213 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) (1) What is meant by Linkage Equilibrium and Disequilibrium in general?

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

(iii) (3) 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://github.com/MicrosoftGenomics/FaST-LMM) You have to use python 2.7 in order to get FaST-LMM installed. Additional parts that cannot be completed using *plink*, can be performed in MATLAB or similar. 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.

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

(b) Recalibrate the test statistic using λ (genomic control) and perform the Cochran-Armitage test. 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. Comment on your findings.