Bioinformatics and Statistical Genetics	Hardy-Weinberg equilibrium (20p)
Jan Graffelman	November 13, 2018

Name: .....

Name: .....

You can make use of the R-packages **HardyWeinberg** and **genetics** (and other packages) to compute your answers. Prepare a .pdf file with all your answers and figures. Send your work by email to the course instructor (`jan.graffelman@upc.edu`) before the 20^{th} of November 2018.

1. The file YRIChr1.rda contains genotype information (10000 SNPs) of individuals from an African population of unrelated individuals. Load this data into the R environment. The file contains a data objects, `X`, with genotype information. This data is in (0,1,2) format, where 0 and 2 represent the homozygotes AA and BB, and 1 represents the heterozygote AB.

2. (2p) How many individuals does the database contain? What percentage of the variants is monomorphic? Remove all monomorphic SNPs from the data bases. How many variants remain in the database? Determine the genotype counts for these variants, and store them in matrix. Apply a chi-square test without continuity correction for Hardy-Weinberg equilibrium to each SNP. How many SNPs are significant (use $\alpha = 0.05$)? .....

   .....

3. (1p) How many markers of the remaining non-monomorphic markers would you expect to be out of equilibrium by the effect of chance alone? .....

   .....

4. (1p) Apply an Exact test for Hardy-Weinberg equilibrium to each SNP. You can use function `HWExactStats` for fast computation. How many SNPs are significant (use $\alpha = 0.05$). Is the result consistent with the chi-square test? .....

   .....

5. (1p) Apply a likelihood ratio test for Hardy-Weinberg equilibrium to each SNP, using the `HWLratio` function. How many SNPs are significant (use $\alpha = 0.05$). Is the result consistent with the chi-square test? .....

   .....

6. (1p) Apply a permutation test for Hardy-Weinberg equilibrium to the first 10 SNPs, using the classical chi-square test (without continuity correction) as a test statistic. List the 10 p-values, together with the 10 p-values of the exact tests. Are the result consistent? .....

   .....

7. (1p) Depict all SNPs simultaeneously in a ternary plot, and comment on your result (because many genotype counts repeat, you may use `UniqueGenotypeCounts` to speed up the computations) ....

.....

8. (1p) Can you explain why half of the ternary diagram is empty? .....

9. (2p) Make a histogram of the p-values obtained in the chi-square test. What distribution would you expect if HWE would hold for the data set? Make a Q-Q plot of the p values obtained in the chi-square test against the quantiles of the distribution that you consider relevant. What is your conclusion?. .....

.....

10. (1p) Imagine that for a particular marker the counts of the two homozygotes are accidentally interchanged. Would this affect the statistical tests for HWE? Try it on the computer if you want. Argue your answer. .....

.....

11. (3p) Compute the inbreeding coefficient ($\hat{f}$) for each SNP, and make a histogram of $\hat{f}$. You can use function `HWf` for this purpose. Give descriptive statistics (mean, standard deviation, etc) of $\hat{f}$ calculated over the set of SNPs. What distribution do you expect $\hat{f}$ to follow theoretically? Use a probability plot to confirm your idea.....

.....

12. (2p) Make a plot of the observed chi-square statistics against the inbreeding coefficient ($\hat{f}$). What do you observe? Can you give an equation that relates the two statistics? .....

.....

13. (1p) Simulate SNPs under the assumption of Hardy-Weinberg equilibrium. Simulate the SNPs of this database, and take care to match each of the SNPs in your database with a simulated SNP that has the same sample size and allele frequency. You can use function `HWData` of the HardyWeinberg package for this purpose. Compare the distribution of the observed chi-square statistics with the distribution of the chi-square statistics of the simulated SNPs by making a Q-Q plot. What do you observe? State your conclusions .....

.....

14. (2p) We reconsider the exact test for HWE, using different significant levels. Report the number and percentage of significant variants using an exac test for HWE with $\alpha = 0.10, 0.05, 0.01$ and 0.001. State your conclusions.

15. (1p) Do you think genotyping error is a problem for the database you just studied? Explain your opinion. .....