Practical 2. Likelihood ratio tests (LRT)   Hand-in: 19/10/2022

Resolve the following exercises in groups of two students. Write your solution in a Word, Latex or Markdown document and **generate a pdf file** with your solution. Upload the pdf file with your solution to the corresponding task at the Moodle environment of the course, no later than the hand-in date.

Many well-known standard statistical tests are actually LRT tests. We do some exercises with data sets where we apply these LRT tests.

1. (10p) **Likelihood ratio test for Hardy-Weinberg equilibrium**. In a genetic association study, the genotypes of a single nucleotide polymorphism have been determined for a sample of individuals. The genotype data file `snp.dat` contains the genotyping results.

   (a) (1p) Load the data in the R environment, and make a table of the different genotypes. Report the table. What is the sample size of the study?

   (b) (1p) How many alleles does this SNP have? How many genotypes could it theoretically have? Estimate all relative genotype frequencies by maximum likelihood (ML). Report the values of the ML estimators.

   (c) (2p) Count the number of alleles of each type in the sample. Estimate the relative allele frequencies by ML. Report the values of the ML estimators.

   (d) (1p) Which allele is the minor (least common) allele?

   (e) (1p) Do a likelihood ratio test (LRT) for Hardy-Weinberg equilibrium using the `HWLratio` function of the R-package `HardyWeinberg`. Report the likelihood ratio statistic and the p-value.

   (f) (1p) State your conclusion of the LRT.

   (g) (1p) State the distribution the the LR statistic for this problem.

   (h) (1p) Calculate the p-value ”by hand” using the value observed for the LR statistic and its distribution. Show your computations. Do you obtain the same result as the `HWLratio` function?

   (i) (1p) Calculate the expected genotype counts under the assumption of Hardy-Weinberg equilibrium. Compare them with the observed counts. What do you observe?

2. (10p) **Comparison of regression models**. In a study on quality of red wines, a set of physicochemical variables has been collected for a large database of red wines. The variable *quality* is used as the response variable in a multiple regression with the physicochemical variables

as predictors. Most physicochemical variables were log-transformed to reduce skew. The available predictors are *fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates* and *alcohol*. The file `RedWines.dat` contains the data.

(a) Load the data into the R environment with the `read.table` instruction.

(b) (2p) Fit a full model by the regression of *quality* on all physicochemical predictors. Report the adjusted R^2 statistic of this model. Which variables are not significant? (use $\alpha = 0.05$).

(c) (2p) Fit a reduced model, eliminating all insignificant predictors from the regression equation in a stepwise fashion (use $\alpha = 0.05$). Report the adjusted R^2 statistic of this reduced model. Does this model have a better or worse fit, according to this statistic?

(d) (2p) Do a likelihood ratio test (F-test) to see whether the full or reduced model fits the data better. Report the F statistic, its reference distribution and the p-value, and state your conclusion.

(e) (2p) Do simple linear regressions of *quality* on the predictors that you eliminated from the model. Do these regressions confirm that the eliminated predictors do not explain *quality*? State your findings and conclusions.

(f) (2p) Are regression coefficients you found in the different regressions consistent with each other? Comment on your findings.