Bachelor's Degree in Bioinformatics
Statistical Models & Stochastic processes
Academic year 2023-2024 1st Quarter

Practical 2. Likelihood ratio test (LRT)

Hand-in date: 25/10/2023

Resolve the following exercise 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.txt 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 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.
   The outcome or response variable is seize and the explanatory variables or predictors are trt, base and age. Subject contains an ID for every individual.
   The dataset seizures_visit4.xls contains the measures performed at visit 4 in a clinical trial.

The Clinical trial was conducted in m = 59 subjects suffering from simple or partial seizures.

- Patients were randomized to the anti-epileptic drug progabide or placebo (0= Placebo, 1=Progabide; variable trt).
- A baseline measure of each subject's propensity for seizures was recorded, namely, the number of seizures suffered in the 8 weeks leading up to the start of the study (variable base).
- Each subject's age at the start of assigned treatment was also recorded (variable age).
- After initiation of assigned treatment, the number of seizures experienced by each subject in n = 4 consecutive two-week periods was recorded, so that the response is a count measured at week 8 of follow up (variable seize).

  The variable seize is used as the response variable in a multiple regression with the available variables as predictors.

  a. (0p) Load the data into the R environment. Do a summary of the data set.
  b. (2p) Fit a full model by the regression of seize on all predictors available in the data set. 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 seize on the predictors that you eliminated from the model. Do these regressions confirm that the eliminated predictors do not explain seize? State your findings and conclusions.
  f. (2p) Are regression coefficients you found in the different regressions consistent with each other? Comment on your findings.