Exercises for the lecture on

# Statistical Methods

Prof. Dr. Vladimir Spokoiny
Vladimir.Spokoiny@wias-berlin.de
Dr. Thorsten Dickhaus
Thorsten.Dickhaus@wias-berlin.de
www.wias-berlin.de/people/dickhaus/

Humboldt-University Berlin
Department of Mathematics
Winter term 2013 / 2014

## Sheet 6

Solutions are due on Monday, November 25th, 2013, 3:15pm.
Every completely and correctly solved exercise gives 4 points.

## Exercises

21. **Likelihood-based inference.**

    (a) Show that the assertion of Example 3.6.(b) holds true (maximizing the likelihood function is equivalent to minimizing the sum of squares of residuals in a regression model with normally distributed noise). This also proves Theorem 3.18.(c).

    (b) Prove Theorem 3.18.(d).

    (c) Show that the two-sided $t$-test is a likelihood ratio test for the pair of hypotheses $H_0 : \{\mu = 0\}$ versus $H_1 : \{\mu \neq 0\}$ in a Gaussian iid. model with unknown variance.

22. **Asymptotic $\chi^2$-test for independence in contingency tables.** In a large-scale genetic association study, the Wellcome Trust Case Control Consortium (WTCCC, Nature, Vol. 447, 2007) aimed at identifying positions on the human genome that are associated with the risk of developing certain diseases, including Crohn's disease. To this end, blood samples from approximately 1800 Crohn's disease patients and approximately 3000 healthy controls were taken and the genotypes of all these study participants at specific genomic positions (loci) were assessed. The genotype of an individual at a particular locus consists of two alleles $A_1$, $A_2$ with values in $\{A, C, G, T\}$, which are the maternal and the paternal DNA bases at this position for this individual (on one strand of the DNA double helix). The following table summarizes the WTCCC's Crohn's disease data for locus `rs5987140`.

| Genotype | $CC$ | $CT$ | $TT$ | $\sum$ |
|---|---|---|---|---|
| diseased | 101 | 234 | 1413 | 1748 |
| healthy | 196 | 322 | 2413 | 2931 |
| $\sum$ | 297 | 556 | 3826 | 4679 |

    (a) Compute and interpret the $p$-value of an asymptotic $\chi^2$-test for independence of disease status and genotype at locus `rs5987140` for the given data.
    Hint: Make use of Remark 3.13.

    (b) On total, the WTCCC dataset for Crohn's disease comprises data for $m = 455086$ loci and all these $m$ loci are simultaneously to be tested for association with risk for Crohn's disease. How could one "adjust" the $m$ locus-specific $p$-values, such that the probability for at least one type I error among the $m$ tests is bounded by a pre-defined constant $\alpha \in (0, 1)$ if one compares the adjusted $p$-values with $\alpha$?
    Re-interpret the $p$-value from (a) based on this method.

23. **Programming exercise: Rent index data.**
Make yourself familiar with the rent index dataset from Example 3.3 (`mietspiegel99.txt`) which you can download from the lecturer's homepage. If you should encounter problems in downloading the file or if you do not have access to the internet, you may alternatively get the file via USB stick during the lecturer's consulting hour.

(a) Estimate intercept and regression coefficient for the non-linear model

$$Y = \beta_0 + \beta_1 / X + \varepsilon,$$

where $Y$ represents the rent per month and square meter in DM and $X$ the living space in square meters. Provide one solution that makes use of Theorem 3.18.(a) and another solution that uses built-in routines of the employed statistics software.

(b) Compute the residuals of this model and perform a graphical residual analysis.

24. **Multiple Select.** Which of the following statements are true and which are false?
Please give reasons for your respective decisions (one short sentence each is sufficient).

1. Maximum likelihood estimators are always admissible.
2. If the realized value $\hat{\vartheta}(x)$ of an (unrestricted) maximum likelihood estimator lies in $\Theta_0$, then an asymptotic likelihood ratio test for $\Theta_0$ versus $\Theta \setminus \Theta_0$ does not reject for any value of $\alpha \in (0, 1)$.
3. If a categorical covariate with exactly $\ell$ realizable values is represented by $\ell$ indicators, then the assumptions of Model 3.14 are violated.
4. Presence of an interaction effect of two covariates $X_j$ and $X_k$ under Model 3.14 can be tested with an asymptotic likelihood ratio test with one degree of freedom.