UNIVERSITEIT
GENT

Faculteit
Landbouwkundige en
Toegepaste Biologische
Wetenschappen

Academiejaar 2000 – 2001

Nonparametrical tests based on sample space partitions

Niet-parametrische testen gebaseerd op partities van de steekproefruimte

door
ir. OLIVIER THAS

Thesis submitted in fulfillment of the requirements for the degree
of Doctor (Ph.D.) in Applied Biological Sciences

Proefschrift voorgedragen tot het bekomen van de graad van
Doctor in de Toegepaste Biologische Wetenschappen

op gezag van
Rector: **Prof. Dr. ir. J. WILLEMS**

Decaan:
Prof. Dr. ir. O. VAN CLEEMPUT

Promotor:
Prof. Dr. J.P. OTTOY

Abstract

This thesis is concerned with probably one of the oldest problems in statistical science: goodness-of-fit tests. In particular 3 types of goodness-of-fit are discussed: (1) the one-sample problem, (2) the k-sample problem and (3) the independence problem. All three types are frequently used in the statistical analysis of biological data. Although many tests exists already, there is still need for omnibus tests that have good overall small sample power. In this thesis, a new class of nonparametric statistical tests for these problems is proposed, all based on the concept of observation-based sample space partitions (SSP) of size c (SSP tests). Basically the test statistics are averages of Pearson chi-squared statistics that are applied to SSP-induced contingency tables (i.e. tables of which the borders of the cells are completely determined by a subset of the sample). For problems (2) and (3) the tests become rank tests, and, hence, exact permutation distributions may be enumerated or approximated by means of Monte Carlo simulations. For all tests the asymptotic null distribution is proven, as well as their omnibus consistency. In a series of simulation studies the powers of the SSP tests are compared to those of many other tests described in the literature. The study shows that especially for the k-sample and the independence problem very high powers are obtained with the SSP tests as compared to many other tests. Further, the study indicates that the power strongly depends on the SSP-size c, which has to be chosen arbitrarily by the user. In order to overcome this practical problem, data-driven versions are proposed, i.e. the SSP size is estimated from the sample. Their asymptotic null distributions and omnibus consistency are proven. In simulation studies it is shown that the data-driven tests succeed quiet well in selecting a SSP size that results in a high-power SSP test. Further, some extensions to the SSP tests are worked out. First, a generalization of the test statistic is obtained by substituting the core of the SSP statistic, which is the Pearson statistic, by the Cressie-and-Read family of goodness-of-fit statistics, indexed by l. Empirical evidence suggests that the small sample power depends on the choice of l, and thus the power may potentially be increased by a good choice of l. A second extension is the decomposition of the SSP test statistic into interpretable components. This has been worked out in detail for the k-sample problem. It is shows that the i-th component is related to differences in i-th moments between the k samples. This makes the SSP omnibus test also capable of suggesting the direction of deviation from the null hypothesis.