Testing equivalence of the two way contingency tables to the independence model.

January 7, 2018

This package provides two tests to show equivalence of the two way contingency tables to the independence model. The goal is to show that two categorial random variables are approximately independent distributed. The approximate independence can be important for some applications or also greatly simplify calculations. However, the testing of the equivalence is a difficult task.

Let d denote Euclidean distance and let $\mathcal{M}$ be the independence model, which contains all product measures of the corresponding dimensions. Two categorial random variables are considered approximately independent for the tolerance parameter $\varepsilon > 0$, if the joint probability density p fulfills the following condition: There exists a product measure $q \in \mathcal{M}$ such that $d(p, q) < \varepsilon$. The minimum distance between p and the model $\mathcal{M}$ is defined as

$$d(p, \mathcal{M}) = \inf_{q \in \mathcal{M}} d(p, q).$$

The equivalence test problem is formally stated by

$$H_0 = \{d(p, \mathcal{M}) \geq \varepsilon\} \text{ against } H_1 = \{d(p, \mathcal{M}) < \varepsilon\},$$

where $\varepsilon > 0$ is the tolerance parameter. The goal is to reject the hypothesis of the non-equivalency H_0 at a significance level α.

The package provides two tests for that purpose: the asymptotic test and the bootstrap test. Both tests are available as functions in the module “tests two way collapsibility”, which return the class “TestResult” as result. The class “TestResult” contains two public fields only:

- Field “result” is Boolean. The value is true if the test rejects H_0 and false otherwise.
- Field “minEps” is double. This is the smallest tolerance parameter ε, for which the test can reject H_0.

The asymptotic test is based on the asymptotic distribution of the test statistic. Therefore the asymptotic test need some sufficiently large number of the observations. It should be used carefully because the test is approximate and may be anti conservative at some points. In order to obtain a conservative test reducing of α (usually halving) or slight shrinkage of the tolerance parameter ε may be appropriate. The asymptotic test is realized as the function “asymptoticTest”, which has the following parameters:

- p is a two dimensional array of double. It should contain two way contingency table.
- n is the number of observations.
- alpha is the significance level.
- epsilon is the tolerance parameter ε.

The bootstrap test is based on the re-sampling method called bootstrap. The bootstrap test is more precise and reliable than the asymptotic test. However, it should be used carefully because the test is approximate and may be anti conservative at some points. In order to obtain a conservative test reducing of α (usually halving) or slight shrinkage of the tolerance parameter ε may be appropriate. We prefer the slight shrinkage of the tolerance parameter because it is more effective and the significance level remains unchanged. The bootstrap test is realized as the function “bootstrapTest”, which has the following parameters:

- p is a two dimensional array of double. It should contain two way contingency table.
- n is the number of observations.
- alpha is the significance level.
- epsilon is the tolerance parameter ε.
- nDirections is the number of random directions to search for a boundary point of H_0. The number of random directions has a negative impact on the computation time. The number should be set empirically. You can increase it gradually (100, 200, ...) until the minimum tolerance parameter “minEps” does not change anymore. For example, we would recommend to use 100 directions for 2x4 tables and 1000 directions for 4x5 tables.
- nBootstrapSamples is the number of bootstrap samples. The parameter should be at least 1000. However, higher values lead to the better approximation generally. Usually it is not necessary to generate more than 10.000 bootstrap samples.

The bootstrap test needs considerable computation time. For example, it may need few minutes on the usual office computer.