Nowadays the problem of dealing with high-dimensional data arises in many different areas of science, such as genetics, medicine, pharmacy and social science. The main property of high-dimensional data is that the data dimension p, i.e., the number of variables or features (e.g. genes), is large while the sample size is relatively small. The high data dimension has forced statisticians to renovate or rewrite existing methods, or even to propose new ones, since classical methods may be inappropriate in such a setting. In this dissertation new statistical procedures to solve classical k-sample problems in the context of high-dimensional data have been developed and investigated. To distinguish our situation from the classical one, we use the notation *p* for the number of variables instead of k.

The first contribution is to solve the classical k-sample problem but in the dependent high-dimensional framework, i.e, k tends to infinity. Zhan and Hart (2014) thus proposed a test for the null hypothesis that all the variables have the same distribution when the number of variables grows, but the sample sizes remain bounded. The test proposed by Zhan and Hart (2014) is applicable when the p samples are independent of each other. In practice, however, it is possible that different data sets are dependent. Therefore an adaptation of the test of Zhan and Hart (2014) that is suitable under dependence is proposed in this dissertation. The main idea is to propose a new estimator of the variance of the statistic suitable under dependencies which satisfy certain mixing conditions. Asymptotic normality and consistency results under general conditions on the alternative are proved and the new test is applied to some real data and its performance is studied in and extensive simulationn study.

The second contribution is a test for the null hypothesis of equality of the p marginal distributions for two groups (e.g. two tumors). Then, we have a k-sample problem with k=2, but the number of variables p whose marginal distributions are compared goes to infinity. Note that one can regard the (global) null hypothesis of equality between the p marginals as an intersection of the p null hypotheses corresponding to p different two-sample problems. A test statistic motivated by the simple idea of comparing, for each of the p variables, the empirical characteristic functions computed from the two samples is proposed. The asymptotic normality of the test statistic is derived under mixing conditions. In order to obtain a practical test several estimators of the variance are proposed, leading to three somewhat different versions of the test. A simulation study to investigate the finite sample properties of the proposed tests is carried out, and a practical illustration involving microarray data is provided.

The global two-sample test proposed can reject or accept the global null hypothesis. When the global null hypothesis is accepted, the conclusion is that the distribution of each of p variables in the two groups is the same. However, if the test rejects the global null hypothesis, additional investigation about which of the p variables are responsible for the statistical significance is often required. Interestingly, the individual test statistics which the proposed global two-sample test is based on can be used to test each of the p null hypotheses separately. Hence, a permutation test based on such individual test statistic can be defined. To take the multiplicity of tests into account a multiple comparison procedure (MCP) must be applied to the large set of p-values. Nevertheless, due to the special form of these permutation p-values (specifically, they are discrete, uniform and homogeneous), majority of MCP's perform poorly. Hence, in this dissertation several adaptions of the FDR procedure proposed by Storey and Tibshirani (2003) have been propose to accommodate it to the characteristic of the p-values. Finally the behaviour of the two sample permutation test proposed in this dissertation is compared by simulation with other well-known two-sample tests obtaining promising results.

It is important to stand out that the dissertation includes the development of user-friendly software ito facilitate the practical use of the new statistical methods. More precisily, two R packages have been developed to apply each of the global tests.

References

Storey, J. and R. Tibshirani (2003). Statistical signicance for genomewide studies. Proceedings of National Academy of Science 100, 9440-9445.

Zhan, D. and J. Hart (2014). Testing equality of a large number of densities. Biometrika 101, 449-464.