Statistical homogeneity tests applied to large data sets from high energy physics experiments

Trusina J, Franc J, and Kůs V
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Trojanova 13, 120 00 Prague 2, Czech Republic
E-mail: jakub.trusina@fjfi.cvut.cz, jiri.franc@fjfi.cvut.cz, kus@fjfi.cvut.cz

Abstract. Homogeneity tests are used in high energy physics for the verification of simulated Monte Carlo samples, it means if they have the same distribution as a measured data from particle detector. Kolmogorov-Smirnov, $\chi^2$, and Anderson-Darling tests are the most used techniques to assess the samples’ homogeneity. Since MC generators produce plenty of entries from different models, each entry has to be re-weighted to obtain the same sample size as the measured data has. One way of the homogeneity testing is through the binning. If we do not want to lose any information, we can apply generalized tests based on weighted empirical distribution functions. In this paper, we propose such generalized weighted homogeneity tests and introduce some of their asymptotic properties. We present the results based on numerical analysis which focuses on estimations of the type-I error and power of the test. Finally, we present application of our homogeneity tests to data from the experiment DØ in Fermilab.

1. Introduction
Homogeneity tests are used when determining if any two samples originate from the same distribution. We use these tests to verify that data measured at certain high energy physics (HEP) experiment are homogeneous with simulated Monte Carlo (MC) sample (i.e. both samples have the same distribution). Rejection of such a test means remodeling of yield sample contribution. Since an amount of entries produced by these generators is significantly higher than the size of data sample, we have to assign an appropriate weight to every entry which simulates not only various trigger efficiencies in the detector. Classical statistical tests should be generalized in such a way that they can test the homogeneity of weighted samples. In this paper, we present $\chi^2$ and Kolmogorov-Smirnov (KS) tests which we implemented in ROOT, the framework used by HEP community. We also compare these tests with those which are already implemented in TH1 class by estimating a probability of type-I error which is a criterion of tests correctness. Subsequently, we discuss different approaches for handling of these weights and their possible utilization in practice. Finally, we apply the tests to samples from DØ experiment.

2. Homogeneity tests
Despite the fact that we have worked with Anderson-Darling test too, in this paper, we present results of $\chi^2$ and KS homogeneity tests only. The reason is that we can compare our proposed tests with those already implemented in ROOT. Nevertheless, it is worth mentioning that Anderson-Darling test has a higher power of test than KS test as shown in [1].
The $\chi^2$ test is designed for multinomial distributed samples represented by binned data (i.e. we count the number of events belonging to the individual intervals of observed space partition). If both samples are homogeneous, they have similar numbers of events in bins. That means we test whether both samples have multinomial distribution $M(n, p_1, ..., p_k)$, where the parameter $n$ can be different for both samples while parameters $(p_1, ..., p_k)$ are mutual. Let $X = (X_1, ..., X_k)$ and $Y = (Y_1, ..., Y_k)$ be random vectors with the multinomial distributions $M(n, p^{(1)}_1, ..., p^{(1)}_k)$ and $M(m, p^{(2)}_1, ..., p^{(2)}_k)$, respectively. Test statistic of KS test and its asymptotic property are the following:

$$T(X, Y) = \sqrt{\frac{nm}{n+m}} \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)| \overset{D}{\underset{H_0}{\rightarrow}} K(\lambda) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k\lambda^2}$$  (3)

under the assumption of homogeneity hypothesis $H_0$. Replacing $(X, Y, n, m, F_n, G_m)$ by $(W, V, W_i, V_i, F_W, G_V)$ we obtain generalized test statistic. We expected that the asymptotic property would not vanish after this modification; however, similarly as for $\chi^2$ test, such property seems to be true for some class of weights only.

3. Estimating probability of type-I error
To verify the weighted tests, we need to show that probability of the type-I error is equal or lower than the significance level $\alpha$. Nevertheless, in order to maximize power of the
test, the probability should be as close as possible to the significance level. Therefore, we carried out computer experiments and compared both $\chi^2$ (under probabilistic binning) and KS tests with TH1::Chi2Test and TH1::KolmogorovTest [2]. We did not compare it with TMath::KolmogorovTest because it cannot be applied to weighted samples; however, it returns the same p-value as our test, if it is applied to unweighted samples. We produced two iid samples from $N(0, 1)$ of the same size: $n = m \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$. We generated and assigned iid weights to every entry of both samples individually and repeated this procedure 10 000 times. Finally, ratio of the tests that rejected null hypothesis on three significance levels $\alpha = \{0.05, 0.01, 0.001\}$ was found. If the null hypothesis (homogeneity $H_0$) is true then the ratios are distributed by Bi($n, \alpha$)/10000. The number of bins for $\chi^2$ tests was chosen as $[1.88n^{2/5}]$ which is proposed in [3]. The corresponding results are shown in figures 1 and 2. The subfigers 1a, 1b, and 2a show us that some correct class of weights seems to exist.

![Figure 1](image1.png)

(a) All weights are equal to 1. (b) All weights were generated from such $U(a,b)$ that $E[w] = 0.7$ and $\text{Var}[w] = 0.7 - 0.7^2 = 0.21$.

![Figure 1](image2.png)

(c) All weights were generated from $U(0.2,1.2)$. (d) All weights were generated from $U(0,2)$.

Figure 1: Probability of type-I error estimated for various iid weights. $\chi^2_Q$Test and KSTest were implemented by us while $\chi^2_{\text{ROOT}}$Test and KSTH1Test are implemented in TH1 class.

From the rest subplots it is clear that the expected asymptotic property is not valid for some
other classes of weights. As the mean of weights deviates from 1 the ratio deviates in the same direction. This statement is true also for variance being equal to 0 (i.e. $w = c$ almost surely, where $c$ is some constant). Furthermore, this rejects the appropriateness of tests if the weights are not randomly generated but they are chosen as constants. If the variance increases from 0, the ratio increases too. We can also see that $\text{KS}_{\text{TH1}}$ Test returns higher p-value than it should produce and, as a consequence, the ratio is lower. The reason is that the binning reduces the highest difference between two WEDFs in most cases. In [4] it is shown that after transformation of weighted sample to unweighted one by summing up “near” entries to one entry with weight equal to one, the calculated p-value by non-generalized test is not significantly different from p-value computed by generalized test. However, such sample transformation does not solve the issue that the type-I error deviates from significance level.

4. Class of perfect weights

In order to find such a class of $i id$ weights, where the probability of type-I error is equal to the significance level, we carried out another experiment. At first, we produced two samples from $N(0,1)$ containing 10 000 entries. Afterwards, we generated weights from $U(a,b)$, where $a$ and $b$ were chosen in such a way that $(E[W], \text{Var}[W]) = (0.05, 0.075, 0.1, ..., 1) \times (0.01, 0.02, ..., 0.4)$. We applied KS test, repeated this procedure 10 000 times and counted ratio of rejected tests for significance level $\alpha = 0.05$. The result can be seen in figure 3 which indicates that the class of weights, for which the homogeneity tests have the same ratio $r$, preserve the parabolic behavior

$$\text{Var}[W] = a(r)E[W] - E[W]^2,$$

where $a(\cdot)$ denotes a function which was manually estimated and linearly interpolated in figure 4. For $a(r) = 1$, KS test seems to have probability of type-I error equal to significance level $\alpha = 0.05$. The same phenomenon occurred when we generated weights from another distribution, for example $N(p, p(1-p))$ with $p \in [0.05, 1]$. Moreover, from the figures 1b and 2a it seems that even $\chi^2$ test’s ratios of rejection lie on the same parabola; however, from figures 1c or 1d it is
clear that $a(\cdot)$ is unique for every different test (for example $\chi^2$ test). Unfortunately, it is not clear whether $a(\cdot)$ changes with the samples produced from other distributions than $N(0,1)$, or how it changes with different sample sizes or if both samples have different types of weights. Ideally, we estimate mean and variance of weights, we find $a(r)$ from (4), and then we know probability of type-I error of such test when rejecting it for p-value lower than significance level $\alpha$. It is important to note that $a(\cdot)$ is scaled by $\alpha$ and it has to fulfill condition that $a(1) = \alpha$.

Figure 3: Contour plot of ratios of rejected KS tests for weights generated from such $U(a,b)$ that $(E[W],\text{Var}[W]) = (0.05,0.075,0.1,...,1)$ $\times$ $0.01,0.02,...,0.4$, fitted to parabolas which are defined by (4) for $a(r) = \{0.6,0.8,1,1.2,1.4,1.6,1.8\}$.

Figure 4: We estimated $a(r)$ by visually estimating points for which contour plot’s levels fits parabolas (as shown in figure 3).

5. Application of modified tests to DØ data

Proposed weighted homogeneity tests were applied to the top quark pair production search in lepton+jets channel using the data sample collected by the DØ detector at the Tevatron proton-antiproton collider at Fermilab [5]. All 6 different channels were used in the mentioned analysis, however, in this paper, we present only some particular results in $e + 3$jets channel, where the data sample contains 11905 entries and 11905 events while MC sample contains 719515 entries and 11905.5 events. All the weights of data sample are equal to 1, unlike of the MC sample, where the mean and variance are $E[W] = 0.0166$ and $\text{Var}[W] = 0.0036$. Histograms and WEDFs of the selected variable DRminejet can be seen in figure 5.

We already mentioned that the generalized test works well when variance of weights is equal to 0. Since in all HEP analysis the weights of final samples are not generated randomly but are assigned through the results of another analysis, we can use it without any doubts. On the other hand, if we consider our weights as iid random variables, calculated p-values would not be true p-values since asymptotic distribution of all four tests is violated. In addition, if we suppose the same function $a(\cdot)$ as in the previous experiment then p-values lie on parabola $\text{Var}[W] = 0.23E[W] - E[W]^2$. Because $0.23 < 1$, probability of type-I error is smaller than significance level and the test is correct; however, we admit that power of the test can be low.

In table 1 we can see results of homogeneity tests of MC vs. data for 8 variables with highest separation power between signal and background that were used as an input to BDT in the MVA analysis to determine final discrimination. Description of each variable can be found in [5]. Notice the difference in p-values between KS tests and $\chi^2$ tests in the Mva_max variable, which is an output from another MVA analysis determining the probability that a given event
contains a jet originating in b-quark. Since it is artificial variable, it is expected to have worse agreement between MC and data. On the other hand, whole analysis is done with binned data stored in histograms, therefore the results of $\chi^2$ test is more reliable.

6. Conclusion
In this paper, we showed that expected asymptotic distributions of generalized $\chi^2$ and KS tests with random weights are not correct; however, for the class of iid weights which fulfills condition that $\text{Var}[W] = E[W] - E[W]^2$, the probability of type-I error seems to be equal to significance level. Moreover, our conjecture is that the relation between weights and probability of type-I error $\tilde{r}$ is given by $\text{Var}[W] = a(\tilde{r})E[W] - E[W]^2$, where $a(\cdot)$ is unknown function which can be estimated for particular test setup. Some more detailed numerical analysis is needed to be carried out. It is unknown how the probability of type-I error would change if the sample sizes are different or whether $a(\cdot)$ changes if we would produce samples from another distribution than $N(0, 1)$. Proving a statement that probability of type-I error is equal to significance level iff $\text{Var}[W] = E[W] - E[W]^2$, would be a great success.

Acknowledgments
This work was supported by the CZ grants LG15047 (MYES), LM2015068 (MYES), SGS15/214/OHK4/3T/14 (CTU), and GA16-09848S (GACR).

References
[1] Engmann S and Cousineau D 2011 Journal of Applied Quantitative Methods 6 117
[2] ROOT Reference Documentation https://root.cern.ch/doc/master/index.html accessed: 2017-01-04
[3] D’Agostino R and Stephens M 1986 Goodness-of-Fit Techniques (New York: Dekker) p 70 ISBN 0824774876
[4] Bour P and Kiss V 2015 SPMS 2015 proceedings (FNSPE CTU in Prague)
[5] D0 Collaboration Abazov V M et al 2016 Phys. Rev. D 94