Generalization of Homogeneity Tests for Weighted Samples and Their Implementation in ROOT

Jakub Trusina, Jiří Franc, Adam Novotný

Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, The Czech Republic

E-mail: jakub.trusina@fjfi.cvut.cz, jiri.franc@fjfi.cvut.cz, adam.novotny@fjfi.cvut.cz

Abstract. In High Energy Physics, tests of homogeneity are used primarily in two cases: for verification that data sample does not differ significantly from numerically produced Monte Carlo sample and for verifying separation of signal from background. Since Monte Carlo samples are usually weighted, it is necessary to modify classical homogeneity tests in order to apply them to weighted samples. In ROOT, the only homogeneity tests that allow testing weighted samples are implemented for binned data. However, after the data are binned the full information is lost. Therefore we compare these ordinary tests with modified versions of the Kolmogorov-Smirnov, Anderson-Darling, and Cramér-von Mises tests that use full sample information. The proposed tests are compared by estimating a probability of type-I error which is crucial for a test's reliability.

1. Introduction

Comparing the properties of two datasets is one of the most common tasks in statistics. Homogeneity tests allow us to determine whether two or more populations have the same probability distribution. In High Energy Physics (HEP), we use these tests, among others, to compare the distribution of measured data and simulated Monte Carlo (MC) samples. Since the MC generators can produce much more records than the real experiment and on top of that these entries can be modified by weights, we need to use a generalization of classical homogeneity tests. In this paper, we present the homogeneity tests which can be applied to weighted unbinned data samples in ROOT. We verify that proposed generalized test statistics have their presumed asymptotic distribution in a large simulation study.

1.1. Homogeneity tests currently available in ROOT

Several homogeneity tests are implemented in ROOT [1], some can be found in the TH1 library for histograms (binned data with multinomial distribution), some in TMath library. The list of all available homogeneity tests with their basic properties is presented below:

- **TH1::Chi2Test** allows testing weighted samples, but it can be applied only to binned data. Results for data from continuous distribution are unreliable when sample sizes are significantly different. Various binning can lead to different test's conclusion.
- **TH1::KolmogorovTest** is a modification of the Kolmogorov-Smirnov (KS) test that can be applied to binned weighted data; however, returned p-value is higher than the true one

- **TH1::AndersonDarlingTest** is a modification of the Anderson-Darling (AD) test, it can be applied to binned unweighted data only
- **TMath::KolmogorovTest** is the classical KS test which can be applied only to unweighted and unbinned data
- **ROOT::Math::GoFTest** This class contains implementations of KS and AD test, where both tests are applicable to unweighted and unbinned samples. AD test can be applied also to binned data.

We can see from this list that only TH1::Chi2Test and TH1::KolmogorovTest tests allow testing weighted data, but these tests can be used only with binned data and are unreliable in many situations. We present the problem with binned, unweighted data in the following example.

Suppose two samples produced from normal distributions $\mathcal{N}(0,1)$ and $\mathcal{N}(0.1,1)$. We use two different binning configurations for the same samples, first one with $nbins = 10, min = -2.5, max = 2.5$ and second one with $nbins = 11, min = -2.45, max = 2.55$. As we can see from figure 1 we obtained different results from the χ^2 test. The p-value in the first approach is approximately 0.01 and in the second one 0.23. If we choose the significance level $\alpha = 0.05$, we reject the null hypothesis in the first case and fail to reject in the second one.

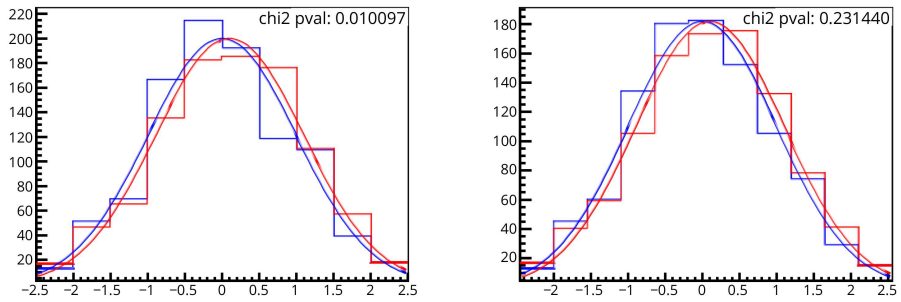

Figure 1. An example of various binning of the same sample and its effect.

This problem is common for all tests with binned data that lose the information of the sample's distribution within each bin and can lead to a different decision if the user adjusts the binning configuration.

On the other hand, tests based on empirical distribution function (EDF) keep complete information. Every difference inside bin's interval can be counted (such as Cramér-von Mises (CvM) or Anderson Darling (AD) test). Binned KS test does not find the true maximum distance between EDFs but the maximum distance between cumulative histograms which is most likely lower.

2. Generalized homogeneity tests

In order to apply homogeneity tests to weighted samples, the classical unweighted homogeneity tests need to be generalized. Generalizations for one-sample KS and AD tests are suggested in [2]. We suggest modifications of KS, CvM and AD homogeneity test statistics. Let $(\boldsymbol{X}, \boldsymbol{W}) = ((X_1, ..., X_n)', (W_1, ..., W_n)')$ be first sample with its weights and $(\boldsymbol{Y}, \boldsymbol{V}) = ((Y_1, ..., Y_m)', (V_1, ..., V_m)')$ be the second one. Let $W_\cdot = \sum_{i=1}^{n} W_i$, $\mathbf{1}_A(\cdot)$ be indicator function of set A, and $K_{1/4}(\cdot)$ be Bessel function of the third kind. To propose appropriate and weighted test statistics, we need to define the following:

Weighted EDF	**Effective sample size**	**Mixed sample's WEDF**
$F_n^{\boldsymbol{W}}(x) = \frac{1}{W.} \sum_{i=1}^{n} W_i \mathbf{1}_{(-\infty, X_i]}(x)$	$n_e = \frac{\left(\sum_{i=1}^{n} W_i\right)^2}{\sum_{i=1}^{n} W_i^2}$	$H_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}(x) = \frac{n_e F_n^{\boldsymbol{W}}(x) + m_e F_m^{\boldsymbol{V}}(x)}{n_e + m_e}$

Now we can present test statistics and presumed asymptotic distribution (a. d.) of generalized homogeneity tests. Details for unweighted samples are in [3] or [4], among others, and in [5] for weighted samples.

- **Kolmogorov-Smirnov test**

Test statistic $T_{n,m}^{\boldsymbol{W},\boldsymbol{V}} = \sqrt{\frac{n_e m_e}{n_e+m_e}} \sup_{x \in \mathbb{R}} \left| F_n^{\boldsymbol{W}}(x) - F_m^{\boldsymbol{V}}(x) \right|$

Presumed a. d. $K(\lambda) = 1 - 2 \sum_{k=1}^{+\infty} (-1)^{k+1} e^{-2k^2\lambda^2}$

- **Cramér-von Mises test**

Test statistic $T_{n,m}^{\boldsymbol{W},\boldsymbol{V}} = \frac{n_e m_e}{n_e+m_e} \int_{\mathbb{R}} \left(F_n^{\boldsymbol{W}}(x) - F_m^{\boldsymbol{V}}(x) \right)^2 \, \mathrm{d}H_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}$

Presumed a. d. $L_{\mathrm{CvM}}(z) = \frac{1}{\pi\sqrt{z}} \sum_{k=0}^{+\infty} (-1)^k \binom{-\frac{1}{2}}{k} \sqrt{1+4k} \, \exp\left(-\frac{(1+4k)^2}{16z}\right) K_{\frac{1}{4}}\left(\frac{(1+4k)^2}{16z}\right)$

- **Anderson-Darling test**

Test statistic $T_{n,m}^{\boldsymbol{W},\boldsymbol{V}} = \frac{n_e m_e}{n_e+m_e} \int\limits_{0 < H_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}(x) < 1} \frac{\left(F_n^{\boldsymbol{W}}(x) - F_m^{\boldsymbol{V}}(x)\right)^2}{H_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}(x)\left(1 - H_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}(x)\right)} \, \mathrm{d}H_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}$

Presumed a. d. $L_{\mathrm{AD}}(z) = \frac{\sqrt{2\pi}}{z} \sum_{k=0}^{+\infty} \binom{-\frac{1}{2}}{k} (1+4k) \exp\left(-\frac{(1+4k)^2\pi^2}{8z}\right) \int_0^{+\infty} \exp\left(\frac{z}{8(w^2+1)} - \frac{(1+4k)^2\pi^2 w^2}{8z}\right) \, \mathrm{d}w$

3. Numerical verification of presumed distributions

It is necessary to verify whether generalized test statistics have the same asymptotic distribution as the original test statistics. Since no theoretical proof of asymptotic properties has been done yet, we can demonstrate them numerically. If we consider data as random variables, distribution of test statistic is a continuous function, and if the null hypothesis is true then

$$p\text{-value} \doteq 1 - F_T\left(T_{n,m}^{\boldsymbol{W},\boldsymbol{V}}\right) \sim \mathrm{U}(0,1).$$

We carried out many experiments in which we produced two samples from different distributions and assigned them weights in such a way that their WEDFs converge to the same distribution. Afterward, we applied homogeneity tests. In figure 2 we present results from two such experiments, in which we compare two samples with 1000 observations and different weights with TH1::Chi2test. The number of bins is chosen by the rule described in [6]. It is obvious that the χ^2 test from TH1 library significantly underestimates the true p-value.

Another approach for verification of the p-value's computation uses the fact that $\mathbb{P}[p\text{-value} < \alpha] = \alpha$ where α is significance level. As α can be any value between 0 and 1, we obtain distribution function of U(0,1). We repeated the whole procedure 10 000 times with selected parameters. Then we plotted EDF of each test's p-values and compared it to CDF of U(0,1). In the figure 3 we can see results of this experiment, where the first sample

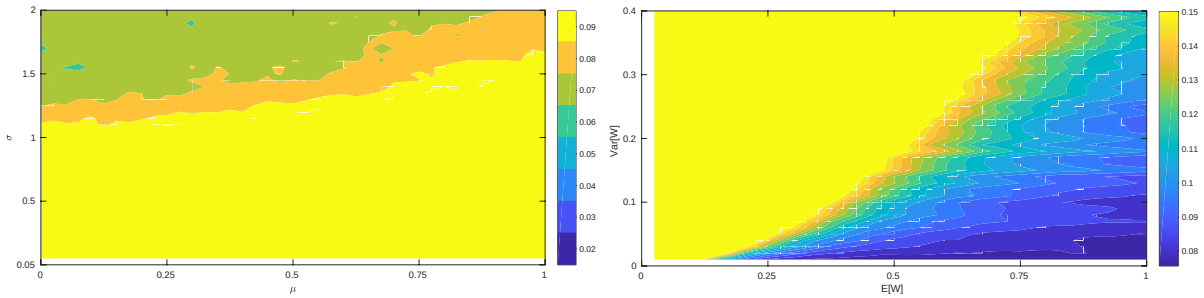

Figure 2. The ratios of rejection (on significance level 0.05) for two different experiments. In the left one, the first sample $X_i \sim \mathcal{N}(0,1)$ with weights $W_i = 1$ and the second $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ with weights $V_i = \sigma\,\varphi(Y_i)\left(\varphi\left(\frac{Y_i-\mu}{\sigma}\right)\right)^{-1}$ where φ is the standard normal distribution. In the right one, the first sample $X_i \sim \mathcal{N}(0,1)$ with weights $W_i = 1$, the second $Y_i \sim \mathcal{N}(0,1)$ and $V_i \sim \text{Gamma}(k,\theta)$ with various $\mathbb{E}[V_i]$ and $\text{Var}[V_i]$. Ratios were counted out of 10 000 repetitions.

$X_i \sim \mathcal{N}(0,1)$ with weights $W_i = 1$ while the second sample $Y_i \sim \mathcal{N}(0.3, 1.1^2)$ with weights $V_i = 1.1\,\varphi(Y_i)\left(\varphi\left(\frac{Y_i-0.3}{1.1}\right)\right)^{-1}$. While p-values from generalized KS, AD, and CvM tests are uniformly distributed if the null hypothesis is true, p-values from χ^2 are not and the type-I error is larger than it should be for arbitrary significance level.

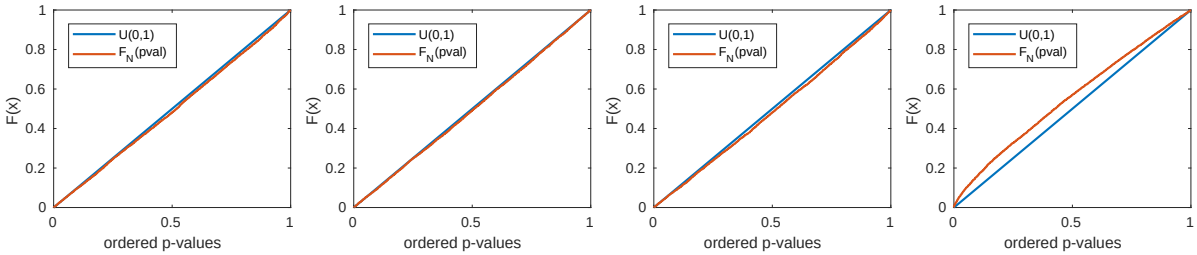

Figure 3. From left to right: Generalized KS, CvM, AD tests and χ^2 test from TH1 class.

4. Power of test comparison

Power of test, defined as a probability that we reject the test while the null hypothesis is not true, differs for various experiments' setting. We carried out another experiment in which we observed the effect of six parameters on it.

We produced two samples from $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu_s, (1+\sigma_s)^2)$. All weights of the first sample are equal to 1 while the weights of the second sample were independently generated from $\text{Gamma}(k,\theta)$. Parameters k and θ will be represented by mean (μ_w) and variance (σ_w) of weights. The first sample's size is equal to n while the other sample's is equal to $k{\cdot}n$. For every setting of (μ_s, σ_s, μ_w, μ_w, σ_w, n, k) we repeated procedure 1000 times and calculated ratio of rejected tests (r) on significance level $\alpha = 0.05$ which is power of test's estimate. In figure 4 we illustrate how the power of test behaves with shifting mean and standard deviation of the second sample. In figure 5 we present the change in power of test while changing the sample size of both samples. In all experiments, the AD test is the most powerful among the presented tests.

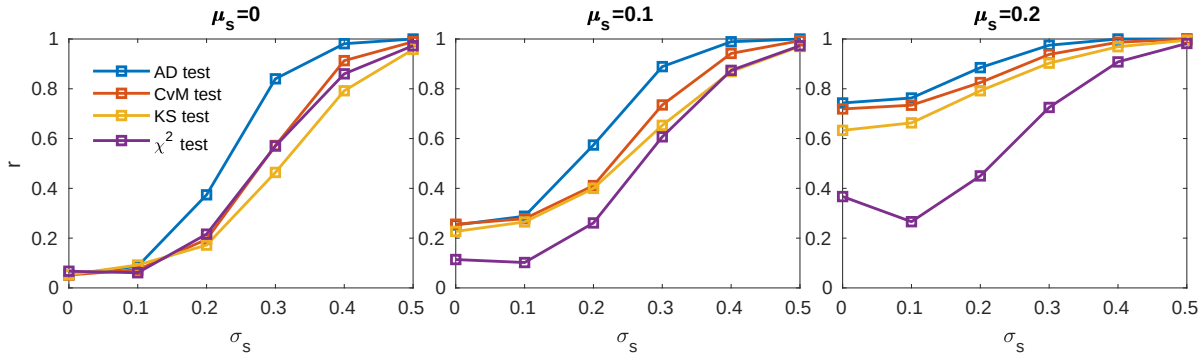

Figure 4. Parameter setting: $(\mu_w, \sigma_w, n, k) = (0.3, 0.1, 200, 10)$. AD test has the highest ratio of rejected tests for both changing parameter μ_s and σ_s. This is also true for $\mu_s = 0.3, 0.4, 0.5$.

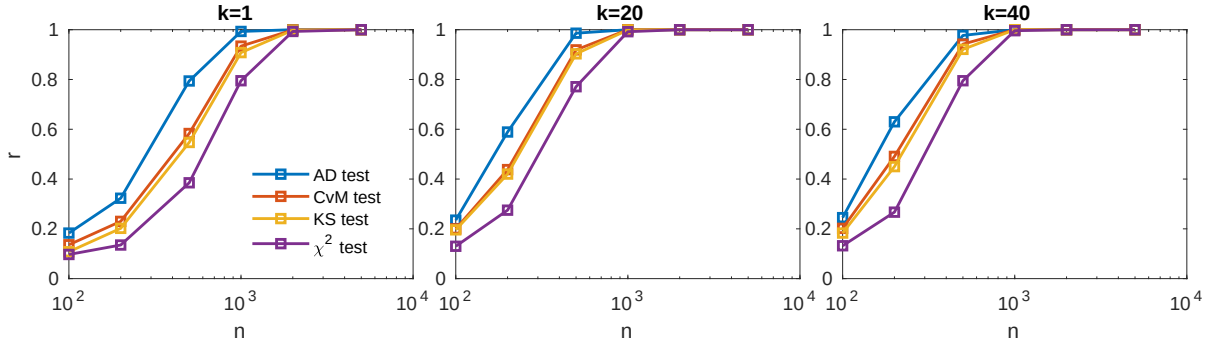

Figure 5. Parameter setting: $(\mu_s, \sigma_s, \mu_w, \sigma_w) = (0.1, 0.2, 0.4, 0.01)$. AD test has again the highest ratio of rejected tests for both changing parameter k and n.

5. Results

In this paper, we have presented part of results from the large simulation study that verified the correctness of test statistics for the proposed generalized homogeneity test. All three generalized tests have been implemented in C++ (ROOT) and the code can be downloaded from [7]. The integration of this code into the new version of ROOT is in the process. For weighted data samples originating from a continuous distribution, we recommend to use presented generalized AD test, which is more powerful than KS or CvM in most cases.

Acknowledgments

We acknowledge support from the Czech CTU grant SGS18/188/OHK4/3T/14 and MEYS grants LM2015068 and LTT18001.

References

[1] ROOT project "ROOT" [software] version 6.04.00 2015 Available from `https://github.com/root-project/root/releases/tag/v6-04-00` [accessed 2019-05-28]

[2] Monahan J F 2011 *Numerical Methods of Statistics* (Cambridge University Press) p 358 2nd ed

[3] Massey Jr F J 1951 *J. Amer. Statist. Assoc.* **46** 68–78

[4] Anderson T W and Darling D A 1952 *Ann. Math. Statist.* **23** 193–212

[5] Trusina J 2019 *Application of homogeneity testing and event classification in neutrino physics* diploma thesis Czech Technical University in Prague

[6] D'Agostino R and Stephens M 1986 *Goodness-of-Fit Techniques* (New York: Dekker) p 70

[7] Trusina J "Generalized KS, CvM, and AD homogeneity tests" [software] 2019 Available from `http://gams.fjfi.cvut.cz/homtests` [accessed 2019-05-28]