Computer simulation on homogeneity testing for weighted data sets used in HEP

Petr Bouř and Václav Kůs

Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Trojanova 13, 12000 Praha 2, Czech Republic

E-mail: petr.bour@fjfi.cvut.cz

Abstract. Modified statistical homogeneity tests on weighted data samples are commonly used in high energy physics applications. We do typically apply the tests in order to test homogeneity of weighted and unweighted samples, e.g. Monte Carlo simulations compared to the real data measurements. The asymptotic approximation of p-value of our weighted variants of homogeneity tests are investigated by means of simulation experiments. The simulation is performed for various probability sample distributions. We show that the asymptotic characteristics of the weighted homogeneity tests are valid for the specific distribution of weights.

1. Introduction

In high energy physics, homogeneity testing precedes many analysis and modeling techniques, particularly in machine learning (ML) applications. It is the case when we apply a data preprocessing procedure called data weighting. Via assigning weights $w_1, \ldots, w_n > 0$ to simulated observations $x_1, \ldots, x_n$, $n \in \mathbb{N}$, we are able to fine-tune our Monte Carlo (MC) simulation data set with respect to our requirements. However, statistical theory concerning homogeneity tests does not handle any weighting procedures, nor associates weights with observations, except for sporadic works on weighted histograms, as e.g. [1]. Therefore, the classical homogeneity tests must be adjusted for weighted data sets. Despite relatively straightforward incorporation of weights into the classical homogeneity tests and their modification, asymptotic properties of these tests can be no longer guaranteed. Thus, our goal is to investigate the validity of asymptotic properties of homogeneity testing for weighted observations through computer simulation.

The need for homogeneity testing may typically arise from a simple signal/backgrounds binary classification task. In this common ML application, we often use MC simulation for both ML classifier training and testing. We may then apply the trained classifier to real measured data set (DATA). Naturally, we expect both MC $\sim F$ and DATA $\sim G$ to be identically distributed: $F \equiv G$. Otherwise, the classification model will not perform well.

2. Weighted tests of homogeneity

Let us assume it is vital to guarantee homogeneity of DATA and MC distributions prior to subsequent utilization of some ML methods. For this purpose, we first define an analogy with empirical distribution function (EDF) for weighted data set. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be iid random variables distributed by cumulative distribution function (CDF) $F(x)$ and let

$(w_1, \ldots, w_n)$ be their corresponding weights, where $W = \sum_{i=1}^{n} w_i$. We define the weighted empirical distribution function (WEDF) to be

$$F_n^W(x) = \frac{1}{W} \sum_{i=1}^{n} w_i I_{(-\infty,x]}(X_i), \tag{1}$$

where $I_A(X)$ is the indicator of the set A. Note that in the case of $w_i = 1$ for all $i = 1, \ldots, n$ (i.e. unweighted DATA), the definition of WEDF goes over to usual EDF.

In order to avoid an investigation of an unknown parametric family, we shall pursue our homogeneity testing only with nonparametric approaches. Thus, proceeding further in this section, we present the Kolmogorov-Smirnov test based upon EDFs of two data sets $\boldsymbol{X}_1 = \left(X_1^{(1)}, \ldots, X_{n_1}^{(1)}\right)$, $\boldsymbol{X}_2 = \left(X_1^{(2)}, \ldots, X_{n_2}^{(2)}\right)$, with respective distribution functions F, G. By the homogeneity hypothesis [3], as our null hypothesis is H_0, we understand

$$H_0 : F = G \quad \text{vs} \quad H_1 : F \neq G \quad \text{at significance level} \quad \alpha \in (0, 1). \tag{2}$$

We require our homogeneity tests to meet the condition $P(W_C|H_0) \leq \alpha$, where W_C is a critical region for the specific test statistic T. We reject hypothesis H_0 if $T \in W_C$. The nature of homogeneity testing prompted us to look for the p-value, i.e., the lowest significance level α for which we reject hypothesis H_0. Thus, for every $\alpha >$ p-value we may automatically reject hypothesis H_0.

2.1. Two sample Kolmogorov-Smirnov test

Let F_{n_1}, G_{n_2} denote the EDFs of two data samples $\boldsymbol{X}_1, \boldsymbol{X}_2$ with respective sample sizes n_1, n_2. We consider the test statistic

$$D_{n_1,n_2} = \sup_{x \in \mathbb{R}} |F_{n_1}(x) - G_{n_2}(x)|. \tag{3}$$

It is clear from the Glivenko-Cantelli lemma [2] that, under the true H_0, it holds that $D_{n_1,n_2} \xrightarrow{a.s.} 0$ for $n_1, n_2 \to \infty$. Furthermore, due to [4], it follows that for the true H_0 and $\lambda > 0$

$$\lim_{n_1,n_2 \to \infty} P\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1,n_2} \leq \lambda\right) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2\lambda^2}. \tag{4}$$

Therefore, we obtain the approximate p-value as $2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2\lambda_0^2}$, where $\lambda_0 = \sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2}$. However, for weighted data sample we are forced to replace EDFs F_{n_1}, G_{n_2}, and the numbers of entries n_1, n_2, with their respective WEDFs $F_{n_1}^{W_1}, G_{n_2}^{W_2}$, and the sums of weights W_1, W_2 in (3) and (4). Instead of (3), we thus obtain the test statistic

$$D_{n_1,n_2}^{W_1,W_2} = \sup_{x \in \mathbb{R}} \left|F_{n_1}^{W_1}(x) - G_{n_2}^{W_2}(x)\right|. \tag{5}$$

The definition (1) of WEDF makes it clear that the statistic $D_{n_1,n_2}^{W_1,W_2} \xrightarrow{a.s.} 0$ for $n_1, n_2 \to \infty$ and $W_1, W_2 \to \infty$. Nevertheless, it is important to notice some of the weaknesses inherent in the above approach. This modified test for the weighted data sample does not have to obey the asymptotic property (4). Let us emphasize that the p-value obtained using the statistic $D_{n_1,n_2}^{W_1,W_2}$ cannot be considered a regular approximate p-value without subsequent detailed research. At this point, we are not able to present a rigorous mathematical proof. However, we intent to supply the HEP community with some recommendations on the suitability of the weighted tests. This is why we propose partial validation of our approach in section 3. A key insight we provide is the stable numerical verification performed for a few fundamental data distributions.

3. Simulation experiments

We have already mentioned the problem of insecure asymptotic properties when applying weighted modifications of the standard tests. We now turn our attention at the numerical simulation. For our purposes here, the best way would be to validate the asymptotic properties using standard unweighted tests. This requires us to plug into the testing an unweighted data set (only instead of weighted MC; DATA is unweighted already). We shall do this by appropriate transformation of the weighted ensemble MC into the unweighted ensemble MC$_\dagger$ [5].

3.1. Rearranging technique

We make two requirements for the transformation. Firstly, we desire to preserve or exploit information contained in MC weighting since the weight assigned to an observation refer to what extent the distribution should be present in the neighbourhood of this observation. Secondly, we require that the sum of weights in MC corresponds to the number of observations in the unweighted MC$_\dagger$. Continuing in this manner, we now proceed as follows. Denote by $\boldsymbol{X} = \left(X_{(1)}, \dots, X_{(n)}\right)$ the ordered sample in MC with weights $(w_1, \dots, w_n)$ and let $W = \sum_{i=1}^{n} w_i$. Let $N = \lfloor W \rfloor$ denote the desired number of observations in the new transformed ensemble MC$_\dagger$. Given both our requirements regarding MC$_\dagger$, we are constructing special weighted averages from $\boldsymbol{X}$. For simplicity, we presume $0 \leq w_i \leq 1$ for all $i = 1, \dots, n$. Into the set of the first weighted average we include the smallest possible number of observations $\left(X_{(1)}, \dots, X_{(k_1)}\right)$ such that $1 \leq \sum_{i=1}^{k_1} w_i < 2$. Thereby, $\sum_{i=1}^{l} w_i < 1$ for all $l < k_1$. The portion of weight w_{k_1} of the observation $X_{(k_1)}$ which contributes above 1 to the sum $\sum_{i=1}^{k_1} w_i$ will not be included into the first weighted average. Hence, we denote this residual portion by $r_{k_1} = \sum_{i=1}^{k_1} w_i - 1$. Thereafter, the first observation $Y_{(1)}$ in MC$_\dagger$ can be defined as the following weighted average

$$Y_{(1)} = \frac{\sum_{i=1}^{k_1} X_{(i)} w_i - X_{(k_1)} r_{k_1}}{\sum_{i=1}^{k_1} w_i - r_{k_1}}. \tag{6}$$

From the definition of r_{k_1} we arrive at

$$Y_{(1)} = \sum_{i=1}^{k_1} X_{(i)} w_i - X_{(k_1)} r_{k_1} = \sum_{i=1}^{k_1 - 1} X_{(i)} w_i + X_{(k_1)} (w_{k_1} - r_{k_1}). \tag{7}$$

The residual portion r_{k_1} will be added to the next weighted average for $Y_{(2)}$. In general, for $Y_{(j)}$ we write

$$r_{k_j} = \sum_{i=k_{j-1}+1}^{k_j} w_i - r_{k_j - 1} - 1 \tag{8}$$

$$Y_{(j)} = X_{(k_{j-1})} r_{k_{j-1}} + \sum_{i=k_{j-1}+1}^{k_j - 1} X_{(i)} w_i + X_{(k_j)} (w_{k_j} - r_{k_j}). \tag{9}$$

Repeating the same steps we transform the original weighted ensemble MC with $\boldsymbol{X} = \left(X_{(1)}, \dots, X_{(n)}\right)$ into the new unweighted ensemble MC$_\dagger$ with $\boldsymbol{Y} = \left(Y_{(1)}, \dots, Y_{(\widetilde{n})}\right)$. We have distributed the weights from the MC so that there is the unit weight for each observation $Y_{(j)}$. Therefore, we are authorized to apply standard homogeneity tests, which guarantees the asymptotic properties.

3.2. Generic p-value validation

We have verified an eligible usage of modified weighted tests in previous section with data sets originating from high energy physics [5]. Now, we provide a reader with more general verification. In our simulation, we consider several different distributions for $\boldsymbol{X} = (X_1, \ldots, X_n)$: Beta, Cauchy, Exponential, Laplace, Logistic, Lognormal, Normal, Uniform and Weibull, whilst the weights $\boldsymbol{W} = (W_1, \ldots, W_n)$ are taken from the Beta distribution, as we can easily tune the expected value,

$$W \sim Beta(\alpha, \beta) \implies E\left[W\right] = \frac{\alpha}{\alpha + \beta}. \tag{10}$$

The appropriate number of data points was determined by preliminary convergence studies. Otherwise, the simulation steps proceed as follows:

- (i) Generate n random weighted data points $(\boldsymbol{X}, \boldsymbol{W})$, e.g. $n = 3500000$.
- (ii) Estimate weighted distribution from all the observations $(\boldsymbol{X}, \boldsymbol{W})$ using kernel density estimator.
- (iii) Repeat all the following items k times, e.g. $k = 1000$:
  - (a) Choose $m_w = \frac{n}{k}$ weighted observations from $(\boldsymbol{X}, \boldsymbol{W})$ as the current MC sample, e.g. $m_w = 3500$.
  - (b) Generate $m_u \approx \sum_{i=1}^{m_w} w_i$ unweighted observations from the estimated weighted distribution and consider them to be the current DATA sample, e.g. $m_u = 1000$.
  - (c) Apply weighted homogeneity test MC vs DATA.
  - (d) Rearrange MC into unweighted sample MC$_\dagger$ and apply standard unweighted test.

Thus, we obtain k individual p-values from the weighted tests and also another k corresponding p-values from the unweighted tests. We may now check asymptotic properties of both weighted and unweighted tests.

For all the distributions under consideration we arrived at two main results. First, the condition $P(W_C|H_0) \leq \alpha$ for significance level α is uniformly satisfied as it is shown in figure 1, i.e. both EDFs are located under diagonal line in the graph. Second, both weighted modifications and unweighted tests have the same resulting p-value distribution. This was tested via common classical homogeneity tests for unweighted data. Nevertheless, the extraordinary correspondence is already obvious from the graph in figure 1. Quite similar results were achieved for all other simulated distributions: Beta, Cauchy, Exponential, Laplace, Logistic, Normal, Uniform and Weibull.

4. Discussion

We carried out a numerical validation of the modified statistical homogeneity tests for generated data sets under Beta-weighting. Our simulation verifies that the approximate asymptotic properties remain the same for both weighted and unweighted tests. In practice, we may either utilize modified weighted tests or we may apply the rearranging technique from section 3.1 directly with the unweighted standard tests (where the asymptotics are theoretically proven). However, our verification was performed for the specific choice of weights distribution only. In the future research, we aim to investigate the effect of various homogeneity tests and different weights distribution on the overall significance and power. We also plan to explore the possibility of theoretic validation of some representative weighted tests for wide-ranging data distribution families. It may be reached by appropriate restrictions imposed on the distribution of weights, since, in practice, there exists only a limited number of high energy physics applications under rather specific weighting.

Please note we only aimed to provide empirical recommendations for weighted tests of homogeneity. A reader may use these only for an experimental setup with data and weights distributions similar to ours. On the other hand, this may cover large number of cases as HEP data often come from these fundamental distributions or their mixture. In any case, the reader may always utilize our validation procedures to verify whether the output of his test corresponds to a regular approximate p-value.

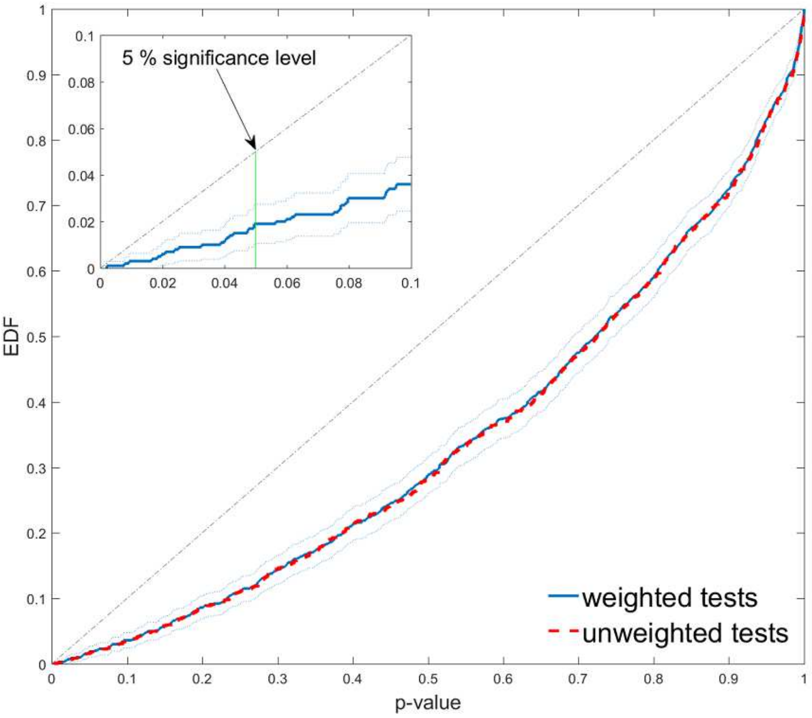

Figure 1. EDF of p-value for weighted and unweighted tests of homogeneity. Homogeneity of p-value distributions was tested via common classical homogeneity tests for unweighted data. Underlying data are taken from the lognormal distribution.

Acknowledgments

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

References

[1] Gagunashvili N D 2015 Chi-square goodness of fit tests for weighted histograms. Review and improvements *JINST* **10** P05004

[2] van der Vaart A W 1998 *Asymptotic Statistics* (Cambridge: Cambridge University Press) p 265

[3] DeGroot M H and Schervish M J 2010 *Probability and Statistics* (Boston: Pearson) p 647

[4] Smirnov N J 1944 *Usp. Mat. Nauk.* **10** 179–206

[5] Bouř P 2016 *Development of statistical nonparametric and divergence methods for data processing in D0 and NOvA experiments* (Prague: FNSPE CTU) p 66