Simulation on Homogeneity Testing for Weighted Data Sets

Petr Bouř, Václav Kůs
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague
petr.bour@fjfi.cvut.cz

Summary

We introduce several modifications of classical statistical tests applicable to weighted data sets 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 and power of our weighted variants of homogeneity tests are investigated by means of simulation experiments. The simulation is performed for various probability distributions of samples. Finally, our methods of homogeneity testing are applied to Monte Carlo samples and real DATA sets measured at the particle accelerator Tevatron in Fermilab at DZero experiment originating from top-antitop quark pair production in two decay channels (electron, muon) with 2, 3, or 4+ jets detected. Consequently, the final variable selection is carried out and the resulting subsets chosen from 46 dimensional physical parameters are recommended for further top quark cross section analysis. Our variable selections differ from the ROOT TMVA based selection for the top-antitop decay channels.

I. INTRODUCTION

Homogeneity testing is an important step in many analysis techniques, particularly in machine learning (ML) applications in physics research. However, statistical homogeneity tests need to be adjusted for weighted datasets. Our goal is to verify the validity of asymptotic properties of homogeneity testing for weighted observations.

Example utilization for ML applications

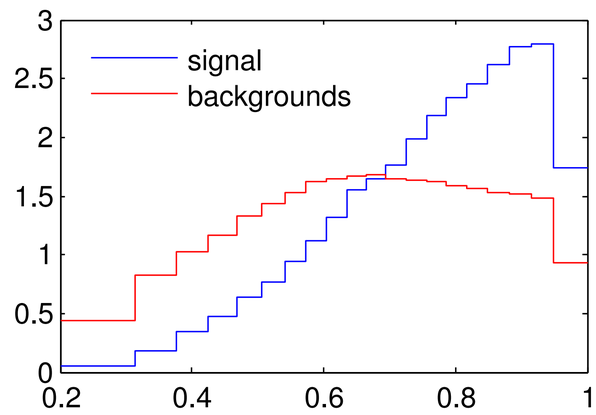


Figure: The underlying problem may be a signal/backgrounds classification task.

- We often use a Monte Carlo simulation dataset (MC) for both training and testing our ML classifier. We may then apply the trained classifier to a real measured dataset (DATA).
- 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.
- Thus, we need to test homogeneity of MC and DATA prior to the modelling.

Homogeneity testing

$H_0 : F = G \quad \text{vs} \quad H_1 : F \neq G \quad \text{for } \alpha \in (0,1)$

- Significance level: $\alpha = P(W|H_0)$
- Critical region: W (rejection of H_0)
- We demand: $P(W|H_0) \leq \alpha$

Two sample Kolmogorov-Smirnov test:

$$p\text{-}value = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2\lambda^2},$$

- $\lambda = \sqrt{\frac{nm}{n+m}}\sup_{x\in\mathbb{R}}|F_n(x) - G_m(x)|$

Weighting

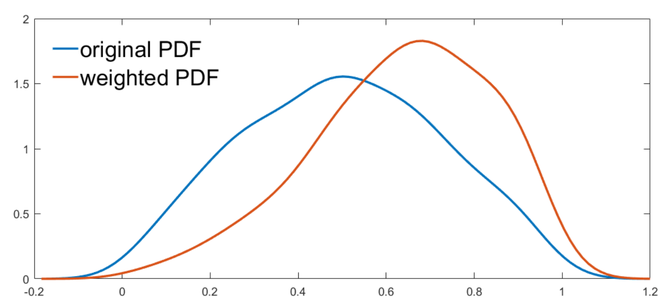


Figure: Example of weighting. Weighted distribution is positively skewed.

How does the asymptotics change when using weights? $(x_1, w_1), \ldots, (x_n, w_n)$:

$$n \mapsto W_F, W_F = \sum_{i=1}^{n} w_i \implies \lambda = \lambda(W_F, W_G)$$

II. RESULTS

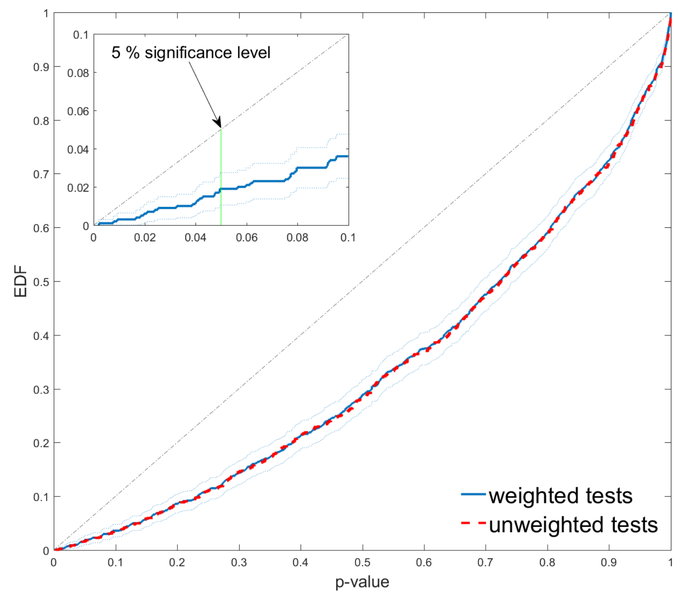
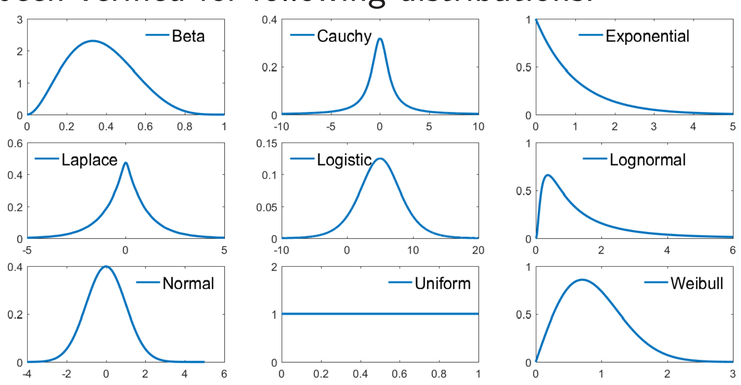


Figure: Empirical distribution functions (EDF) of p-value for weighted and unweighted tests of homogeneity (underlying data are taken from lognormal distribution).

1. Significance level condition is uniformly satisfied:
$$P(W|H_0) \leq \alpha$$
2. Both weighted modifications and unweighted tests have the same p-value distribution.

Probability distributions in simulation

Weighted modifications of homogeneity tests have been verified for following distributions:



Rearranging technique

- Technique aggregates weighted observations into data points with the weight $w_i = 1$.
- By using the weighted average and redistributing the weights, we obtain an equivalent but unweighted dataset.
- Local information from data is preserved.
- Thus, rearranging technique may be used directly with the unweighted standard tests.
- ...where the asymptotics are proven!

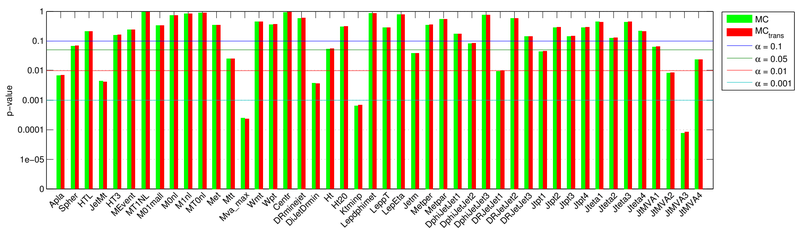


Figure: Modified weighted Kolmogorov-Smirnov test: p-value for all variables in the weighted MC channel Electron 4+ Jets and the rearranged unweighted ensemble.

III. APPROACH

Weights W are taken from the Beta distribution as we may easily tune the expected value:

$$W \sim Beta(\alpha, \beta) \implies E[W] = \frac{\alpha}{\alpha+\beta}$$

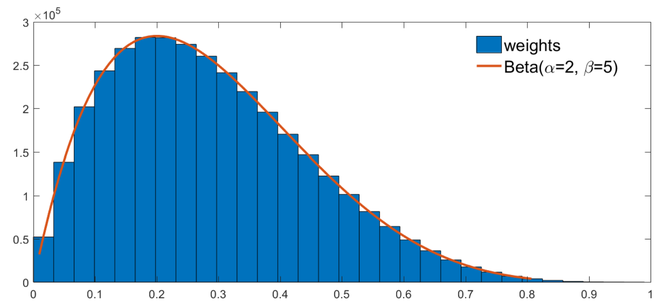


Figure: Histogram of weights used in the simulation.

- The appropriate number of simulation data points was determined by preliminary convergence studies.

Simulation steps

1. Generate n random weighted data points (X, W), e.g. $n = 3,500,000$.
2. Estimate weighted distribution from all the observations (X, W) (using kernel density estimation). Repeat all the following k times, e.g. $k = 1,000$:
3. Draw $m_w = \frac{n}{k}$ weighted observations (X, W) as your current MC sample, e.g. $m_w = 3,500$.
4. Generate $m_u \approx \sum_{i=1}^{m_w} w_i$ unweighted observations from estimated weighted distribution as your current DATA sample, e.g. $m_u = 1,000$.
5. Apply weighted homogeneity test MC vs DATA.
6. Rearrange MC into unweighted sample and apply standard unweighted test.

$\longrightarrow$ Thus, we obtain k p-values from the weighted tests and also k p-values from the unweighted tests.

IV. DISCUSSION

- effect of various homogeneity tests

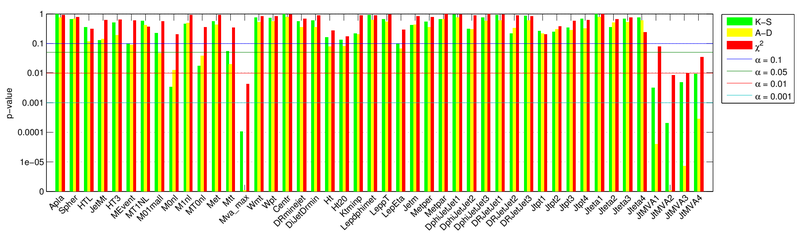


Figure: Homogeneity tests of MC vs DATA univariate distributions: p-value for all variables in the MC channel Muon 4+ Jets.

- effect of weights distribution (!)
- different motivations for weighting
- validity for arbitrary distribution
- multivariate testing