Simulation on Homogeneity Testing for Weighted Data Sets

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

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

![Graph showing signal and backgrounds](image)

- 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 ~ F and DATA ~ G to be identically distributed: F ≡ 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

\( H_0 : F = G \) vs \( H_1 : F \neq G \) 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:

\[
p\text{-value} = 2 \sum_{k=1}^{\infty} \left( -1 \right)^{k-1} e^{-2k^2\lambda^2} \]

\[
\lambda = \sqrt{\frac{m}{n}} \sup_{x \in \mathbb{R}} |F(x) - G(x)|
\]

Weighting

![Graph showing original PDF vs weighted PDF](image)

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

\[
n \rightarrow W_f, W_r = \sum_{i=1}^{n} w_i \implies \lambda = \lambda(W_f, W_r)
\]

II. RESULTS

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

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

![Graph showing empirical distribution functions (EDF) of p-value for weighted and unweighted tests of homogeneity (underlying data are taken from lognormal distribution).](image)

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:

- Beta
- Cauchy
- Exponential
- Laplace
- Lognormal
- Normal
- Uniform
- Weibull

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!

![Graph showing homogeneity tests of MC vs DATA univariate distributions: p-value for all variables in the MC channel Electron 4+ Jets and the rearranged unweighted ensemble.](image)

III. APPROACH

- 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 \( m_w = \frac{W}{\sum W} \) weighted observations \((X, W)\) as your current DATA sample, e.g. \( m_w = 3,500 \).
4. Generate \( m_w \approx \sum m_i \) unweighted observations from estimated weighted distribution as your current DATA sample, e.g. \( m_w = 1,000 \).
5. Apply weighted homogeneity test MC vs DATA.
6. Rearrange MC into unweighted sample and apply standard unweighted test.

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

![Graph showing homogeneity tests of MC vs DATA univariate distributions: p-value for all variables in the MC channel Electron 4+ Jets and the rearranged unweighted ensemble.](image)