Homogeneity tests in high energy physics

Adam Novotný

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

June 21, 2018

Overview

Motivation

- modify and verify four homogeneity tests on weighted samples:
  - Kolmogorov–Smirnov
  - Anderson–Darling
  - Cramer–von Mises
  - Pearson χ^2
- application of these homogeneity tests on data from DØ experiment:
  - data (unweighted)
  - Monte Carlo simulated samples (weighted)

Homogeneity tests

- nonparametric tests, testing whether two samples come from the same distribution
- crucial to find
  - test statistic
  - limiting distribution
- test statistic and limiting distribution allow to evaluate the test
- output of every test is p-value

Tests working with EDF

- Kolmogorov–Smirnov (nonintegral)
- Anderson–Darling (integral)

$$W^2_{m,n} = N \int_{B_{m+n}} \frac{(F_m(x) - G_n(x))^2}{H_{m+n}(x)(1 - H_{m+n}(x))} \, \mathrm{d}H_{m+n}(x) \tag{1}$$

- Cramer–von Mises (integral)

$$T^2_{m,n} = N \int_{B_{m+n}} (F_m(x) - G_n(x))^2 \, \mathrm{d}H_{m+n}(x) \tag{2}$$

where $B_{m+n} = \{x \in \mathbb{R} : H_{m+n} < 1\}$ and $H_{m+n}(x) = \frac{mF_m(x)+nG_n(x)}{m+n}$

Anderson–Darling test

- we find the limiting distribution using Laplace transformation

$$L_{\mathrm{AD}}(z) = \frac{\sqrt{2\pi}}{z} \sum_{i=0}^{\infty} \binom{-\frac{1}{2}}{i} (4i+1) \exp\left(-\frac{(4i+1)^2\pi^2}{8z} \right) \int_0^{\infty} \exp\left(\frac{z}{8(w^2+1)} - \frac{(4i+1)^2\pi^2 w^2)}{8z} \right) \mathrm{d}w \quad (3)$$

- we find p-value

$$p\text{-value} = 1 - L_{\mathrm{AD}}(W^2_{m,n}) \quad (4)$$

Cramer-von Mises test

- we find the limiting distribution using Laplace transformation

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

$$K_\nu(y) = \frac{\pi^{1/2} y^\nu}{2^\nu \Gamma(\nu + \frac{1}{2})} \int_0^\infty \sinh^{2\nu}(u) \exp(-y\cosh(u))\,\mathrm{d}u, \quad y > 0, \tag{5}$$

- we find p-value

$$p\text{-value} = 1 - L_{\mathrm{CvM}}(T^2_{m,n}). \tag{6}$$

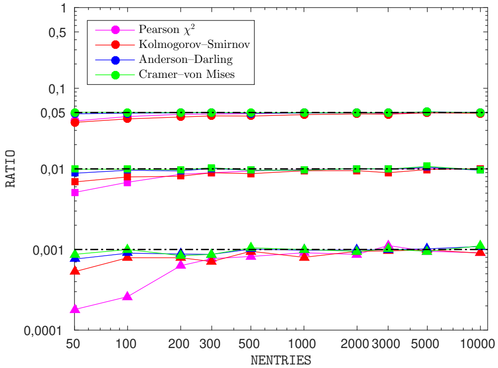
Modification of homogeneity tests

- replace EDF (empirical distribution function) with WEDF (weighted empirical distribution function)
- use effective sampling

Verification of modified tests

- verify tests on generated entries from different distributions and weights from different distributions
- two samples with entries from $\mathcal{N}(0, 1)$, weights from Γ distribution
- repeat each test 100,000 times
- `RATIO` - ratio of rejected tests to not rejected tests, `NENTRIES` - number of entries
- log-log plot
- power of test

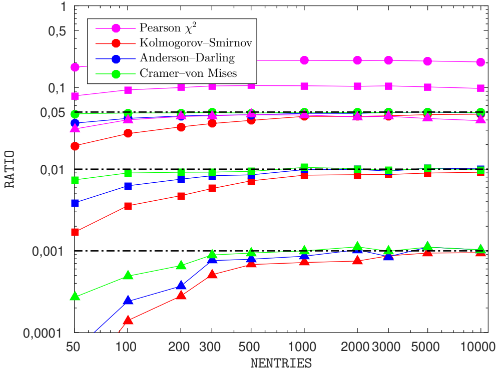
Table: Asymptotic behaviour of homogeneity tests applied to generated entries from $\mathcal{N}(0,1)$ and weights from $\Gamma(10,0.1)$.



RATIO - ratio of rejected tests to not rejected tests

NENTRIES - number of entries

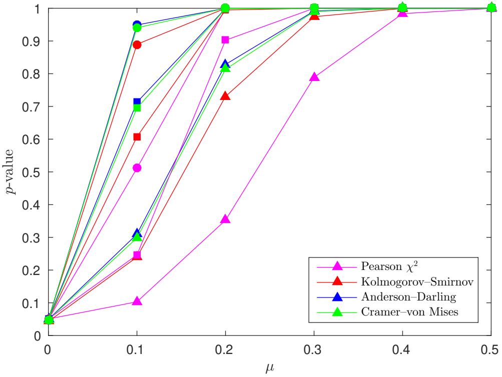
Table: Asymptotic behaviour of homogeneity tests applied to generated entries from $\mathcal{N}(0,1)$ and weights from $\Gamma(0.25,1)$.



`RATIO` - ratio of rejected tests to not rejected tests

`NENTRIES` - number of entries

Table: Power of test of homogeneity tests applied to generated entries from $\mathcal{N}(0,1)$ and weights from $\Gamma(10, 0.1)$ with different number of entries.



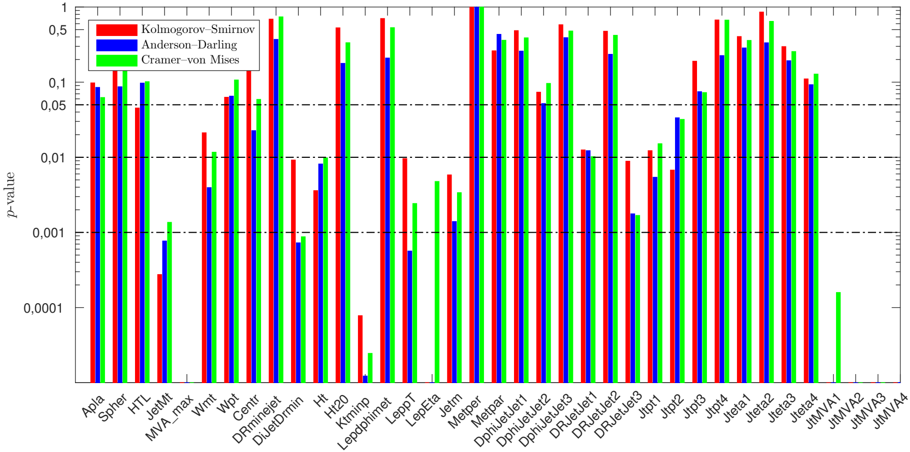
NENTRIES - 500, 1,500, 3,000

Experiment DØ

- two datasets:
  - data (unweighted)
  - Monte Carlo simulated samples (weighted)
- each dataset divided into 6 channels
- 37 variables

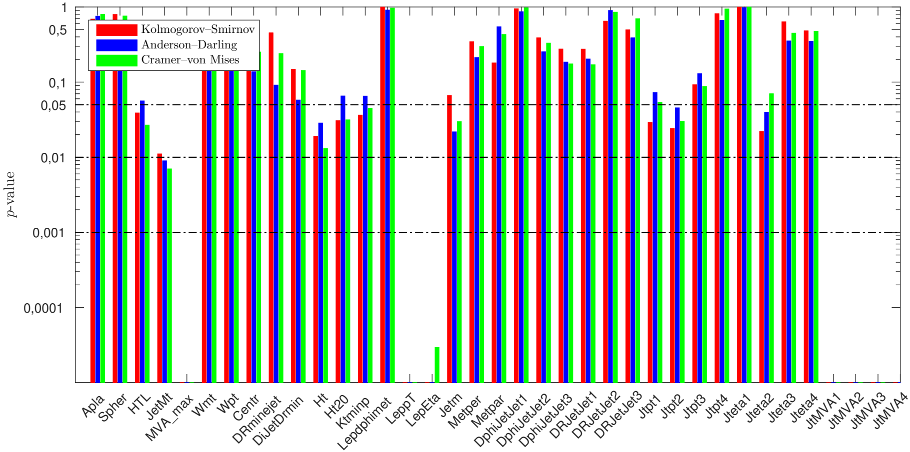
Results of tests applied to DØ datasets

Table: Homogeneity tests applied to data and Monte Carlo in ele+4jets channel.



Results of tests applied to DØ datasets

Table: Homogeneity tests applied to data and Monte Carlo in muo+4jets channel.



Conclusion

- we modified and verified four homogeneity tests on weighted samples
- we applied these homogeneity tests on data from DØ experiment

Thank you for your attention