Program for statistical comparison of histograms

S. Bityukov (IHEP, INR), N. Tsirova (NPI MSU)

Scope

- **Introduction**
- **Statistical comparison of two histograms**
- **Normalized significance**
- **Example**
- **"Distance measure"**
- **Internal resolution of the method**
- **Script stat_analyzer.C**
- **Missing E_T: large difference between histograms**
- **P_T light jet: only statistical difference**
- **Output of the script**
- **Conclusion**

Introduction

Sometimes we have two histograms and are faced with the question: Are the realizations of random variables in the form of these histograms taken from the same statistical population?

There are many approaches for comparison of two histograms: Kolmogorov-Smirnov test, Cramer-von-Mises test, Anderson-Darling test, Likelihood ratio test for shape and for slope, ...

The review on this subject is in the paper
“Testing Consistency of Two Histograms”
by F. Porter.

http://arxiv.org/abs/0804.0380

Usually, a special test-statistic is used for comparing the two histograms
(for example, χ^2, NIM, A614 (2010) 287 **).**

Statistical comparison of two histograms

A method, which uses the distribution of test-statistics for each bin of histograms, is proposed in note arXiv:1302.2651.

These test-statistics obeys the distribution close to standard normal distribution if both histograms are obtained from the same statistical population by the same analyzer of data. Correspondingly, the joint distribution must be close to standard normal distribution.

For example, if we consider observed values which are produced by the same Poisson flow of events during equal independent time ranges then a test statistic

$$\hat{S}_{c12} = 2 \cdot (\sqrt{\hat{n}_{i1}} - \sqrt{\hat{n}_{i2}}),$$

where $\hat{n}_{i1}$ is a number events in bin i of histogram 1, $\hat{n}_{i2}$ is a number events in bin i of histogram 2, is the test statistic of such type. It is shown in paper PoS(ACAT08)118.

Often test-statistics of such type are named as "significances of a deviation" or "significances of an enhancement".

Normalized significance

Let us consider a model with two histograms where the random variable in each bin obeys the normal distribution

$$\varphi(x|n_{ik}) = \frac{1}{\sqrt{2\pi}\sigma_{ik}}\ e^{-\frac{(x-n_{ik})^2}{2\sigma_{ik}^2}}. \qquad (1)$$

Here the expected value in the bin i is equal to n_{ik} and the variance $\sigma^2_{n_{ik}}$ is also equal to n_{ik}. k is the histogram number ($k = 1, 2$).

Let integral in the histogram 1 equals N_1 and integral in the histogram 2 equals N_2 (for example, if we have different integrated luminosity in experiments).

We define the normalized significance as

$$\hat{S}_i(K) = \frac{\hat{n}_{i1} - K\hat{n}_{i2}}{\sqrt{\hat{\sigma}^2_{n_{i1}} + K^2\hat{\sigma}^2_{n_{i2}}}}. \qquad (2)$$

Here $\hat{n}_{ik}$ is an observed value in the bin i of the histogram k, $\hat{\sigma}^2_{n_{ik}} = \hat{n}_{ik}$ and coefficient of normalization $K = \dfrac{N_1}{N_2}$.

Example

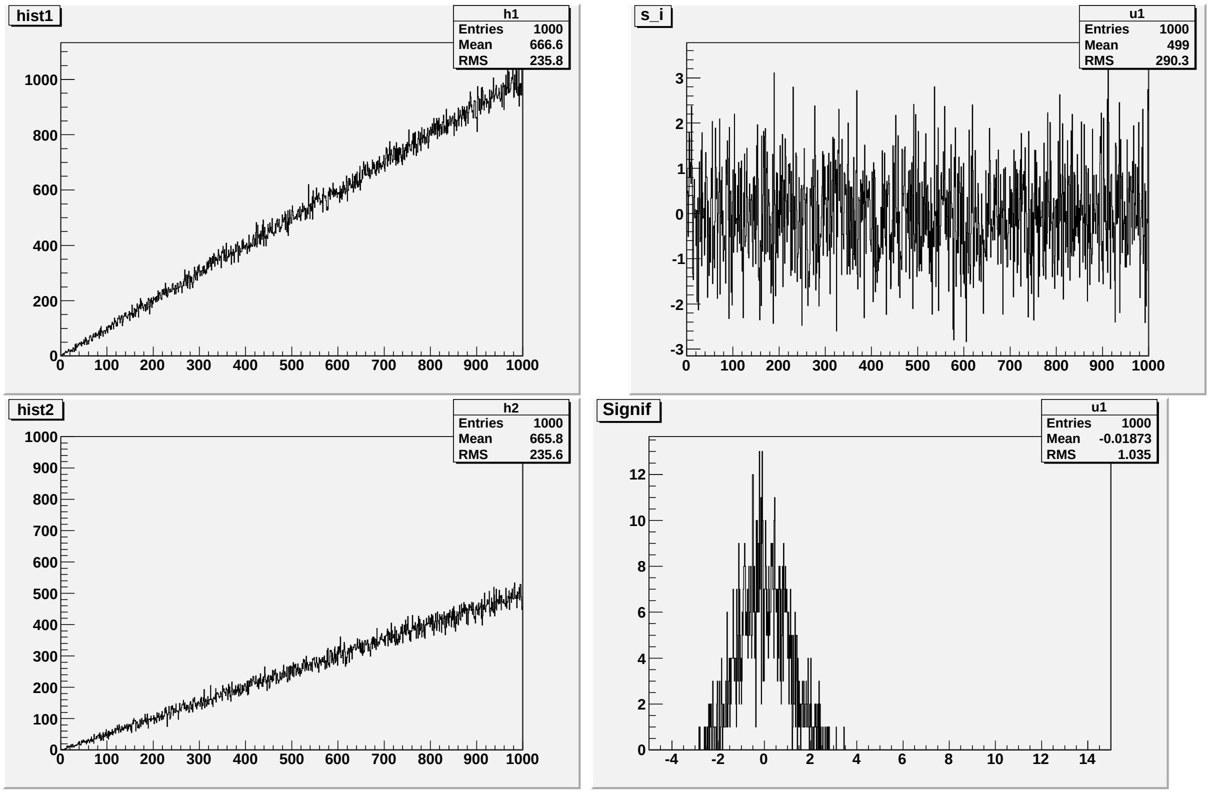


Figure 1: Triangle distributions ($K = 2$, $M = 1000$): the observed values $\hat{n}_{i1}$ in the first histogram (left,up), the observed values n_{i2} in the second histogram (left, down), observed normalized significances $\hat{S}_i$ bin-by-bin (right, up), the distribution of observed normalized significances (right, down).

The example with histograms produced from the same events flow during unequal independent time ranges shows that the standard deviation (RMS in the ROOT notation) of the distribution in the picture (right, down) can be used as estimator of the statistical difference between histograms (this distribution is close to standard normal distribution in our example).

“Distance measure”

The RMS as a “distance measure” in our case has a clear interpretation (in fact, we set a scale of this “differmeter”):

- $RMS = 0$ – **histograms are identical;**
- $RMS \sim 1$ – **both histograms are obtained (by the using independent samples) from the same parent distribution;**
- $RMS >> 1$ – **histograms are obtained from different parent distributions.**

An accuracy (internal resolution) of the method depends on the number of bins M in histograms and on the normalized coefficient K. The accuracy can be estimated via Monte Carlo experiments.

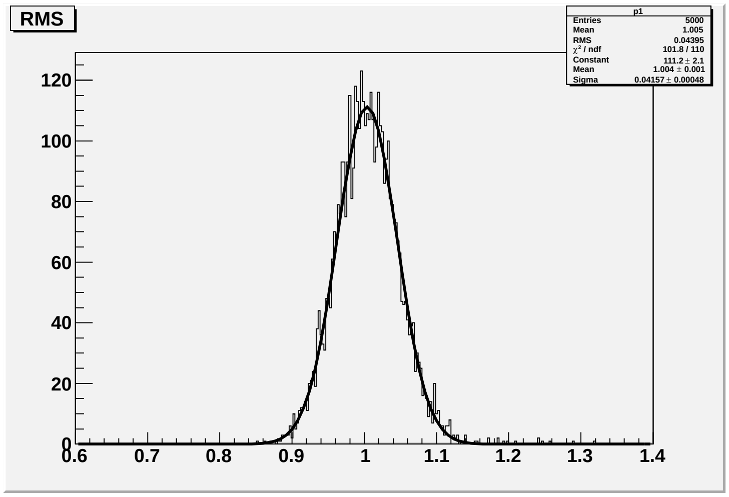


Figure 2: Distribution of RMS for 5000 comparisons of histograms (triangle distribution, $K = 1$, $M = 300$).

Internal resolution of the method

The dependencies of the internal resolution on the bin numbers M and on the value of coefficient of normalization K are shown in Fig. 3.

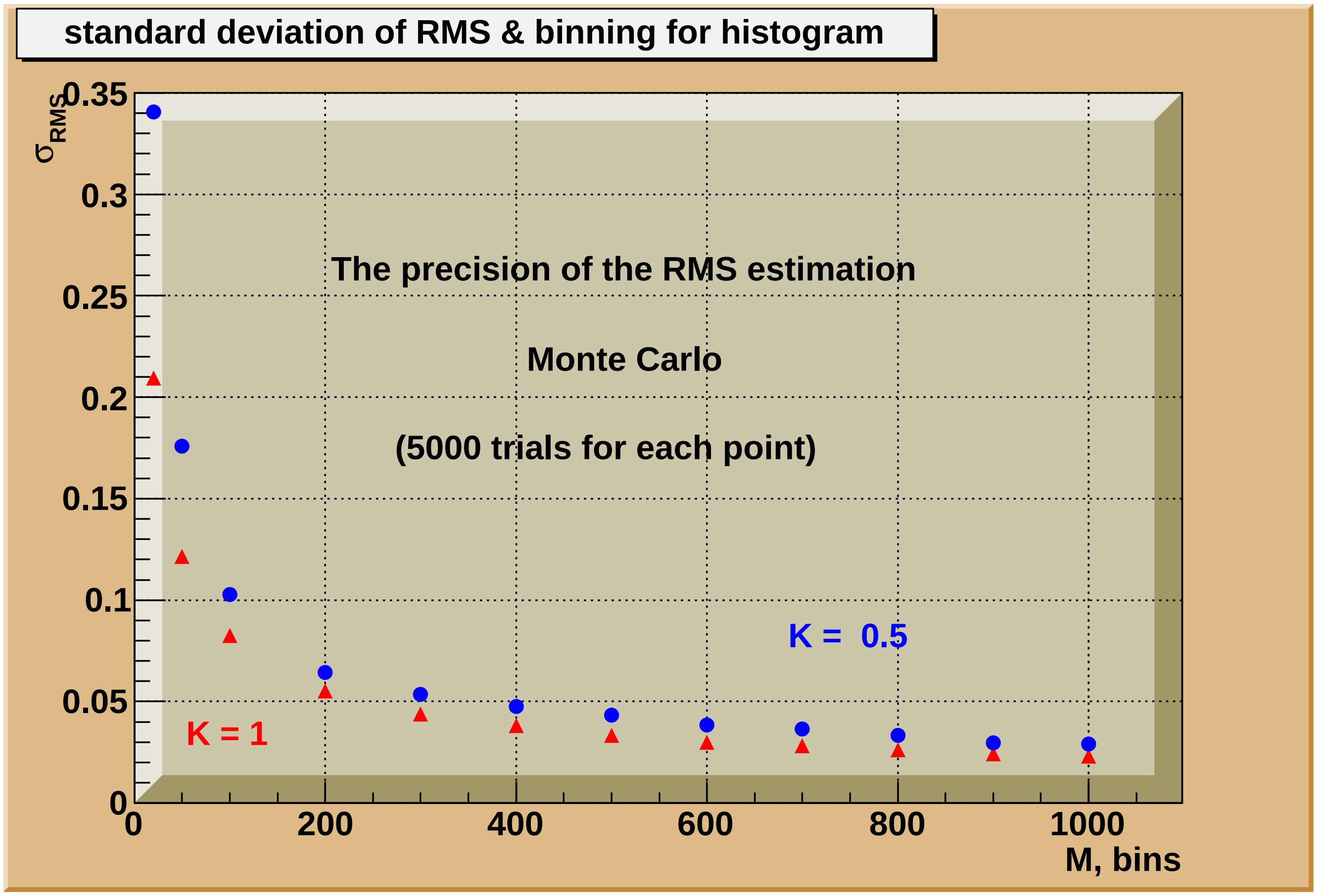


Figure 3: The dependence of the standard deviation of RMS on number of bins M. This dependence is shown for two values of normalized coefficient K ($K = 1$ and $K = 0.5$).

Script stat_analyzer.C

Two input files (*.root) with a set of TH1D histograms to compare.
User should indicate which histograms he/she wants to compare.

Processing:
– calculate mean value, RMS and p-value for each pair of histograms;
– sort variables by RMS in descending order;
– print sorted results
(variable – p-value – RMS – mean).

Output:
– to screen;
– to text file.

Commands for start-up of script in ROOT

```
root[0] .L stat_analyzer.C++
root[1] stat_analyzer(”signal_rv”,”signal”)
```

Missing E_T: large difference between histograms

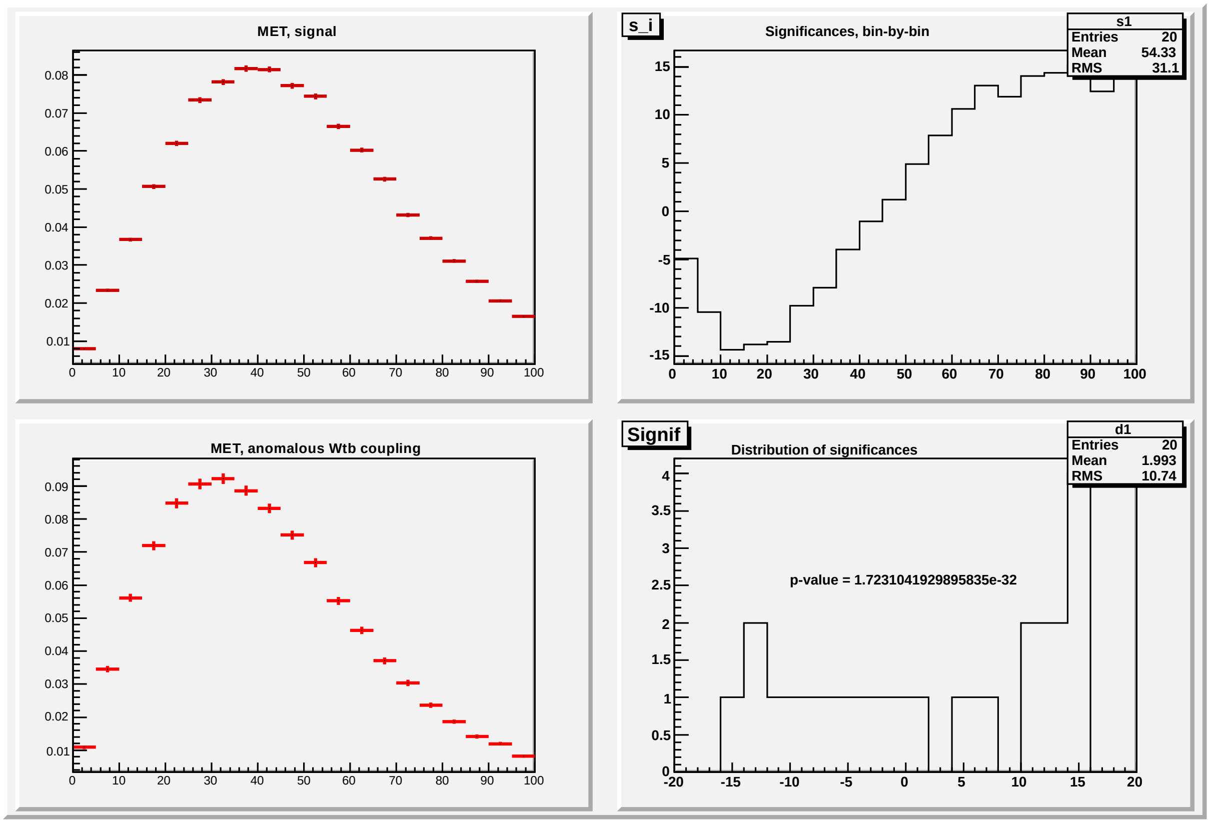


Figure 4: MET: the observed values in the first histogram (left,up), the observed values in the second histogram (left, down), observed normalized significances bin-by-bin (right, up), the distribution of observed normalized significances (right, down).

Let us consider the distributions of the probability (with errors) of the missing E_T to be in corresponding bin of histogram in the case of registration of Standard Model events and the events which are produced due to the presence of anomalous Wtb coupling. The comparison of these distributions is shown in Fig. 4 (RMS = 10.74, Mean = 1.99).

P_T light jet: only statistical difference

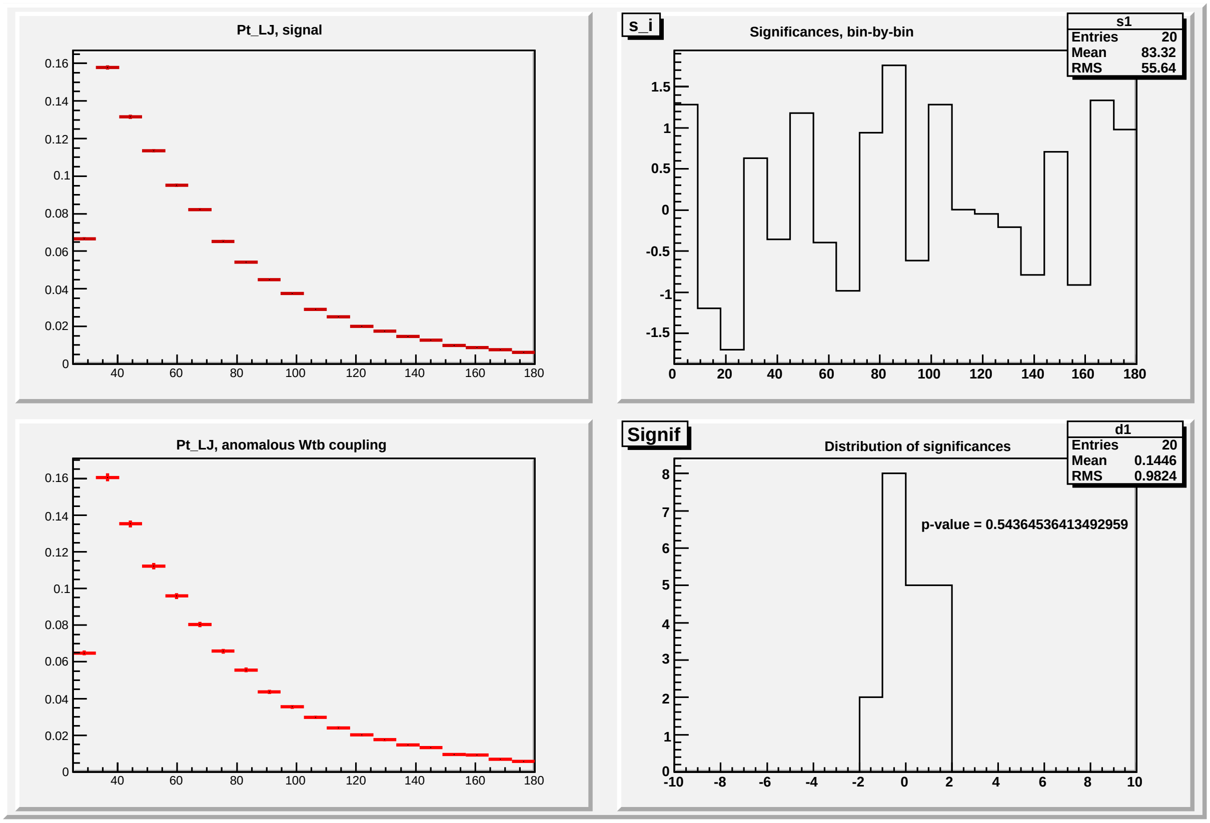
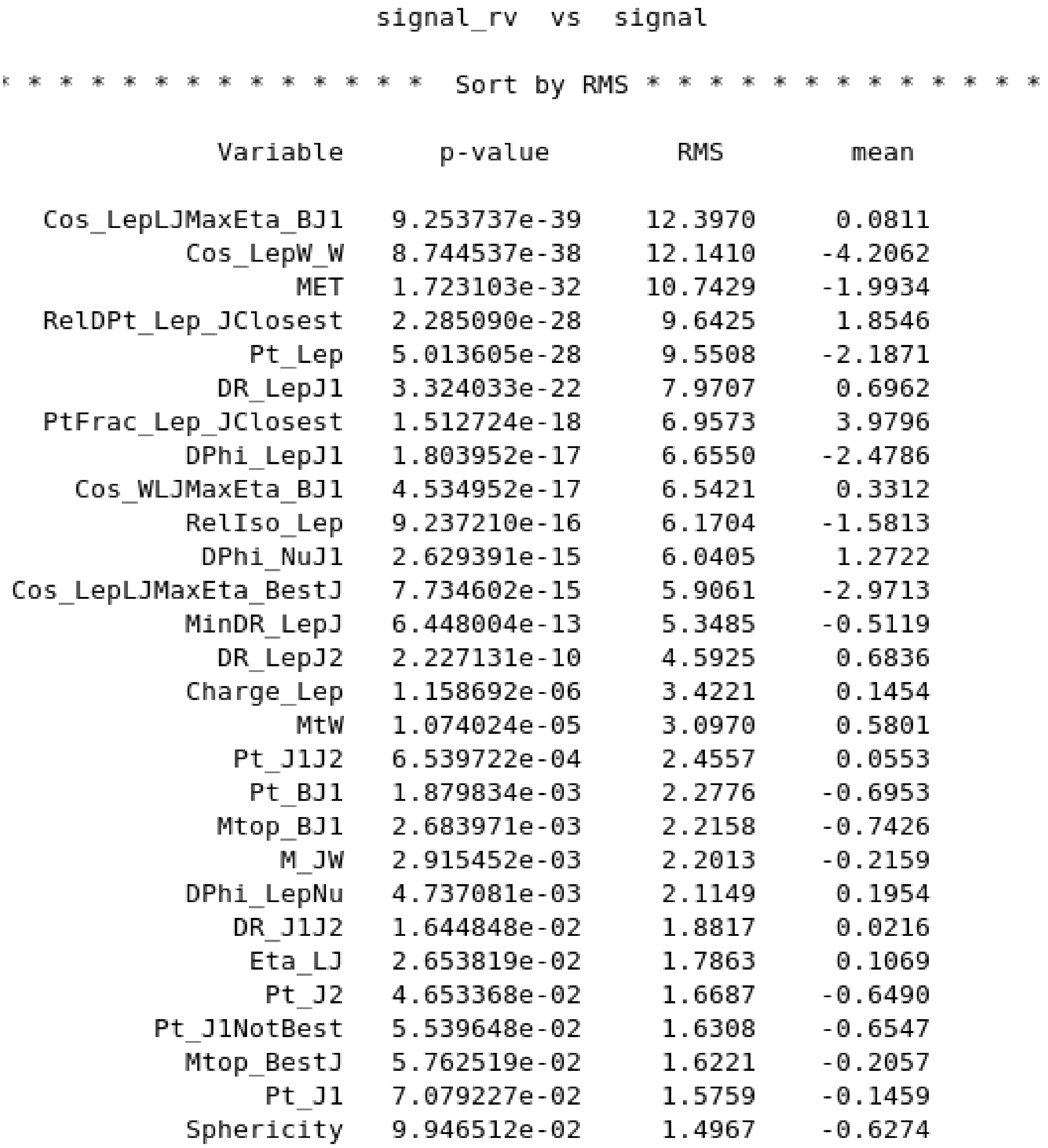


Figure 5: Pt_LJ: the observed values in the first histogram (left,up), the observed values in the second histogram (left, down), observed normalized significances bin-by-bin (right, up), the distribution of observed normalized significances (right, down).

The case of the absence of the difference between distributions for the production of single top quark in frame of SM and model with anomalous Wtb coupling is shown in Fig. 5 for P_t distribution of jet from light quark (RMS = 0.98, Mean = 0.14).

Output of the script



```
                         signal_rv   vs   signal

* * * * * * * * * * * * * *  Sort by RMS * * * * * * * * * * * * *

                 Variable       p-value          RMS          mean

      Cos_LepLJMaxEta_BJ1   9.253737e-39     12.3970        0.0811
               Cos_LepW_W   8.744537e-38     12.1410       -4.2062
                      MET   1.723103e-32     10.7429       -1.9934
      RelDPt_Lep_JClosest   2.285090e-28      9.6425        1.8546
                   Pt_Lep   5.013605e-28      9.5508       -2.1871
                 DR_LepJ1   3.324033e-22      7.9707        0.6962
      PtFrac_Lep_JClosest   1.512724e-18      6.9573        3.9796
               DPhi_LepJ1   1.803952e-17      6.6550       -2.4786
        Cos_WLJMaxEta_BJ1   4.534952e-17      6.5421        0.3312
               RelIso_Lep   9.237210e-16      6.1704       -1.5813
                DPhi_NuJ1   2.629391e-15      6.0405        1.2722
    Cos_LepLJMaxEta_BestJ   7.734602e-15      5.9061       -2.9713
               MinDR_LepJ   6.448004e-13      5.3485       -0.5119
                 DR_LepJ2   2.227131e-10      4.5925        0.6836
               Charge_Lep   1.158692e-06      3.4221        0.1454
                      MtW   1.074024e-05      3.0970        0.5801
                  Pt_J1J2   6.539722e-04      2.4557        0.0553
                   Pt_BJ1   1.879834e-03      2.2776       -0.6953
                 Mtop_BJ1   2.683971e-03      2.2158       -0.7426
                     M_JW   2.915452e-03      2.2013       -0.2159
               DPhi_LepNu   4.737081e-03      2.1149        0.1954
                  DR_J1J2   1.644848e-02      1.8817        0.0216
                   Eta_LJ   2.653819e-02      1.7863        0.1069
                    Pt_J2   4.653368e-02      1.6687       -0.6490
             Pt_J1NotBest   5.539648e-02      1.6308       -0.6547
               Mtop_BestJ   5.762519e-02      1.6221       -0.2057
                    Pt_J1   7.079227e-02      1.5759       -0.1459
               Sphericity   9.946512e-02      1.4967       -0.6274
```

Figure 6: Output of the script

Conclusions

- We discussed the possible tool for comparison of histograms in frame of the ROOT system (script stat_analyzer.C).
- This tool very easy in use and very easy to understand the results.
- This tool can be used in tasks of monitoring of the equipment. In this moment the method is used for choice of the most significant variables in multivariate analysis.
- This tool requires the additional development (now the method works for histograms TH1D).

η of lepton: the difference exits

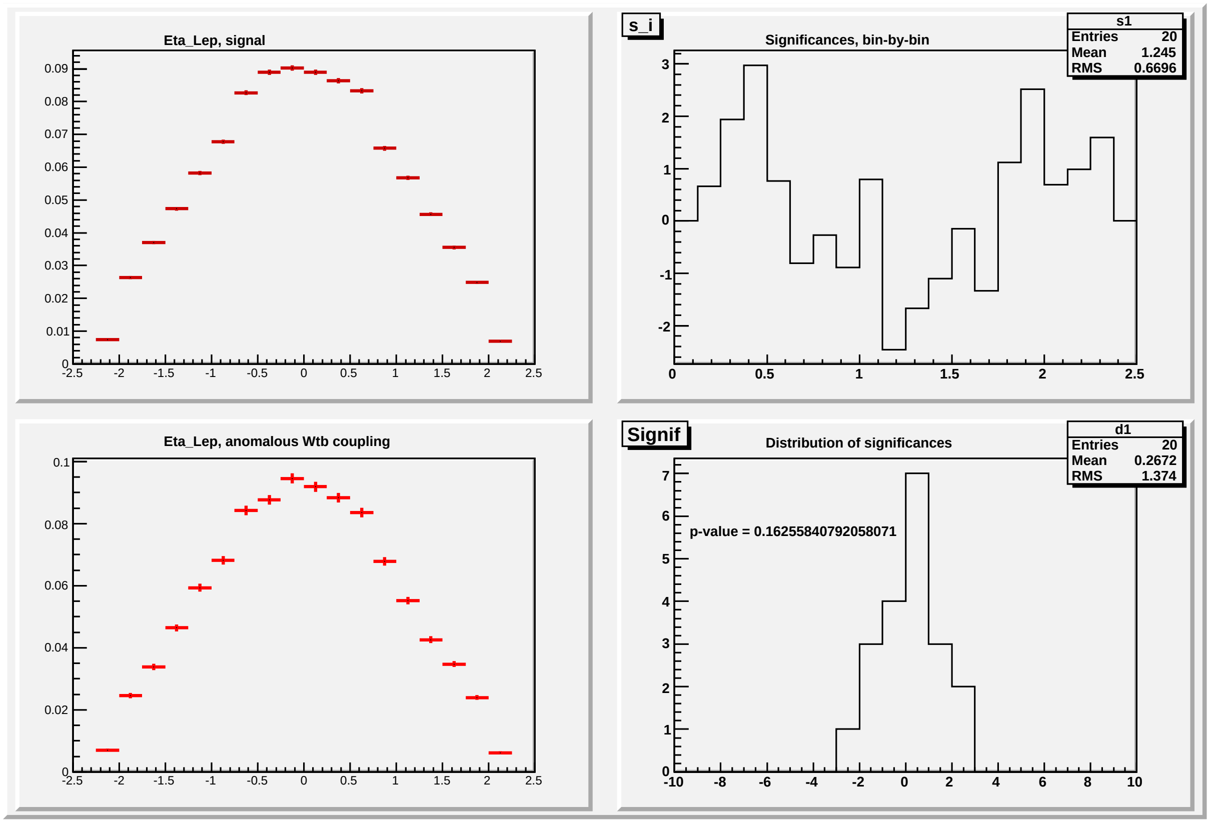


Figure 7: Eta_Lep: the observed values in the first histogram (left,up), the observed values in the second histogram (left, down), observed normalized significances bin-by-bin (right, up), the distribution of observed normalized significances (right, down).

The case of small difference between the histograms is shown in Fig. 7 (RMS = 1.37, Mean = 0.29).