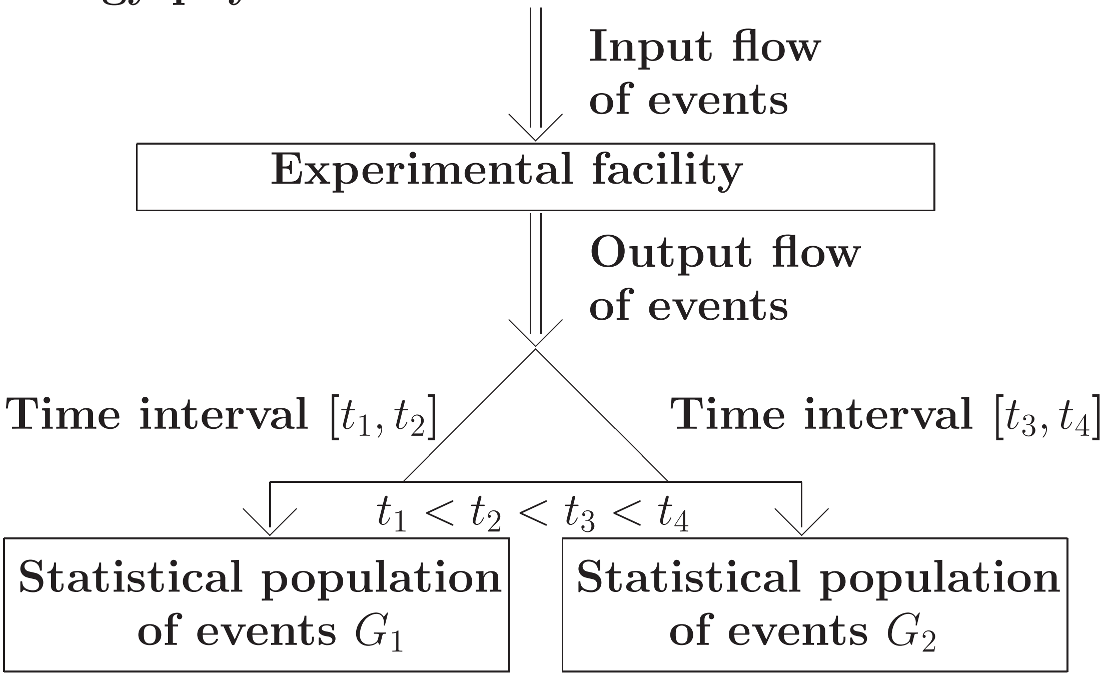
A method for statistical comparison of histograms

S. Bityukov, V. Smirnova (SRC IHEP),
N. Krasnikov (INR RAS, JINR),
A. Nikitenko (Imperial College, SRC ITEP)

- **Introduction**
- **Common scheme of the monitoring**
- **“Distance measures”**
- **Distribution of the test statistics**
- **Normalized significance**
- **Example**
- **Rehistogramming**
- **Distinguishability of histograms**
- **Case A: the same statistical population**
- **Case B: small difference between histograms**
- **Power of the test and probability of the correct decision**
- **Conclusion**

Introduction

The problem of the testing the hypothesis that two histograms are drawn from the same distribution is a very important problem in many scientific researches. For example, this problem exists for the monitoring of the experimental facility during experiments in high energy physics.



Input flow of events

Experimental facility

Output flow of events

Time interval $[t_1, t_2]$

Time interval $[t_3, t_4]$

$t_1 < t_2 < t_3 < t_4$

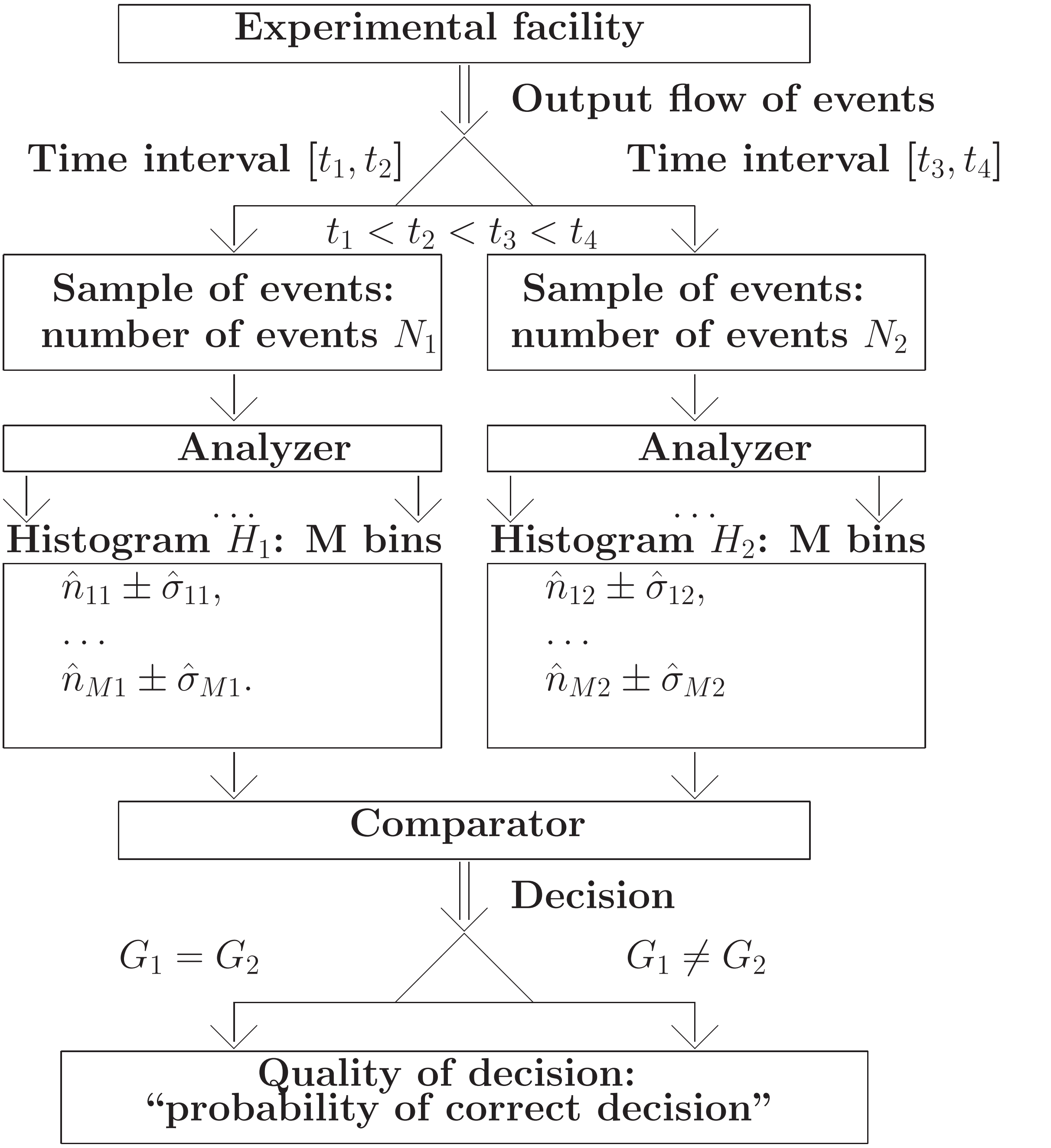
Statistical population of events G_1

Statistical population of events G_2

If facility is in norm during both time intervals then $G_1 = G_2$**.**

If facility is out of norm during one of time intervals then $G_1 \neq G_2$**.**

Common scheme of the monitoring



Experimental facility

Output flow of events

Time interval $[t_1, t_2]$

Time interval $[t_3, t_4]$

$t_1 < t_2 < t_3 < t_4$

Sample of events: number of events N_1

Sample of events: number of events N_2

Analyzer

Analyzer

. . .

. . .

Histogram H_1**: M bins**

$\hat{n}_{11} \pm \hat{\sigma}_{11}$,

. . .

$\hat{n}_{M1} \pm \hat{\sigma}_{M1}$.

Histogram H_2**: M bins**

$\hat{n}_{12} \pm \hat{\sigma}_{12}$,

. . .

$\hat{n}_{M2} \pm \hat{\sigma}_{M2}$

Comparator

Decision

$G_1 = G_2$

$G_1 \neq G_2$

Quality of decision: "probability of correct decision"

"Distance measures"

The most of methods for comparison of histograms use the "distance measure", for example,

- **the "χ^2 distance measure" between two observed histograms –**

$$\chi^2 = \sum_{i=1}^{M} \frac{\left(\frac{\hat{n}_{i1}}{N_1} - \frac{\hat{n}_{i2}}{N_2}\right)^2}{\frac{\hat{n}_{i1}}{N_1^2} + \frac{\hat{n}_{i2}}{N_2^2}} = \sum_{i=1}^{M} \hat{S}_i^2,$$

  $\hat{S}_i$ **in the case of the Poisson flows (G_1 and G_2) is a "normalized significance of deviation" for bin#i,**
  $N_1 = \sum_{i=1}^{M} \hat{n}_{i1}$ **- total number of events in histogram#1,**
  $N_2 = \sum_{i=1}^{M} \hat{n}_{i2}$ **- total number of events in histogram#2.**

- **"Bhattacharyya distance measure" –**

$$T_{\mathrm{BDM}} = \sqrt{\frac{\hat{n}_1}{N_1} \cdot \frac{\hat{n}_2}{N_2}} = \left(\sum_{i=1}^{M} \frac{\hat{n}_{i1}\hat{n}_{i2}}{N_1 N_2}\right)^{1/2}.$$

More examples can be found in *F. Porter, Testing Consistency of Two Histograms*, **arXiv:0804.0380.**

Distribution of the test statistics I

We propose to use several statistical moments of the distribution of test-statistics $\hat{S}_i,\ i = 1, M$ (normalized significances of deviation) as “a distance measure”.

If condition $G_1 = G_2$ is performed then each of test-statistics $(\hat{S}_i,\ i = 1, M)$ obeys the distribution which is close to standard normal distribution $\mathcal{N}(0, 1)$.

In this case the distribution of these test-statistics (S_i is calculated for each bin i of comparing histograms) also close to standard normal distribution.

In the report we consider the bidimensional “distance”

$$SRMS = (\bar{S}, RMS), \tag{1}$$

where $\bar{S} = \dfrac{\Sigma_{i=1}^{M} \hat{S}_i}{M}$ is mean value of the distribution of the $\hat{S}_i$ and $RMS = \sqrt{\dfrac{\Sigma_{i=1}^{M} (\hat{S}_i - \bar{S})^2}{M}}$ is the root mean square.

Distribution of the test statistics II

$SRMS$ **has a clear interpretation:**

- **if** $SRMS = (0,0)$ **then histograms are identical;**
- **if** $SRMS \approx (0,1)$ **then** $G_1 = G_2$ **(if** $\bar{S} \approx 0$ **and** $RMS < 1$ **then samples have overlapping);**
- **if previous conditions is not performed then** $G_1 \neq G_2$**.**

Note, the relation

$$RMS^2 = \frac{\chi^2}{M} - \bar{S}^2, \qquad (2)$$

where $\chi^2 = \sum\limits_{i=1}^{M} \hat{S}_i^2$**, exists for the distribution of significances.**

It shows that test-statistic χ^2 **is a combination (usually, non-optimal) of two test-statistics** RMS **and** $\bar{S}$**.**

Normalized significance

Let us consider a model with two histograms (H_1 and H_2) where the random variable in each bin obeys the normal distribution

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

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

Let we observed the histogram H_1 with N_1 events and the histogram H_2 with N_2 events.

We define the normalized significance as

$$\hat{S}_i = \frac{\hat{n}_{i1} - K\hat{n}_{i2}}{\sqrt{\hat{\sigma}^2_{i1} + K^2\hat{\sigma}^2_{i2}}}. \qquad (3)$$

Here $\hat{n}_{ik}$ is an observed value in the bin i of the histogram k, σ_{ik} is a standard deviation for $\hat{n}_{ik}$ and K is coefficient of normalization (K is defined by the task, for example, $K = \dfrac{N_1}{N_2}$ or $K = \dfrac{t_2 - t_1}{t_4 - t_3}$).

Example

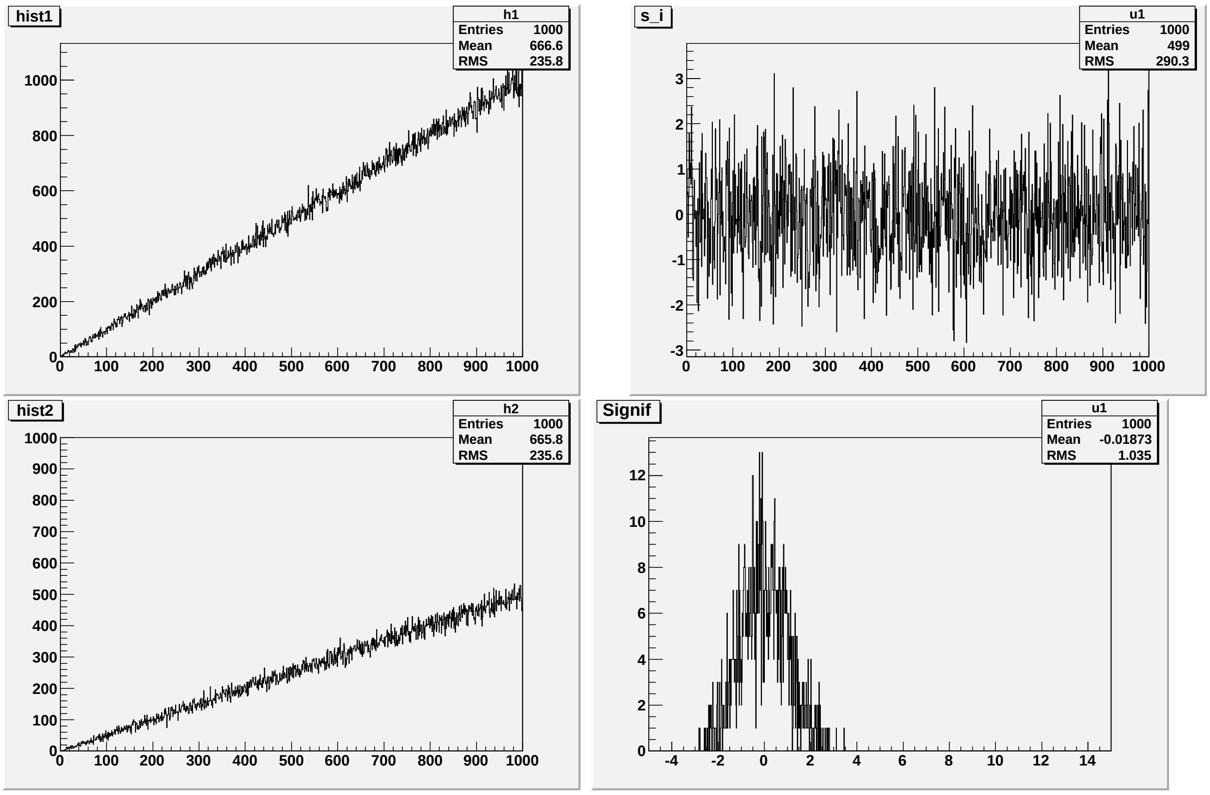


Figure 1: **Triangle distributions ($K = 2$, $M = 1000$): the observed values $\hat{n}_{i1}$ in the first histogram (left,up), the observed values n_{i2} in the second histogram (left, down), observed normalized significances $\hat{S}_i$ bin-by-bin (right, up), the distribution of observed normalized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

The example with histograms produced from the same events flow during unequal independent time ranges shows that the mean value and the standard deviation of the distribution of the $\hat{S}_i$ can be used as estimator of the statistical difference between histograms (the distribution of the S_i is close to $\mathcal{N}$(0,1)).

Rehistogramming

Two models of the statistical populations (pseudo populations) can be produced. Each of models represents one of the histograms.

In considered examples below 49999 clones for each of histograms are produced by the Monte Carlo simulation of content for each bin i of histogram k due to the law $\mathcal{N}(\hat{n}_{ik}, \hat{\sigma}_{ik}),\ i = 1, M,\ k = 1, 2$. As a result there are 50000 pairs of histograms for comparisons.

The comparison is performed for each pair of histograms (50000 comparisons in our examples). The distribution of the significances $\hat{S}_i$ is obtained as a result of each comparison. After that the moments of this distribution are calculated (in our case $\bar{S}$ and RMS). It is allow to estimate the error in determination of distribution moments.

This procedure can be named as "rehistogramming" in analogy with "resampling" in bootstrap method.

Distinguishability of histograms I

The estimation of the distinguishability of histograms is performed with the using of hypotheses testing.

"A probability of correct decision" $(1-\tilde{\kappa})$ about distinguishability of hypotheses is used as measure of the potential in separation of two histograms.

It is probability of the correct choice between two hypotheses "the histograms are produced by the treatment of events from the same event flow (the same statistical population)" or "the histograms are produced by the treatment of events from different event flows". This value can characterize the distinguishability of two histograms.

If $1-\tilde{\kappa}=1$ then the distinguishability of histograms is 100%, i.e. histograms are produced by the treatment of events from different event flows.

If $1-\tilde{\kappa}=0$ then it is impossible to separate these histograms, i.e. histograms are produced by the treatment of events from the same event flow.

Distinguishability of histograms II

The probability of correct decision $1 - \tilde{\kappa}$ is formed by the Type I error (α) and by the Type II error (β) in hypotheses testing.

α (Type I error) is the probability to accept the alternative hypothesis if the main hypothesis is correct.

β (Type II error) is the probability to accept the main hypothesis if the alternative hypothesis is correct.

If critical region (critical value, critical line, ...) is used correctly, i.e. if $\alpha + \beta \leq 1$, then

$$1 - \tilde{\kappa} = 1 - \frac{\alpha + \beta}{2 - (\alpha + \beta)} \tag{4}$$

(more details in *S.I. Bityukov, N.V. Krasnikov, Distinguishability of Hypotheses,* **Nucl.Inst.&Meth. A 534 152 (2004)).**

Case A: the same statistical population

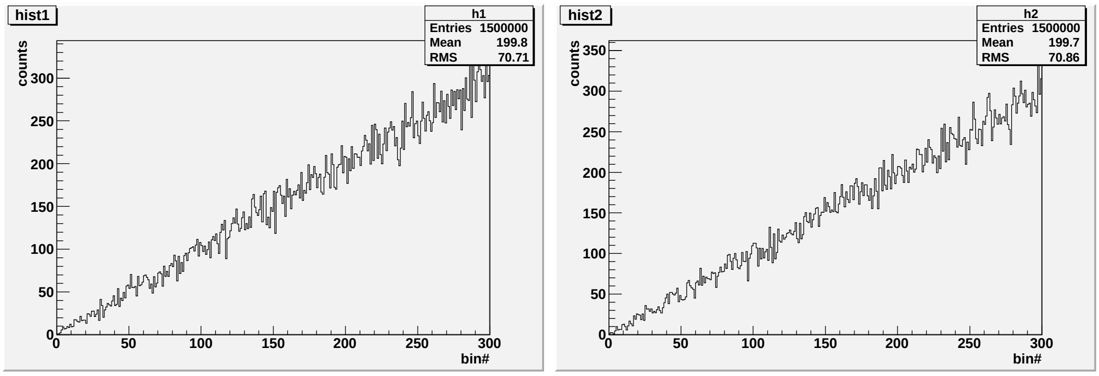


Figure 2: **Case A: input histograms – the same triangle distributions, M=300, K=1.**

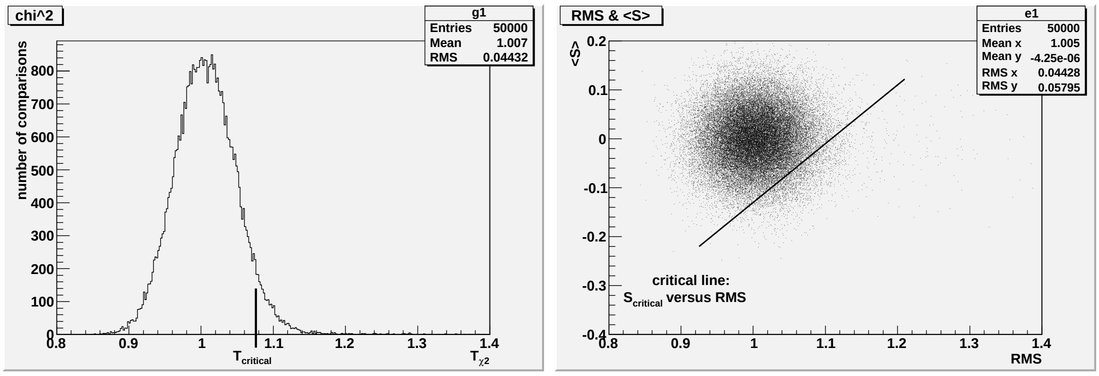


Figure 3: Case A: 50000 comparisons – $\sqrt{\frac{\chi^2}{M}}$ for each trial (left), RMS & $\bar{S}$ (right).

The Case A can be considered as a self-calibration of this method before applying it to the Case B.

Case B: small difference between histograms

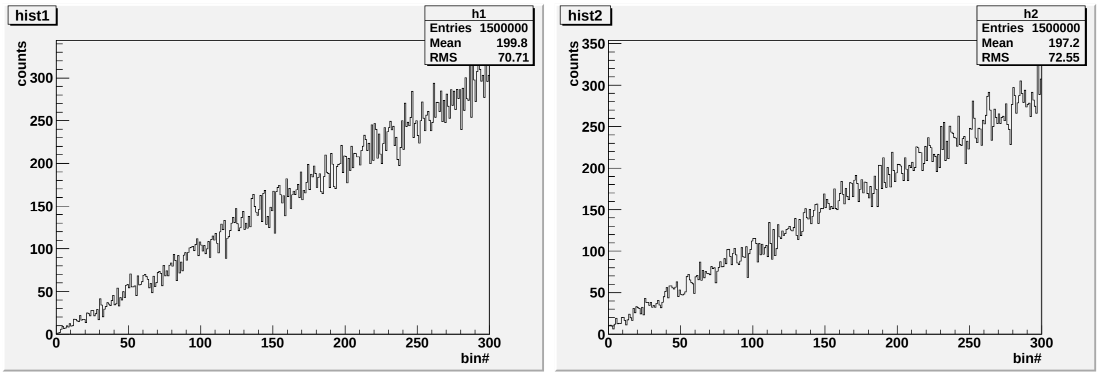


Figure 4: **Case B: input histograms – difference in slope of second histogram, M=300, K=1.**

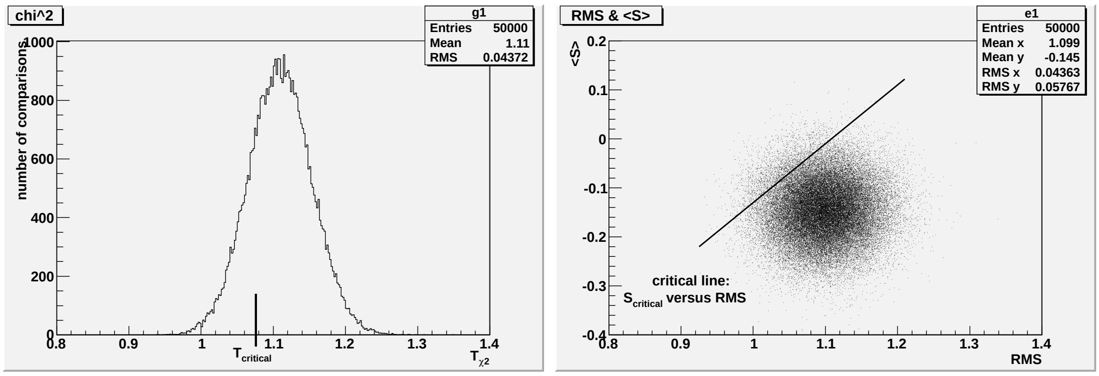


Figure 5: Case B: 50000 comparisons – $\sqrt{\frac{\chi^2}{M}}$ for each trial (left), RMS & $\bar{S}$ (right).

The power of the T_{χ^2} test = 0.7797.
The power of the $SRMS$ test = 0.9574.

Power of the test and probability of the correct decision

Accepted	In reality Case A	 Case B	Power of test	Probability of correct decision
Case A	**47499**	**11014**		
Case B	**2501**	**38986**		
	α	β	$1-\beta$	$1-\tilde{\kappa}$
	0.05	**0.2203**	**0.7797**	**0.8437**

Table 1: $\sqrt{\frac{\chi^2}{M}}$ **- 50000 decisions. Critical value** $\chi^2_{critical}$ **= 1.07576.**

Accepted	In reality Case A	 Case B	Power of test	Probability of correct decision
Case A	**47502**	**2132**		
Case B	**2498**	**47868**		
	α	β	$1-\beta$	$1-\tilde{\kappa}$
	0.05	**0.0426**	**0.9574**	**0.9515**

Table 2: $RMS\&\bar{S}$ **- 50000 decisions. The critical line is used for separation of two-dimensional distributions:** $S_{critical} = 1.2 * RMS_{critical} - 1.33$.

One can see that the method, which uses RMS and $\bar{S}$, gives better distinguishability of histograms than the χ^2 method.

Note, in this study are used only two moments of the distribution of significances.

Conclusions

- The proposed approach allows to perform the comparison of histograms in more detail than the methods which use one-dimensional test statistics.
- This method can be used in the task of the monitoring of equipment during experiments.
- The main items of the considerations are
  - the "normalized significance of deviation" provides us the distribution which is close to $\mathcal{N}(0,1)$ if $G_1 = G_2$;
  - the "rehistogramming" provides us the tool for an estimation of the accurace in the determination of statistical moments and, correspondingly, for testing the hypothesis about distinguishability of histograms;
  - the probability of correct decision gives us the estimator of the decision quality.