On one method of comparison experimental and theoretical data

S.I. Bityukov, NRC "Kurchtov Institute" – IHEP, Protvino & INR RAS, Moscow

N.V.Krasnikov, INR RAS, Moscow & JINR, Dubna

V.V. Smirnova, NRC "Kurchtov Institute" – IHEP, Protvino

Outline

1. Introduction
2. Statistical duality
   a) Confidence density
   b) Confidence distribution
3. Significance of difference
   a) Classification of significances
   b) Asymptotical normality of significances
4. Comparison of histograms
   a) Consistency or distinguishability of histograms
   b) Distance between histograms: bidimentional test statistics
5. Comparison of dependences
   a) Hypotheses testing
   b) Method of repeated dependence
6. Example: Statistical comparison of data sets
7. Conclusion: Advantages of this approach

Introduction

The method for statistical comparison of data sets (experimental and theoretical) is discussed.
The method now is in development.
The key parts of the method are presented in the report.

References on the method description:

[1] S. Bityukov, N. Krasnikov, A. Nikitenko, V. Smirnova, A method for statistical comparison of histograms, arXiv:1302.2651,2013

[2] S. Bityukov, N. Krasnikov, A. Nikitenko, V. Smirnova, On the distinguishability of histograms, Eur.Phys. J. Plus (2013) 128:143.

[3] S.I. Bityukov, N.V. Krasnikov, A.V. Maksimishkina, V.V. Smirnova, Multidimensional test statistics and statistical comparison of histograms, Int. journal of economics and statistics, 4 (2016) 98-101.

[4] S.I. Bityukov, N.V. Krasnikov, A.V. Maksimishkina, A.N. Nikitenko, V.V. Smirnova, A method for statistical comparison of data sets and its uses in analysis of nuclear physics data, Izvestiya vysshikh uchebnykh zavedenij. Yadernaya ehnergetika, 3 (2014) 44-51.

Statistical duality I

Definition: Let us a function $f(x,\lambda)$ can be expressed as a family of probabilities densities for variable x with given parameter λ, $p(x|\lambda)$, and as a family of densities for variable λ with given parameter x, $p(\lambda|x)$, so that $f(x,\lambda) = p(x|\lambda) = p(\lambda|x)$. These distributions have a property of *statistical duality* and they can be named as *statistically dual distributions.*

This properties take place for several statistically dual and self-dual distribitions:

Poisson versus *Gamma(1, x+1):* $f(x,\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$,

normal versus normal ($\sigma = const$): $f(x,\lambda) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\lambda)^2}{2\sigma^2}}$,

Cauchy versus Cauchy (b = const): $f(x,\lambda) = \frac{b}{\pi(b^2+(x-\lambda)^2)}$,

Laplace versus Laplace (b = const): $f(x,\lambda) = \frac{1}{2b} e^{-\frac{|x-\lambda|}{b}}$, $\ldots$.

Statistical duality II

The notion is introduced in note (S. Bityukov, V. Smirnova, V. Tapereckina, arXiv: math/0411462, 2004).

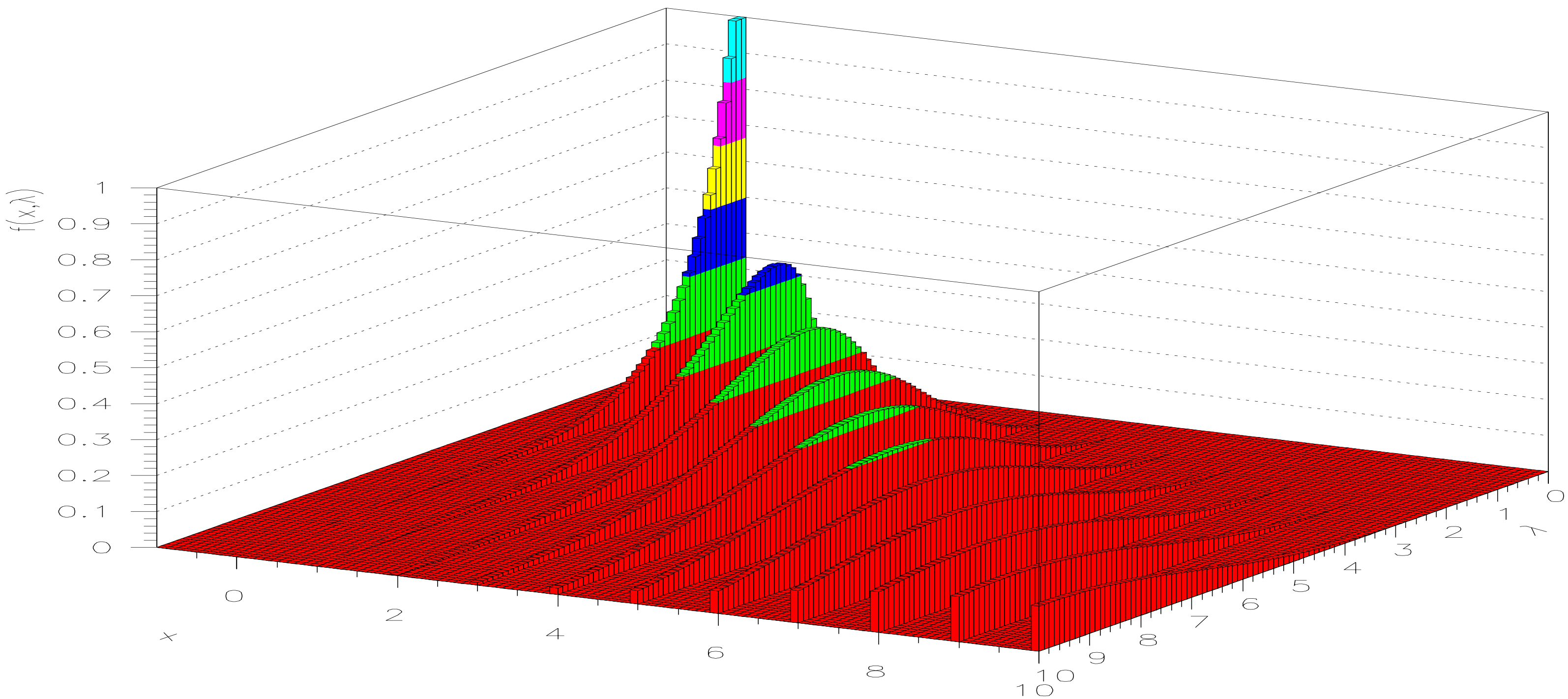
It is used in paper (M. Blackwell et al., Sociological methods & research, 46 (3) (2017) 342-369) for construction of a unified approach to measurement error and missing data.
Also, this notion is used for analysis of household sizes in paper (V.E. Jennings, C.W. Lloid-Smith, Mathematical Scientist, 40 (2015) 103-117). Authors of paper changed their notation *Poisson-Gamma frame* in their previous papers to our *statistical duality*.

We used the statistical duality to prove the uniqueness of corresponding confidence intervals for parameter via construction of *confidence density* of parameter.

Confidence densities I

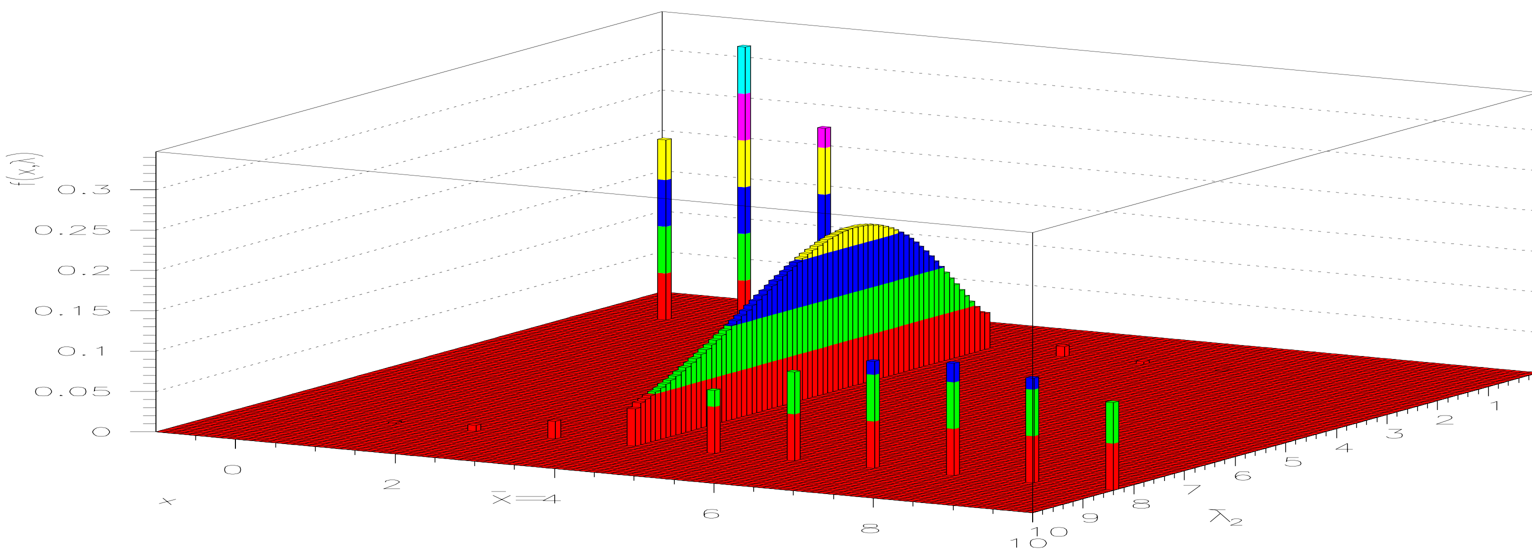
Let us construct the bidimensional function (here *x* is integer) for the case of

Poisson versus *Gamma(1, x+1):* $f(x,\lambda) = \dfrac{\lambda^x e^{-\lambda}}{x!}$.



• Confidence densities II

Let us x is a random variable and λ is a parameter. Then fix value $x = 4$ (for example, the number of observed events equal 4). The function $f(4, \lambda)$ (i.e. *Gamma(1,5)*) is a *confidence density of parameter* λ. Chose the upper and lower limits along λ axis we can construct any confidence interval for parameter λ if $x = 4$ (see, figure below) which contents the parameter with given confidence (here 90%).

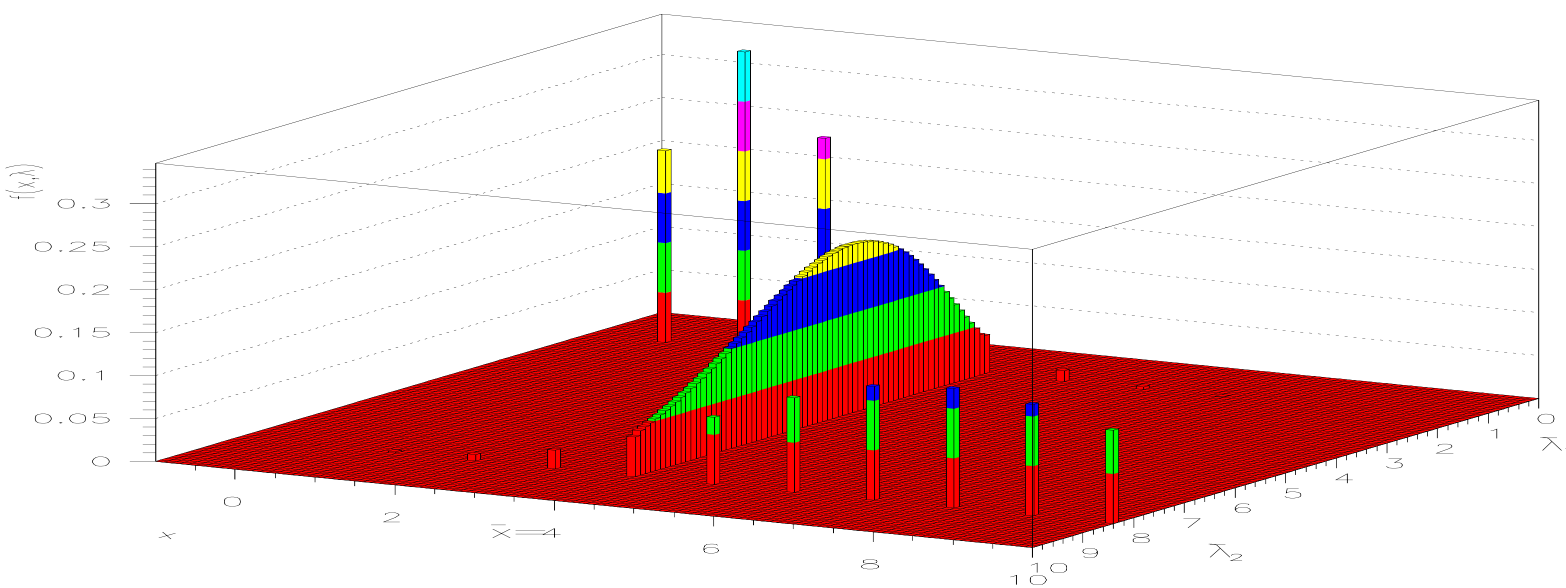


Confidence density III

The identity

$$\sum_{k=x+1}^{\infty} f(k,\lambda_{lower}) + \int_{\lambda_{lower}}^{\lambda_{up}} f(x,\lambda)d\lambda + \sum_{k=0}^{x} f(k,\lambda_{up}) = 1$$

does not leave a place for any prior except uniform in construction of confidence intervals (S. Bityukov, V. Medvedev, V. Smirnova, Yu. Zernii, arXiv:physics/0403069, 2004). It means that *Gamma(1,x+1)* is the confidence density of parameter λ if we observed x events in Poisson flow of events.



Confidence distribution

The uniqueness of confidence densities is true for other statistically dual distributions (S. Bityukov, N. Krasnikov, S. Nadarajah, V. Smirnova, AIP Conf.Proc 1305 (2011) 346-353). This allows to construct and use *confidence distribution* of parameter λ under estimation via the measurement of the random variable x.

More details can be found in reviews (M. Xie & K. Singh, International Statistical Review, 81 (2013) 3) and (S. Nadarajah, S. Bityukov, N. Krasnikov, Statistical Methodology, 22 (2015) 23-46).

Note, *statistical duality* is duality between c*onfidence* and *probability*. In this sence the equation in previous slide can be considered as *law of conservation*.

Significance of difference I

The concept of ``*the significance*" of a signal in presence of background in experiment (Y. Zhu, arXiv:physics/0507145 [physics.data-an]) (or, more precisely, "*the significance of the difference*" between the number of signal events and zero) is widely used in data analysis in high-energy physics.

Let a sample (or samples) of realizations of some random variable be obtained from an infinite population within a given time. Each realization is called as an event. Number of realizations, which *determine by some of conditions* (for example, *cuts*), can be either a background events, or a signal events, which are indistinguishable.

Significance of difference II

Several methods exist to quantify the statistical ``*significance*'' of an signal (expected or estimated) in this sample. Following the conventions in high energy physics, the term significance usually means the ``*number of standard deviations*'' of an expected or observed signal is above expected or estimated background. It is understood implicitly that ``*significance*'' should follow a Gaussian distributions with a standard deviation which equals one.

In the simplest case, the concept ``*significance*'' can be described with the help of two numbers: ***b*** - the number of background events and ***s*** - the number of signal events (signal and background events are indistinguishable) that appeared during the given time.

Significance of difference III

The distributions of the observed number of background events $\hat{\boldsymbol{b}}$ and the observed number of signal events $\hat{\boldsymbol{s}}$ usually obey Poisson distributions with parameters $\boldsymbol{b}$ (expected number of background events) and $\boldsymbol{s}$ (expected number of signal events), respectively. Note, the realization of random variable allows to estimate the parameter of Poisson distribution. It means that we must compare the estimated parameters of Poisson flows of events when we compare two samples.

For example, to assess the uncertainties that arise after (or before) the measurements, the significance of $\boldsymbol{S_1} = \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{b}}}$ or $\boldsymbol{S_2} = \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{s}+\boldsymbol{b}}}$ was often used. With a small number of events, significances $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ give incorrect results.

Let us $\boldsymbol{S}$ characterizes the significance of signal. The choice of significance to be used depends on the study. There are three types of significances $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ (S. Bityukov, N. Krasnikov, A. Nikitenko, V. Smirnova, Proc. of Sci. (ACAT2008) 118, 2008).

Classification of significances I

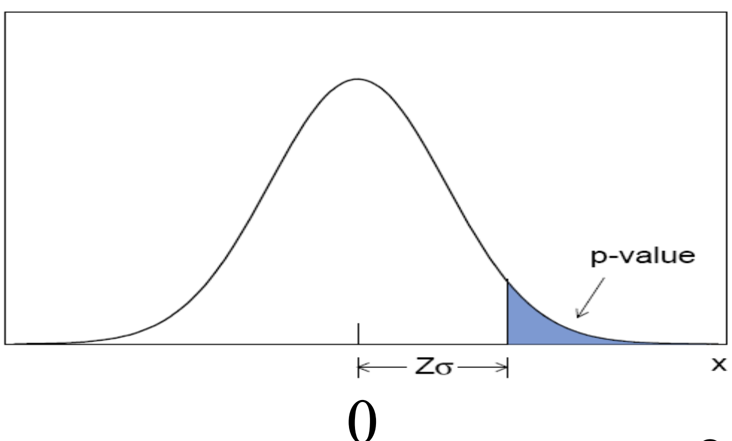
(A) Expected significance: if ***s*** and ***b*** are expected values then we take into account both statistical fluctuations of signal and of background. Before observation we can calculate only an expected significance ***S*** which is a parameter of experiment. ***S*** characterizes the quality of experiment:

$$\boldsymbol{S_{c12}} = 2\,(\sqrt{\boldsymbol{s}+\boldsymbol{b}} - \sqrt{\boldsymbol{b}}\,)$$ as an example.

Classification of significances II

(B) Observed significance: if $\widehat{\boldsymbol{s}+\boldsymbol{b}}$ is observed value and $\boldsymbol{b}$ is expected value then we take into account only the fluctuations of background. In this case we can calculate an observed significance $\widehat{\boldsymbol{S}}$ which is an estimator of expected significance of experiment $\boldsymbol{S}$. $\widehat{\boldsymbol{S}}$ characterizes the quality of experimental data.

For example, $\hat{Z}$ (or $\widehat{\boldsymbol{S}}_{\boldsymbol{cP}}$). This significance corresponds a probability to observe number of events equal or greater than $\widehat{\boldsymbol{s}+\boldsymbol{b}}$ in sample with Poisson distribution with mean $\boldsymbol{b}$ which converted to equivalent number of sigmas of a Gaussian distribution, i.e. $1 - \Phi(\hat{Z}) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\hat{Z}} e^{-\frac{t^2}{2}}dt$.



Significance $\hat{Z} = 5$ corresponds p-value = $2.87 \cdot 10^{-7}$.

Classification of significances III

(*C*) If $\widehat{s+b}$ and $\hat{b}$ are observed values of signal+background and background with known errors of measurement then we can use the standard theory of errors to estimate the significance of signal S_d. In case of normal distribution of errors the formula for S_d looks as

$$S_d = \frac{\widehat{s+b} - \hat{b}}{\sqrt{\sigma_{s+b}^2 + \sigma_b^2}} \ ,$$

where σ_{s+b}^2 and σ_b^2 are corresponding variances of error distributions.

If samples for estimation of $\widehat{s+b}$ and $\hat{b}$ have different volumes (different integrated luminosities of experiments) then formula for significance looks as

$$S_d = \frac{\widehat{s+b} - K\hat{b}}{\sqrt{\sigma_{s+b}^2 + K^2\sigma_b^2}} \ ,$$

where K is a ratio of integrated luminosities of experiments.

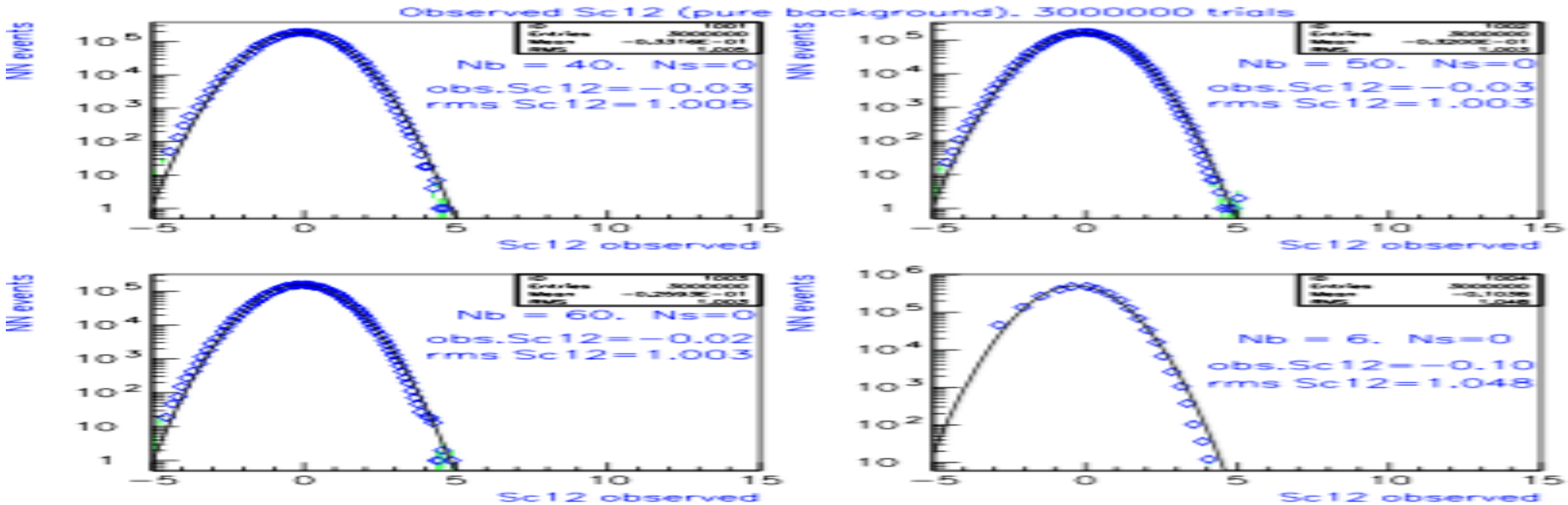
Asymptotical normality of significances I

An important property of these significances is property that when comparing two independent samples obtained from the same general population, the distribution of estimates of ``the significance of the difference'', obtained for these samples, is close to the standard normal distribution ***N(0,1)*** (see Figures in next slide).

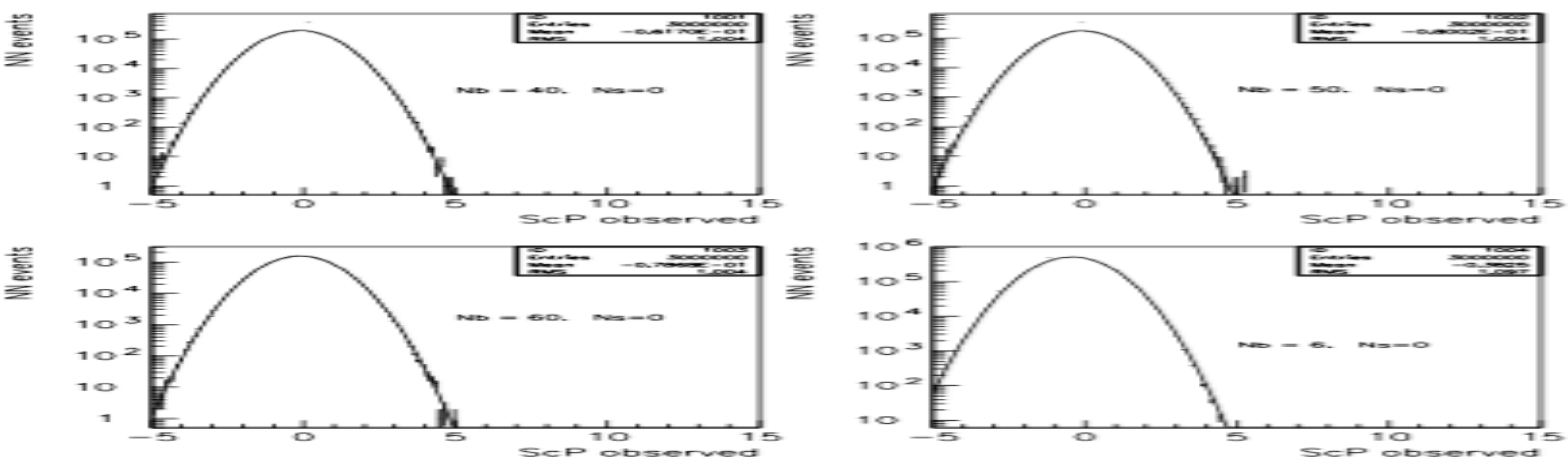
It is shown for several significances in paper (S. Bityukov, N. Krasnikov, A. Nikitenko, V. Smirnova, Proc. of Sci. (ACAT2008) 118, 2008) (significances $\boldsymbol{S_{c12}}$ and $\hat{Z}$) and in paper (S. Bityukov, N. Krasnikov, A. Nikitenko, V. Smirnova, Eur.Phys.J.Plus (2013) 128:143) (significance $\boldsymbol{S_d}$) by Monte Carlo experiments. Fisz (M. Fisz, The limiting distribution of a function of two independent random variables and its statistical applications, Colloquium Mathematicum, 3 (1955) 199-202) shows that the significance $\boldsymbol{S_d}$ in case of Poisson distribution is asymptotically normal ***N(0,1)***.

Asymptotical normality of significances II

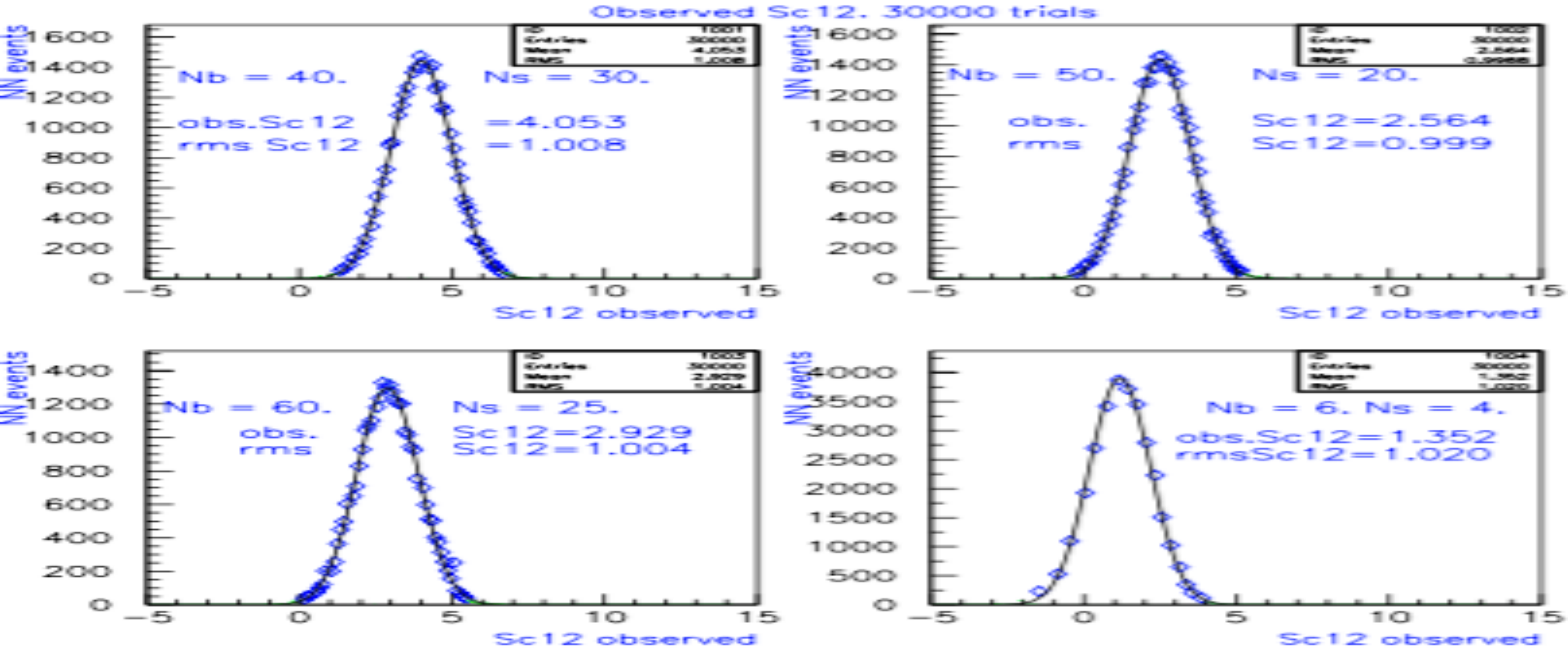
Sc12



ScP



Asymptotical normality of significances III
Sc12



Note, the variance of such type significances close to 1, i.e. the variance is a constant and we have the property of *statistical duality* for significances. It means that *observed significance* (realization of random variable(s)) is unbiased estimation of *expected significance* (parameter of experiment) with variance of confidence density close to *1* (Poisson flows of events).

Comparison of histograms I

This property of significances allows to *unificate* the comparison of corresponding bins of histograms or corresponding points of dependences.

The famous slogan "God made man, but Samuel Colt made them equal" can be rephrased for these significances "Experimenter made measured points, and only significance of difference make them equal".

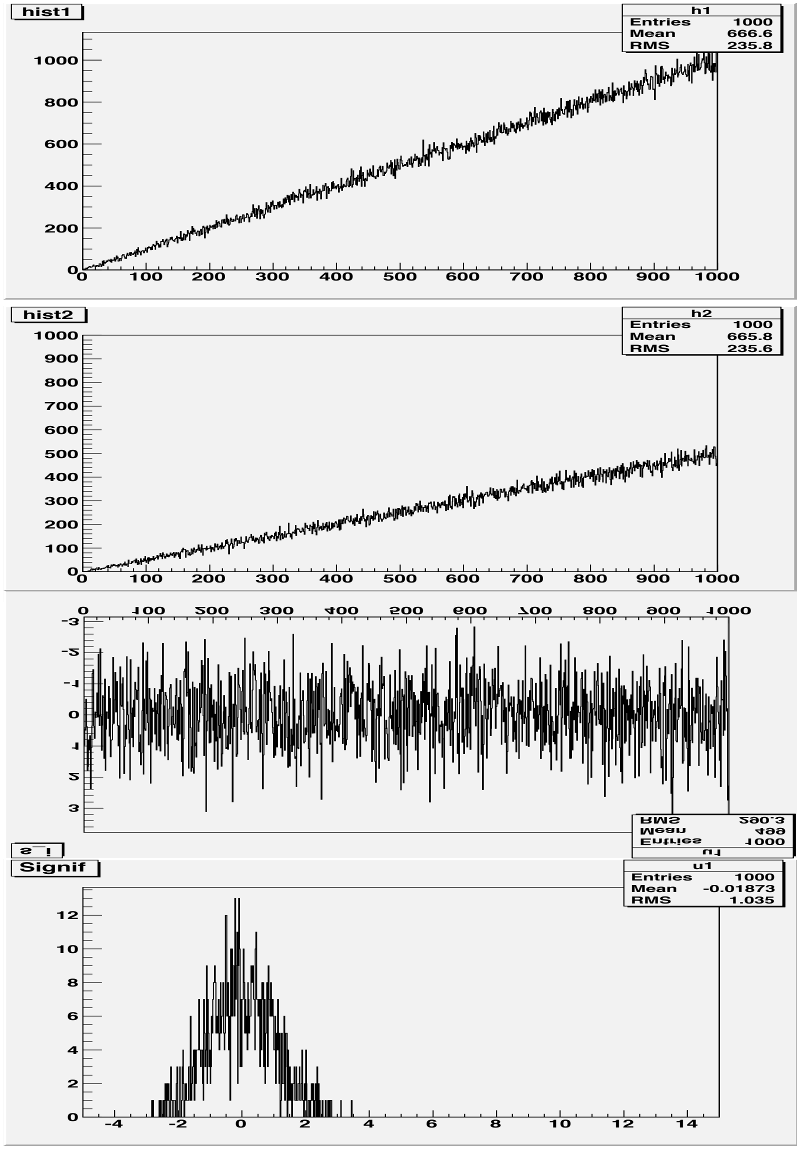
Statistical duality allows to mix frequentist probabilities and confidence densities. This means that we can use the measurement of corresponding random variables as estimators of parameter and confidence density of this parameter.

For example, we plan to compare *reference histogram* or *reference dependence* and *test histogram* or *test dependence*. During comparison we can consider values in *reference histogram* or in *reference dependence* as observed random values. Correspondingly, values in *test histogram* or in *test dependence* we can consider as parameters. And vice versa.

Comparison of histograms II

Suppose, there is given a set of non-overlapping intervals. A histogram represents the frequency distribution of data which populates those intervals. This distribution is obtained during data processing of the sample (which is taken from the Poisson flow of events) with observed values of random variable). These intervals usually are called as bins. Consider the example of histogram comparison by the use formula (see, slide 15)

$\hat{S}_i(K) = \frac{\widehat{n_{i1}} - K\widehat{n_{i2}}}{\sqrt{\sigma^2_{\hat{n}_{i1}} + K^2\sigma^2_{\hat{n}_{i2}}}}$, where i is a number of bin, $K = 2$ is ratio of histogram volumes, $\sigma^2_{\hat{n}_{i1}}, \sigma^2_{\hat{n}_{i2}}$ are variances of number of events in bin#i. Here $\sigma^2_{\hat{n}_{i1}} = \widehat{n_{i1}}$, $\sigma^2_{\hat{n}_{i2}} = \widehat{n_{i1}}$.



Consistency or distinguishability of histograms

Often a goal of histograms comparison is a testing of their *consistency*. *Consistency* here is the statement that both histograms are produced during data processing of independent samples which are taken from the same flow of events (or from the same population of events).

In our paper (EPJ+) is proposed approach which allows to estimate the *distinguishability* of histograms and, correspondingly, the distinguishability of parent events flows (or parent samples). We use the distribution of some test statistics (significances of difference) instead of single test statistics in other methods. This distribution has statistical moments (for example, the mean, RMS, asymmetry, excess, ...), i.e. the distribution can be considered as *multidimensional test statistics* with, for example, statistical moments as coordinates.

Distance between histograms: bidimentional test statistics

We can calculate statistical moments for distribution of $\widehat{S}_i$, and in principle, we have information about distinguasability of samples under testing. In analyses we use only two moments: the mean value of significances distribution $\overline{S}$ and the *rms* of this distribution, i.e. bidimensional test statistics $SRMS = (\bar{S}, rms)$:

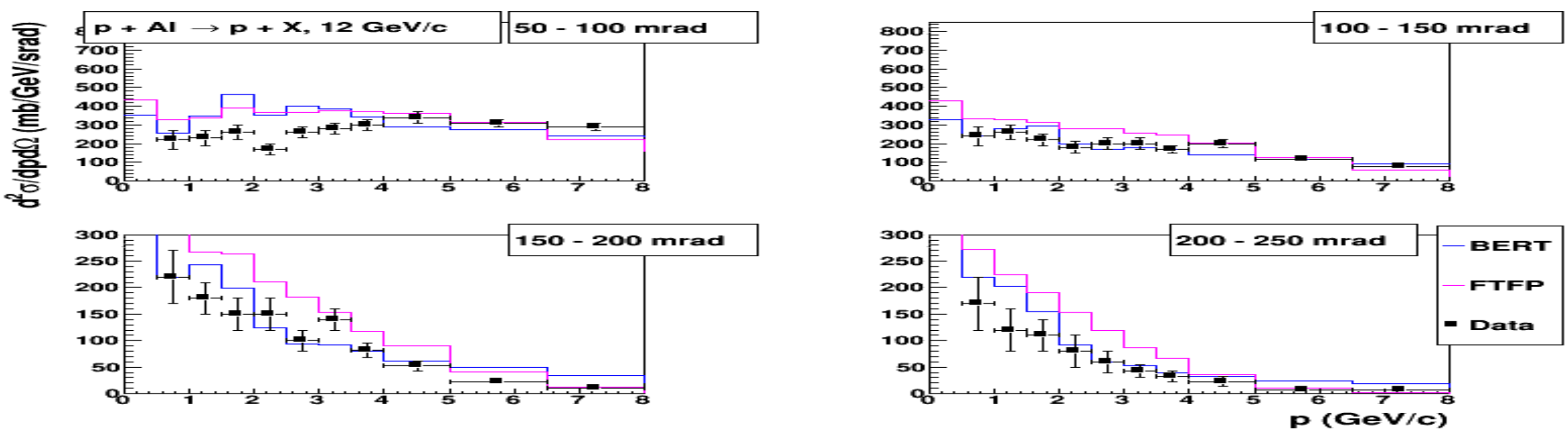
a) if $SRMS = (0, 0)$, then histograms are identical,
b) if $SRMS \approx (0, 1)$, then samples are taken from the same flow of events,
c) if previous conditions are not valid, the flows have difference.

Note, in the case (b) the realisation $\widehat{S}_i$ of random value "significance of difference" for bin # *I* obeys the standard normal distriburion *N(0,1)*.

Comparison of dependences I

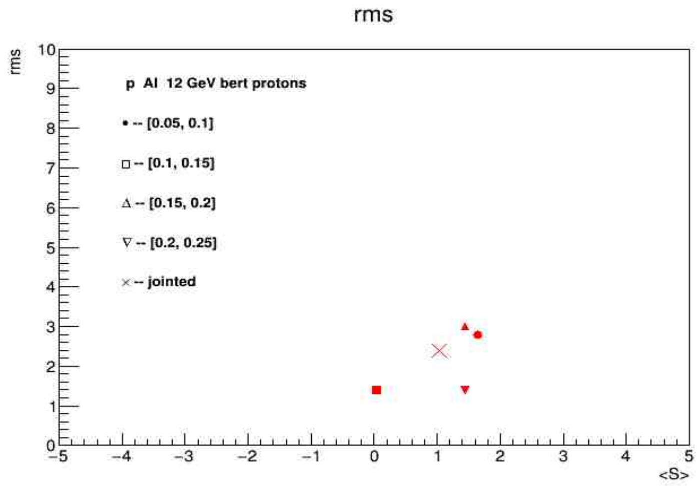
We can applied this approach for comparison of dependences. The another formula for $\widehat{S}_i = \frac{\widehat{n_{i1}} - \widehat{n_{i2}}}{\sqrt{\sigma^2_{\hat{n}_{i1}} + \sigma^2_{\hat{n}_{i2}}}}$ is used in this case.

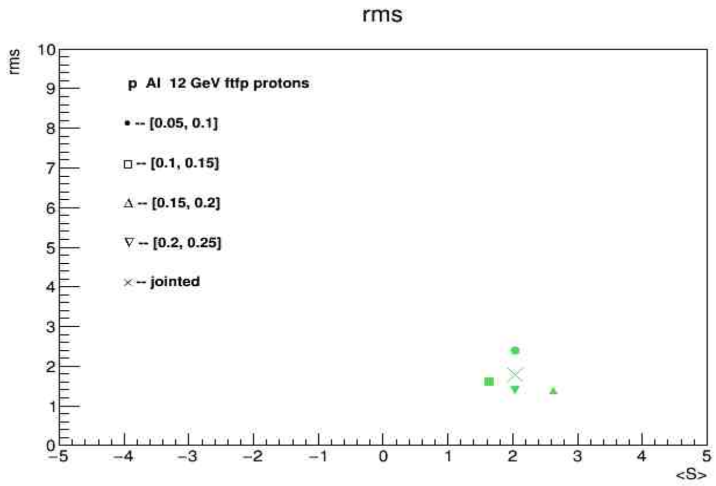
Due to statistical duality we can use test statistics *SRMS* = ($\bar{S}$,*rms*) both as measured random variable and as a parameter which describes this pair of dependences under comparing.

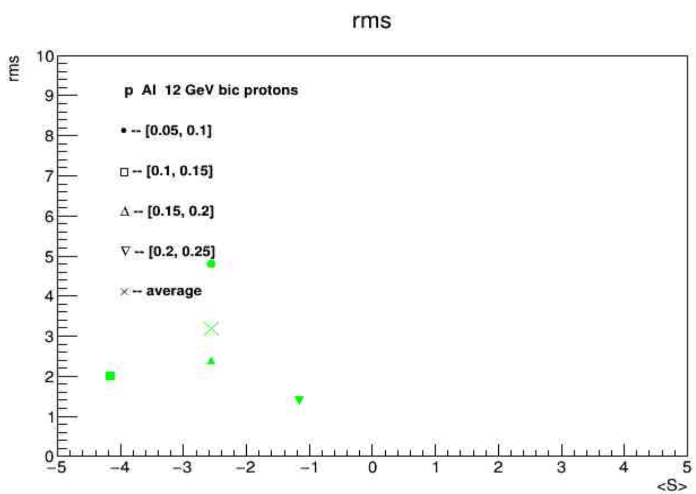


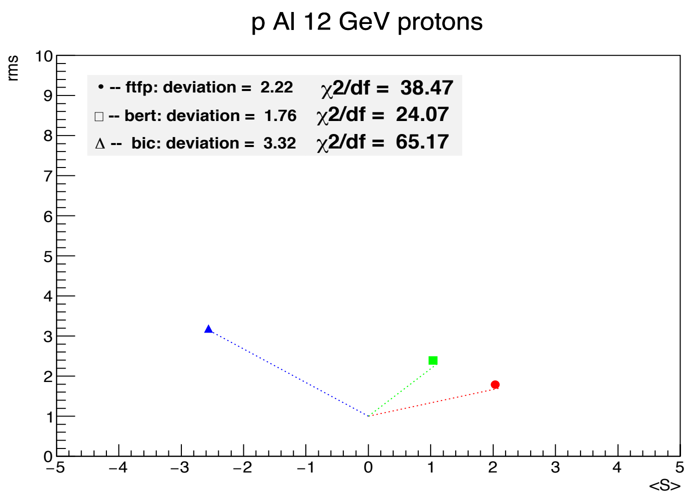
FTFP – FRITIOF Precompound, BERT – Bertini cascade, BIC – binary cascade models. V. Ivantchenko (preliminary)

Comparison of dependences II









Hypotheses testing I

The using of this approach for comparison of dependences has many problems, which can be avoid with the help of the *method of repeated dependence*.

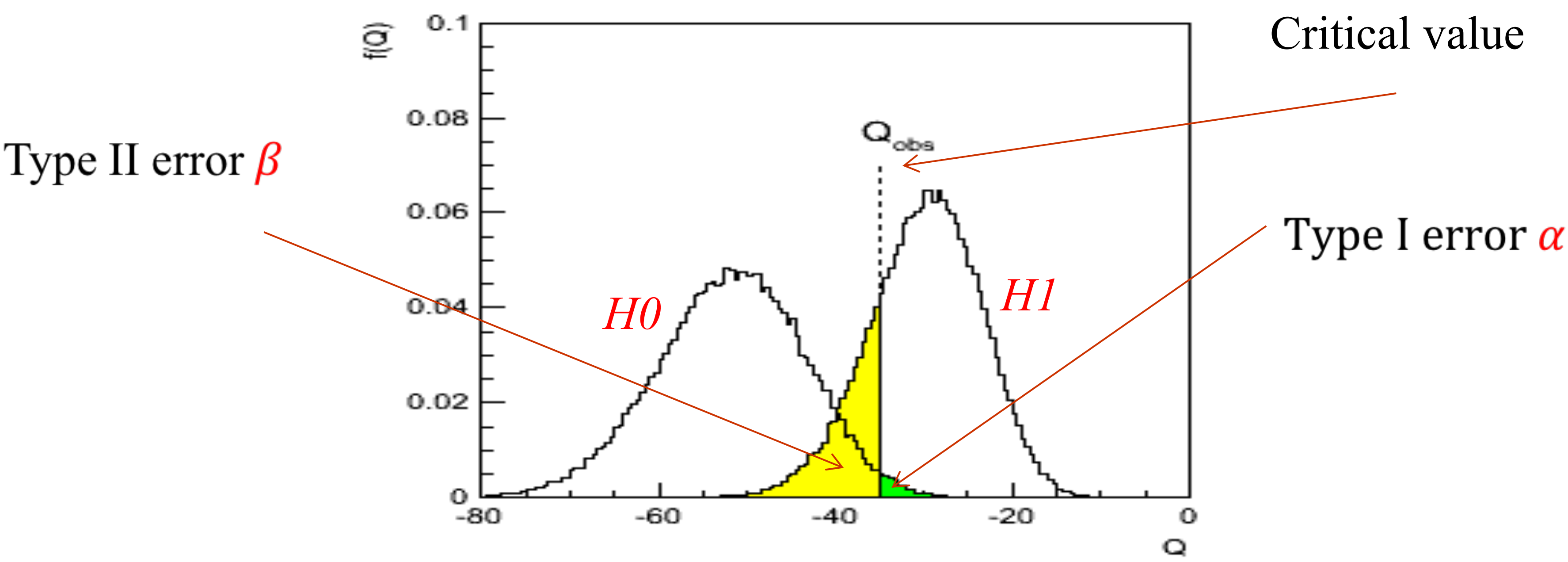
If the goal of the comparison of dependences is the check of their consistency, then task is reduced to hypotheses testing: *main hypothesis* H0 (dependences are produced during data processing of samples taken from the same flow of events) against *alternative hypothesis* H1 (dependences are produced during data processing of samples taken from different flows of events).

The determination of *critical area* allows to estimate *Type I error* (α) and *Type II error* (β) in decision about choice between H0 and H1.

The *Type I error* is a probability of mistake if done choice is H1, but H0 is true.

The *Type II error* is a probability of mistake if done choice is H0, but H1 is true.

Hypotheses testing II



Distribution *H0* is a confidence density of expected value of test statistics (which is used for hypotheses testing) if hypothesis H0 is true, distribution *H1* is a confidence density of expected value of test statistics if hypothesis H1 is true.

Hypotheses testing III

The selection of a *significance level* (α) allows to estimate the *power of the test* (1-β). Usually, values of significance level are 10%, 5%, 1%.

If both hypotheses are equivalent, then other combinations of the α and β are used.

For example, in task about distinguishability of the flows of events works a *relative uncertainty* $\dfrac{\alpha+\beta}{2-(\alpha+\beta)}$ for $\alpha+\beta \le 1$ (S. Bityukov, N. Krasnikov, Proc. on Confidence limits, CERN, Geneva, 2000)

Under the *test of equal tails* the mean error $\dfrac{\alpha+\beta}{2}$ can be used.

Method of repeated dependence I

The hypotheses testing require the knowledge of the distribution of test statistics. The distribution of test statistics can be constructed by Monte Carlo.

If *errors of values in measured points* of at least one of dependence (for example, reference) *are known*, than one can construct the set of similar dependences (*clones*) according with errors, which imitates the population of dependences which produced due to data processing of the samples taken from the same flow of events. This set of dependences is used for construction of the distribution of *reference statistics* for the case of H0 hypothesis (due to comparison of the reference dependence and the produced clones of the reference dependence).

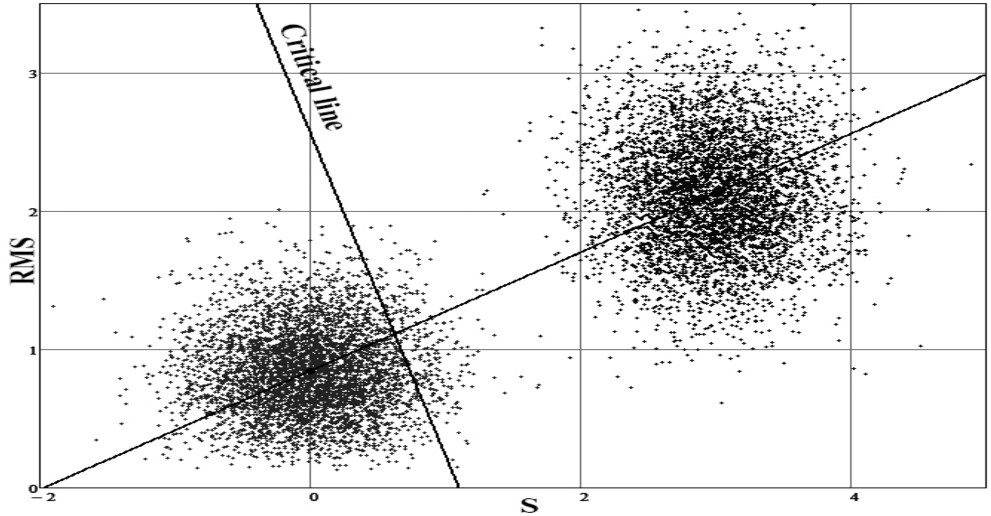
This procedure can be named as ”*method of repeated dependence*" in analogy with ”*method of repeated sample*" in *bootstrap* technique.

Method of repeated dependence II

Further the set of dependences of such type is constructed for test dependence (second dependence under comparing). New set is used for construction of the distribution of *test statistics* for the case of H1 hypothesis. This is comparison of the reference dependence with the produced clones of test dependence (i.e. with clones of second dependence).

The comparison of the distribution of *reference statistics* for the case of H0 hypothesis (imitation population of SRMS, which produced by the comparing of reference dependence and its clones) and the distribution of *test statistics* for the case of H1 hypothesis (imitation population of SRMS, which produced by the comparing of reference dependence and clones of test dependence) allows to estimate the *uncertainty* in hypotheses testing. Note, there are many combinations for comparison depending on task (reference clones with test clones and so on).

Method of repeated dependence III



Critical line with significance level $\alpha = 5\%$.

Spots here can be considered as confidence density of expected parameter SRMS = (S,rms), which corresponds observed random variable $SRMS = (\bar{S}, rms)$.

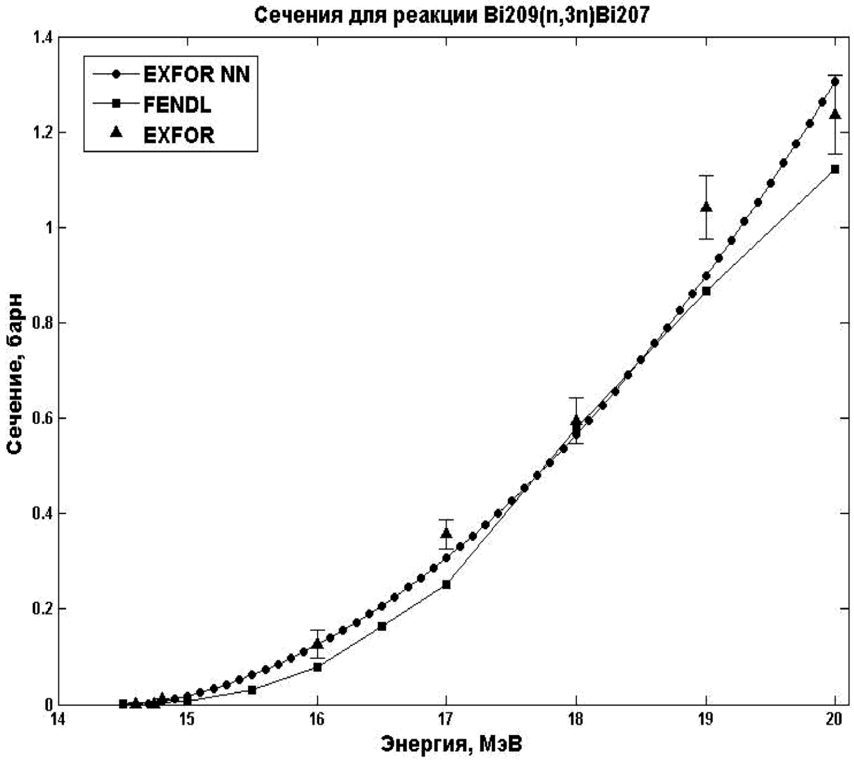
Fraction of points from lower spot, which are over *critical line*, is α.

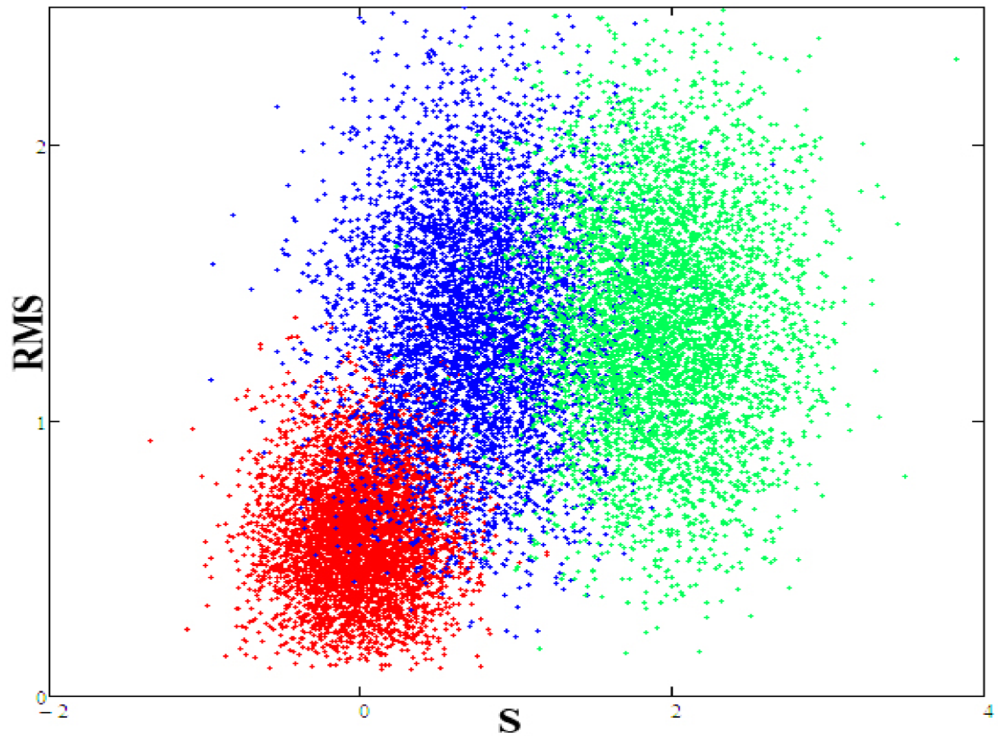
Fraction of points from upper spot, which are under *critical line*, is β.

Then, *relative uncertainty of decision* is $\hat{\varkappa} = \dfrac{\alpha+\beta}{2-(\alpha+\beta)}$.

It is a *measure of distance* between reference and test dependences for given *critical line*.

Example: Statistical comparison of data sets





The comparison of the experimental data (triangles with errors and red spot) with results obtained in frame of two models (blue and green spots) [A. Maksimushkina, V. Smirnova, Method for data statistical visualization, Scientific Visualization, 7, Issue 5 (2015) 26-37].

Conclusion: advantages of this approach

1. We have a *measure* of the "distance" between dependences. It is *relative uncertainty* $\hat{\varkappa}$ of the decision about *consistence of dependences*.

2. We can compare multidimensional dependences likewise as unidimensional dependences.

3. We can compare two sets of several dependences simultaneously likewise as we compare a pair of dependences.

4. We can use any unidimensional test statistics (Kolmogorov-Smirnov, Anderson-Darling, …) as additional dimension in proposed multidimensional (here bidimentional) test statistics.

5. In principle, we can compare sets of heterogenious data.