Two approaches to Combining Significance

S.Bityukov, N.Krasnikov, A.Nikitenko

Introduction

“Suppose one experiment sees a 3-sigma effect and another experiment sees a 4-sigma effect. What is the combined significance? Since the question is ill-posed, the statistics literature contains many papers on the topic … ” (Cousins, 2007).

Methodology for combining findings from repeated studies did in fact begin with the idea of combining independent tests back in the 1930’s (Tippett, 1931; Fisher, 1932; Pearson, 1933). There are many approaches to this subject. Many of them is discussed in cited review of R. Cousins.

We consider the using of one (Stouffer et al., 1949) of these methods for combining of significances. We show the applicability of this method in the case of Poisson flows of events under study. We also discuss the approach based on confidence distributions. This approach shows an applicability of Stouffer’s method (inverse normal method) for combining of significances under certain conditions.

Combination of tests

All the methods of combining tests depend on what is known as a ***P***-value. A key point is that the observed ***P***-values derived from continuous test statistics follow a uniform distribution under the null hypothesis H_0 regardless of the form of the test statistic, the underlying testing problem, and the nature of the parent population from which samples are drawn.

Quite generally, suppose $X_1, \ldots, X_n$ is a random sample from a certain population indexed by the parameter θ, and $T(X_1, \ldots, X_n)$ is a test statistic for testing $H_0{:}\theta=\theta_0$ against $H_1{:}\theta>\theta_0$, where θ_0 is a null value, and suppose also that H_0 is rejected for large values of $T(X_1, \ldots, X_n)$.

There is no general recommendation for the choice of the combination method. All the combination methods are optimal for some testing situations. As an example we consider the method (Stouffer's method) from the class of probability transformation methods.

Inverse normal method (Stouffer et al., 1949)

It is based on fact that the ***z*** value based on the ***P*** value, defined as

$z = \Phi^{-1}(P)$ is a standard normal variable under the null

hypothesis H_0, where $\Phi(.)$ is the standard normal **c**umulative **d**istribution **f**unction (**cdf**). Thus, when, the ***P*** values $P_1, \ldots, P_L$ are converted to the ***z*** values $z_1, \ldots, z_L$, we have **i**ndependent and **i**dentically **d**istribited (**iid**) standard normal variables under H_0. The combined significance test is essentially based on the sum of these ***z*** values, which has a normal distribution under the null hypothesis with mean ***0*** and variance ***L***.

The test statistic $Z = \sum_{i=1}^{L} z(P_i) / \sqrt{L}$ is thus a standard normal variable under H_0, and hence can be compared with the critical values in the standard normal table.

What do we mean by significance? (I)

``Common practice is to express the significance of an enhancement by quoting the number of standard deviations" (Frodesen, et al., 1979)

Let us define a significance ***Z*** (or, often, ***S*** in HEP) (Cousins, 2007):

$$Z = \Phi^{-1}(1-p) = -\Phi^{-1}(p)$$

where

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} \exp(-t^2/2)dt = \frac{1+erf(Z/\sqrt{2})}{2}$$

so that

$$Z = \sqrt{2}erf^{-1}(1-2p)$$

For example, ***Z***=5 corresponds to ***p***=2.87*10E-7. On can see the relation between some uncertainty ***p*** and the corresponding number of standard deviations ***Z*** in the frame of standard normal distributions.

Internal and observed significances

Z characterizes the significance of the deviation of one value from another value (usually, signal ***s*** + background ***b*** from background ***b***). The choice of significance to be use depends on the study:

A) 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 internal (or initial) significance Z_p which is **a parameter of experiment**. Z_p characterizes the quality of experiment.

B) If ***s+b*** is observed value and ***b*** is expected value then we take into account only the fluctuations of background. In this case we can calculate an observed significance Z_e which is **an estimator of internal** significance of experiment Z_p. Z_e characterizes the quality of experimental data.

C) If ***s*** and ***b*** are observed values with known errors of measurement then we can use the standard theory of errors.

Zoo of significances

Many types of significances are used. For example, the significances Z_{Bi} (Binomial)=Z_{Γ} (Gamma), Z_N (Bayes Gaussian), Z_{PL} (Profile Likelihood) were studied in details in paper (Cousins et al., 2008).

As shown in (Bityukov et al., 2006) several types of significances can be considered as normal random variables with variance close to 1. For example, significances S_{c12} and Z_N (or S_{cP}) satisfy this property.

S_{c12} (Bityukov et al., 1998) corresponds to the case of hypotheses testing of two simple hypotheses $H_0:\theta=b$ against $H_1:\theta=s+b$.

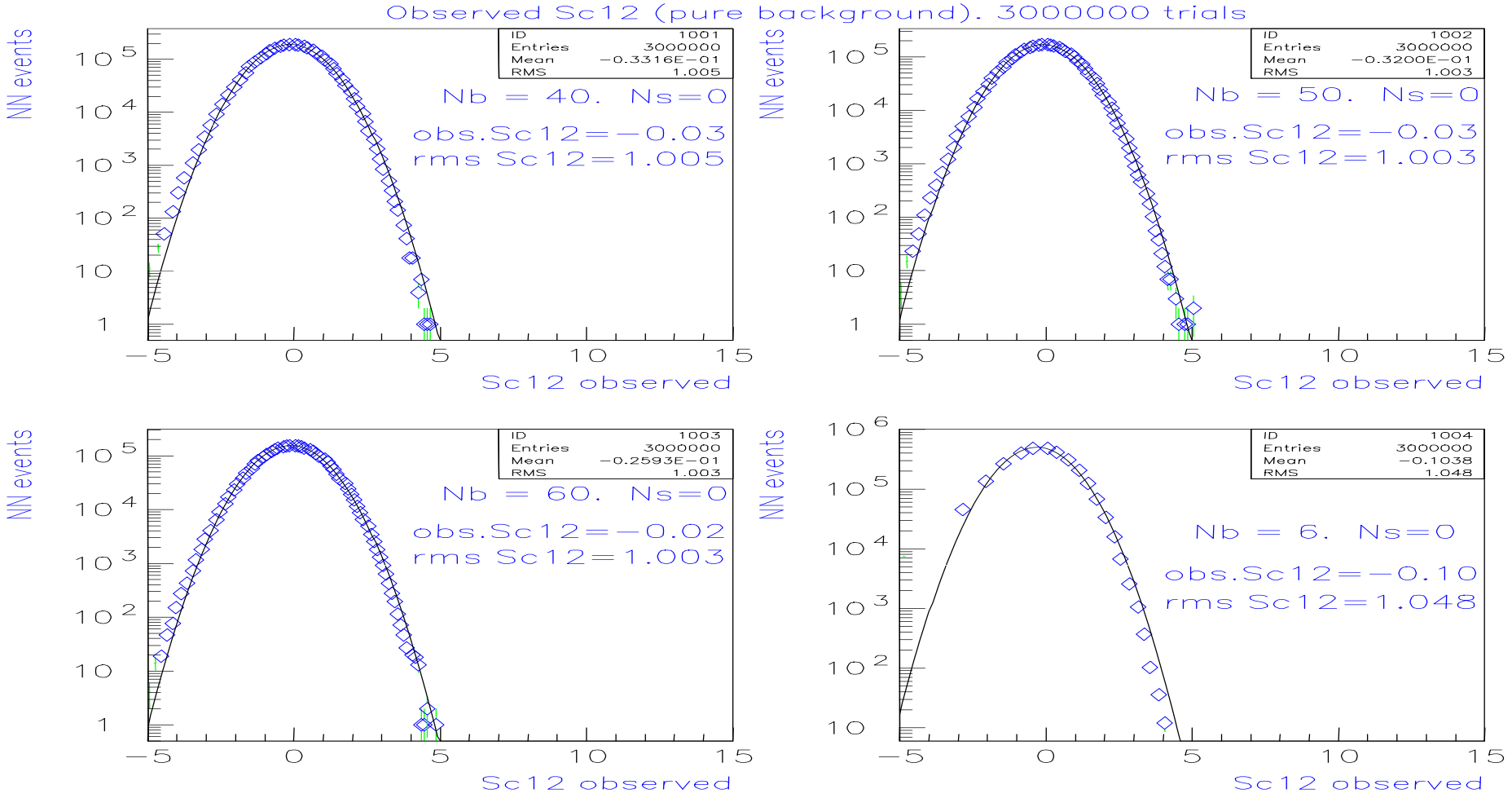
$S_{c12}=2(\sqrt{(s+b)}-\sqrt{(b)})$.

Z_N (Narsky, 2000) is the probability from Poisson distribution with mean b to observe equal or greater than $s+b$ events, converted to equivalent number of sigmas of a Gaussian distribution. It is the case of hypotheses testing with $H_0:\theta=b$ against $H_1:\theta>b$.

Let us show the applicability of the Stouffer's method to significances of such type. We present here only the results for S_{c12}. Results for Z_N (S_{cP}) are analogous.

What do we mean by significance ? (II)

Distributions of observed S_{c12} in the case of signal absence are presented for 3000000 simulated experiments for each value of b (b=40, 50, 60, 6, correspondingly)

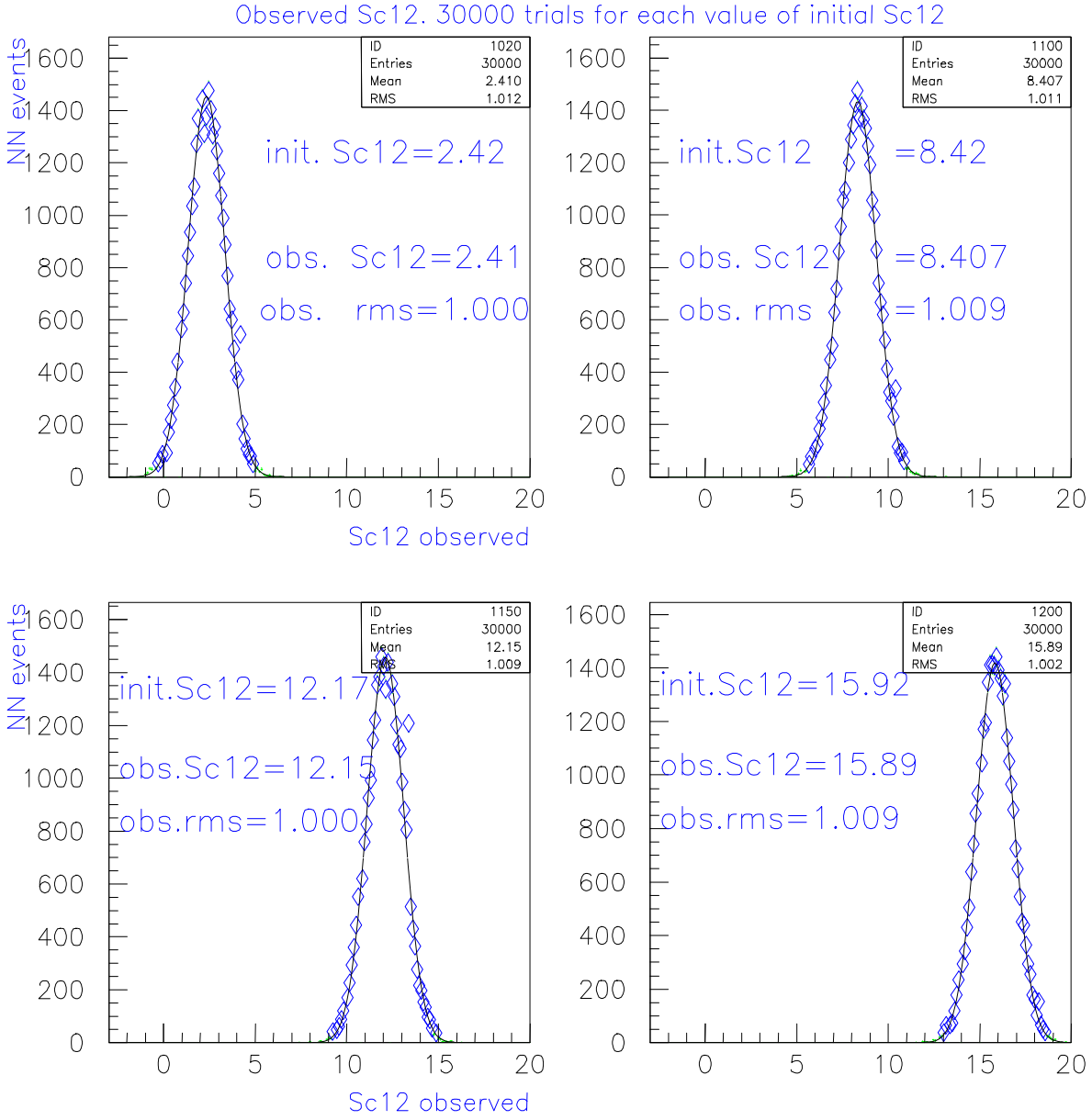


The method of the study

We use the method which allows to connect the magnitude of the observed significance with the confidence density of the parameter “the internal significance”.

We carried out the uniform scanning of internal significance S_{c12}, varying S_{c12} from 1 up to 16, using step size 0.075. By playing with the two Poisson distributions (with parameters s and b) and using 30000 trials for each value of S_{c12} to construct the conditional distribution of the probability of the production of the observed value of significance S_{c12} by the internal significance S_{c12}. Integral luminosity of the experiment is a constant $s+b$. The parameters s and b are chosen in accordance with the given internal significance S_{c12}, the realization N_{obs} (or $s+b$) is a sum of realizations N_s (or s) and N_b (or b).

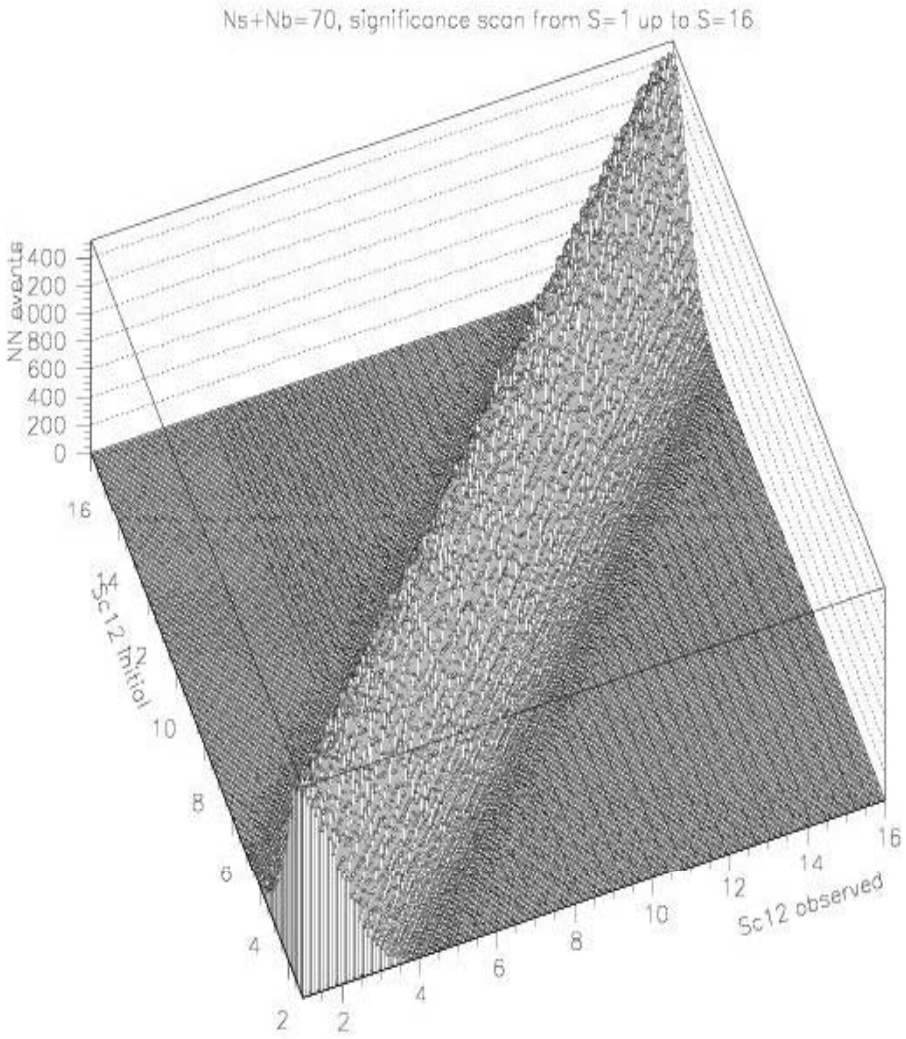
The observed significance



The distributions of S_{c12} of several values of internal significance S_{c12} with the given integral luminosity $s+b=70$ are presented.

The observed distributions of significances are similar to the distributions of the realizations of normal distributed random variable with variance which close to **1.**

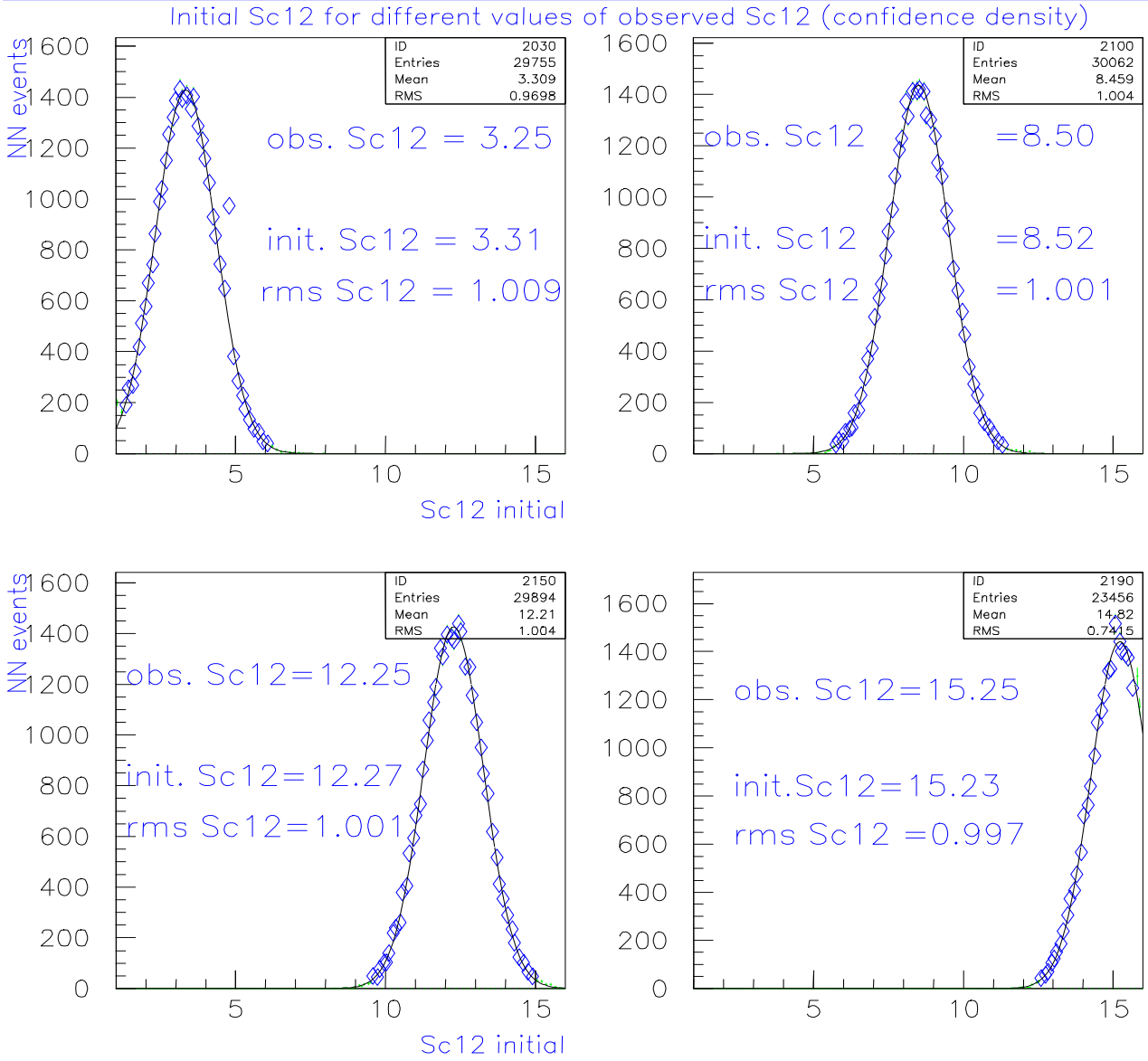
Relationship between internal and observed significances



The distribution of the observed significance S_{c12} versus the internal significance S_{c12} shows the result of the full scanning.

The normal distributions with a fixed variance are statistically self-dual distributions. It means that the confidence density of the parameter "internal significance" Z has the same distribution as the random variable which produced a realization of the observed significance Z.

The internal significance



The several distributions of the probability of the internal significances S_{c12} to produce the observed values of S_{c12} are presented. These figures clearly show that the observed significance S_{c12} is an unbiased estimator of the internal significance S_{c12}.

The statement I

The observed significance S_{c12} (the case of the Poisson flow of events) is a realisation of the random variable which can be approximated by normal distribution with variance close to 1
(for example, it is a standard normal distribution *N(0,1)* in the case of pure background without signal).

It means that with this observed significance one can work as with the realization of the random variable.

The combining significances

Let us define the observed summary significance $\mathbf{Z_{sum}}$, the observed ***combined*** significance $\mathbf{Z_{comb}}$ and the observed mean significance $\mathbf{Z_{mean}}$ for the $\mathbf{L}$ partial observed significances $\mathbf{Z_i}$ with standard deviation $\boldsymbol{\sigma}(\mathbf{Z_i}) \sim \mathbf{1}$:

$$Z_{sum} = \sum_{i=1}^{L} Z_i ,$$

$$\sigma^2(Z_{sum}) = \sum_{i=1}^{L} \sigma^2(Z_i),$$

$$Z_{comb} = \frac{Z_{sum}}{\sqrt{\sigma^2(Z_{sum})}},$$

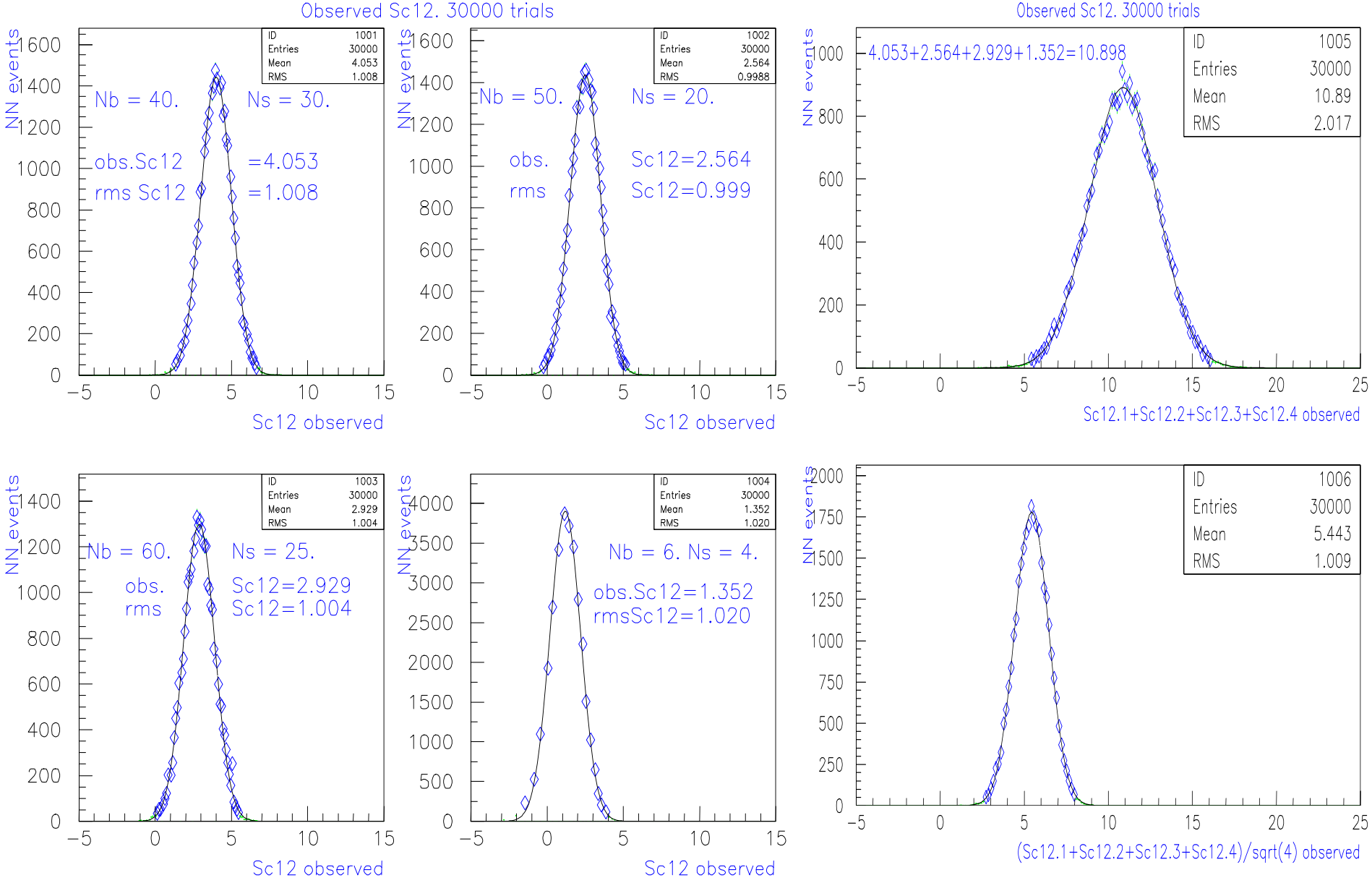
$$Z_{mean} = \frac{Z_{sum}}{L}.$$

The statement II

The ratio of the sum of the several partial observed significances and the standard deviation of this sum is the estimator of the combining significance of several partial observed significances. It is essentially Stouffer's method.

It can also be shown by a Monte Carlo simulation. Let us generate the observation of the significances for four experiments with different parameters ***b*** and ***s*** simultaneously. The results of this simulation (30000 trials) for each experiment are presented in next slide. The distribution of the sums of four observed significances of experiments in each trial and the distribution of these sums divided by 2 (i.e. sqrt(4)) in each trials is shown too.

Sc12 – partial, summary and combined significances



Confidence distributions

The consecutive theory of combining information from independent sources through confidence density is proposed in paper (Singh et al., 2005).

Suppose $X_1, \ldots, X_n$ are ***n*** independent random draws from a population **F** and **χ** is the sample space corresponding to the data set $\mathbf{X_n} = (X_1, \ldots, X_n)'$. Let ***θ*** be a parameter of interest associated with **F**, and let **Θ** be the parameter space.

A function $\mathbf{H_n(.)=H_n(X_n,(.))}$ on $\boldsymbol{\chi}$ x **Θ -> [0,1]** is called a **confidence distribution** (CD) for a parameter ***θ*** if

(i) for each given Xn € $\boldsymbol{\chi}$, $\mathbf{H_n(.)}$ is a continuous **cdf**;

(ii) At the true parameter value $\boldsymbol{\theta=\theta_0}$, $\mathbf{H_n(\theta_0)=H_n(X_n,\theta_0)}$, as a function of the sample $\mathbf{X_n}$, has the uniform distribution **U(0,1)**.

We call, when it exists, $\boldsymbol{h_n(\theta)=H_n'(\theta)}$ a **confidence density.**

Combination of CD (I)

Let $\mathbf{H_1(y), \ldots, H_L(y)}$ be **L** independent CDs, with the same true parameter $\boldsymbol{\theta}$. Suppose $\mathbf{g_c(U_1, \ldots, U_L)}$ is any continuous function from $\mathbf{[0,1]^L}$ to **R** that is monotonic in each coordinate. A general way of combining, depending on $\mathbf{gc(U_1, \ldots, U_L)}$ can be described as follows: Define $\mathbf{Hc(U_1, \ldots, U_L)=G_c(g_c(U_1, \ldots,U_L))}$, where $\mathbf{G_c(.)}$ is the continuous **cdf** of $\mathbf{g_c(U_1, \ldots, U_L)}$, and $\mathbf{U_1, \ldots, U_L}$ are independent **U(0,1)** distributed random variables. Denote $\mathbf{H_c(y)=H_c(H_1(y), \ldots, H_L(y))}$. It is a CD and it is a combined CD.

Let $\mathbf{F_0(.)}$ be any continuous **cdf** and a convenient special case of the function $\boldsymbol{g_c}$ is expressed via inverse function of $\mathbf{F_0(.)}$

$$g_c(U_1,\ldots,U_L)=F_0^{-1}(U_1)+\ldots+F_0^{-1}(U_L).$$

In this case, $\mathbf{G_c(.)=F_0*\ldots*F_L}$, where * stands for convolution.

Combination of CD (II)

This general CD combination recipe is simple and easy to implement. Two examples of $\mathbf{F_0}$ are:

1. $\mathbf{F_0(t)=\Phi(t)}$ is the **cdf** of the standard normal. In this case

$$H_{NM}(y) = \Phi\left(\frac{1}{\sqrt{L}}[\Phi^{-1}(H_1(y)) + \ldots + \Phi^{-1}(H_L(y))]\right).$$

One can see that this formula leads to the formula of Stouffer.

2. $\mathbf{F_0(t)=1-\exp(-t)}$, for $\mathbf{t \geq 0}$, is the **cdf** of the standard exponential distribution (with mean=1). In this case the combined CD is

$$H_{E1}(y) = P\left(\chi^2_{2L} \leq -2\sum_{i=1}^{L}\log(1-H_i(y))\right),$$

It is well known Fisher's omnibus method.

Probability of incorrect decision

The uncertainty in hypotheses testing is determined by two types of errors: Type I error α - probability to accept hypothesis H_1 if hypothesis H_0 is correct and Type II error β - probability to accept hypothesis H_0 if hypothesis H_1 is correct.

In our case by definition $\underline{Z_e}$ corresponds to $\alpha = 1 - \Phi(\underline{Z_e})$ and $\beta = 0.5$ (because the $\underline{Z_e}$ is an unbiased estimator of Z_p, we suppose that 50% of realizations of Z under condition $Z_p = \underline{Z_e}$ will lie below Z_p and 50% of realizations will lie above Z_p). Z_{comb} satisfies the same condition by construction.

If we take the probability of incorrect decision κ as a measure of uncertainty then we have the condition on critical value for minimization of uncertainty (in considered case) $\alpha = \beta$ (Bityukov et al., 2004). This probability for Z_{comb} equals $\kappa = \alpha = 1 - \Phi(Z_{comb}/2)$.

Comment

Comment: ***About weights***. *Partial significances* $\mathbf{Z_1}$ *and* $\mathbf{Z_2}$ *combine with third partial significances* $\mathbf{Z_3}$ *according to formula*

$$\mathbf{((Z_1+Z_2)\,/\,\sqrt{2})\,*\,\sqrt{2}\,/\,\sqrt{3}\;+\,Z_3\,*\,1\,/\,\sqrt{3}\,=\,(Z_1+Z_2+Z_3)\,/\,\sqrt{3}}.$$

Conclusion

As shown, the Stouffer's method of combining significances works for significances which obey the normal distribution.

The significances S_{c12}, Z_N, Z_{Bi}, and Z_{PL} satisfy to the criterion of normality in wide range of values ***s*** and ***b*** in Poisson flows.

The choice of the combination method depends on many factors. As seems, the confidence distributions are often convenient for combining information from independent sources. This approach also leads to the Stouffer's formula in our case.

Note, any method for combining ***P***-values, considered in (Cousins, 2007), can be used for combining significances and vice versa. These methods provide the normality of Z_{comb} if partial ***Z***'s are normal.

Acknowledgments

We are grateful to Vladimir Gavrilov, Vyacheslav Ilin, Andrei Kataev, Vassili Katchanov, and Victor Matveev for the interest and support of this work. We thank Robert Cousins and Sergei Gleyser for stimulating, educational discussions. S.B. would like to thank the Organizing Committee of ACAT 2008 for hospitality and support.

References

S.I. Bityukov, N.V. Krasnikov (1998). Modern Physics Letter A13, 3235.

S.I. Bityukov, N.V. Krasnikov (2004). Nucl.Instr.&Meth., A534, 152-155.

S. Bityukov, N. Krasnikov, and A. Nikitenko (2006). On the combining significances. physics/0612178.

R.D. Cousins (2007) Annotated Bibliography of Some Papers on Combining Significances or p-values, arXiv:0705.2209 [physics.data-an].

Robert D. Cousins, James T. Linnemann, Jordan Tucker (2008) Nucl.Instr. & Meth. A595, 480--501.

R. A. Fisher (1970). Statistical Methods for Research Workers. Hafner, Darien, Connecticut, 14th edition. The method of combining significances seems to have appeared in the 4th edition of 1932.

A.G.Frodesen, O.Skjeggestad, H.Tøft, Probability and Statistics in Particle Physics, UNIVERSITETSFORLAGET, Bergen-Oslo-Tromsø, 1979. p.408.

References

I. Narsky (2000). Nucl.Instr.&Meth. A450, 444.

K. Pearson (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. Biometrika, 25(3/4):379—410.

K. Singh, M. Xie, W. Strawderman (2005). Combining information from independent Sources through confidence distributions. Annals of Statistics, 33, 159-183.

S. Stouffer, E. Suchman, L. DeVinnery, S. Star, and R.W. Jr (1949). The American Soldier, volume I: Adjustment during Army Life. Princeton University Press.

L. Tippett (1931). The Methods of Statistics. Williams and Norgate, Ltd., London, 1st edition. Sec. 3.5, 53-6, as cited by Birnbaum and by Westberg.