On the notion "significance of the difference"

Sergey Bityukov[1,2], Nikolai Krasnikov[2], Nikita Okunev[3] and Vera Taperechkina[3]

[1] National Research Center "Kurchatov Institute" Institute for High Energy Physics, Protvino, Russia
[2] Institute for Nuclear Research Russian Academy of Science, Moscow, Russia
[3] Moscow Technological University (MIREA), Moscow, Russia

E-mail: Serguei.Bitioukov@cern.ch

Abstract. The concept of the significance of a signal in presence of background in experiment is widely used in data analysis in high-energy physics. It is shown that when comparing pairs independent samples obtained from one population (Poisson flow of events), the distribution of estimates of the significance is asymptotically standard normal distribution.

1. Introduction

The concept of "the significance" of a signal in presence of background in experiment [1, 2] (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.

We consider the significance Type C [3]. This significance is used in works [4, 5, 6] (see. also, [7]) for analysis of experimental data.

It is shown that when comparing two independent samples obtained from one population, the distribution of estimates of "the significance of the difference" is asymptotically standard normal distribution.

Let a sample of realizations of some random variable be obtained from an infinite population within a given time. Each realization is called an event. The stream of events in experiment usually is a Poisson flow. 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. 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.

In the simplest case, this 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) which appeared during the given time.

In a real experiment, the numbers of background and signal events are realizations of random variables. The distributions of the observed number of background events $\hat{b}$ and the observed number of signal events $\hat{s}$ in Poisson flow of events obey Poisson distributions with parameters b (expected number of background events) and s (expected number of signal events), respectively.

Note, the realization of random variable allows to estimate the parameter of Poisson distribution. It means that when we compare the estimated parameters of Poisson flows of

events we compare two samples.

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

2. Classification of significances

In paper [3], a classification of significances in accordance with the scope of applicability was proposed. Let us S characterizes the significance of signal. The choice of significance to be used depends on the study. There are three types of significances.

A. If s and b are expected values then we take into account both statistical fluctuations of signal and fluctuations of background. Before observation we can calculate only an expected significance S which is a parameter of experiment. S characterizes the quality of experiment ($S_{c12} = 2(\sqrt{s+b} - \sqrt{b})$ [3] as an example).

B. If $\widehat{s+b}$ 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 $\hat{S}$ which is an estimator of expected significance of experiment S. $\hat{S}$ characterizes the quality of experimental data (S_{cP} as an example). S_{cP}, probable, was proposed for using in HEP in ref. [8]. This significance corresponds a probability to observe number of events equal or greater than $s+b$ in sample with Poisson distribution with mean b which converted to equivalent number of sigmas of a Gaussian distribution.

C. If $\widehat{s+b}$ and $\hat{b}$ are observed values of signal plus background and only background with known errors of measurement then we can use the standard theory of errors to estimate the significance of signal S_d. If measured variables obey Poisson distribution or measured values have normal distributed variances the formula for S_d looks as

$$S_d = \frac{\widehat{s+b} - \hat{b}}{\sqrt{\sigma^2_{\widehat{s+b}} + \sigma^2_{\hat{b}}}}, \tag{1}$$

where $\sigma^2_{\widehat{s+b}}$ and $\sigma^2_{\hat{b}}$ are corresponding variances of errors 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

$$S_d = \frac{\widehat{s+b} - K\hat{b}}{\sqrt{\sigma^2_{\widehat{s+b}} + K^2\sigma^2_{\hat{b}}}}, \tag{2}$$

where K is a ratio of integrated luminosities of experiments.

An important property of these significances is property that when comparing two independent samples obtained from the same population (for example, both samples are without signal events, i.e. $s = 0$) the distribution of estimates of "the significance of the difference" is close to the standard normal distribution $N(0,1)$. It is shown for several significances in paper [3] (significances S_{c12} and S_{cP}) and in paper [6] (significance Eq. 2) by Monte Carlo experiments. Fisz [9] shows that the significance (1) is asymptotically normal $N(0,1)$ in the case $s = 0$ (both samples are taken from the same population).

Let us show the presence of this property for significance in case of Eq. 2.

Let we have two independent samples which are taken from the same stationary Poisson flow of events during time t_1 and t_2, correspondingly. We study subflow with parameter λ, which satisfy the certain conditions (for example, cuts) of this flow. Let us the subflow content only

background events. Let there are choosen $\hat{n}_1$ events in the first sample and $\hat{n}_2$ events in second sample. To determine the significance of the difference we use the formula

$$S_d = \frac{\hat{n}_1 - K\hat{n}_2}{\sqrt{\lambda t_1 + K^2 \lambda t_2}}, \tag{3}$$

where $K = \dfrac{t_1}{t_2}$. Expected numbers of events in samples λt_1 and λt_2 are variances of corresponding Poisson distributions.

Suppose that $\hat{n}_1$ and $\hat{n}_2$ are unbiased estimates of corresponding expected values n_1 and n_2 for given samples (asymptotically it is true [9]), i.e. $E(\hat{n}_i) = n_i = \lambda t_i,\ i = 1, 2$. Note, that $\lambda t_1 + K^2 \lambda t_2$ is constant. Then, due to the stationarity of the flow of events and the independence of samples, we have

$$E(S_d) = E(\frac{\hat{n_1} - K\hat{n}_2}{\sqrt{\lambda t_1 + K^2 \lambda t_2}}) = \frac{E(\hat{n_1}) - E(K\hat{n}_2)}{\sqrt{\lambda t_1 + K^2 \lambda t_2}} = \frac{\lambda t_1 - K\lambda t_2}{\sqrt{\lambda t_1 + K^2 \lambda t_2}} = 0, \tag{4}$$

$$D(S_d) = D(\frac{\hat{n_1} - K\hat{n}_2}{\sqrt{\lambda t_1 + K^2 \lambda t_2}}) = \frac{D(\hat{n_1}) + D(K\hat{n}_2)}{\lambda t_1 + K^2 \lambda t_2} = \frac{\lambda t_1 + K^2 \lambda t_2}{\lambda t_1 + K^2 \lambda t_2} = 1. \tag{5}$$

It means that asymptotical normality of Eq. 1 take place in the case Eq. 2.

3. Conclusion

This property allows to use significance S_d in many applications of data analysis.

Due to this property one can construct the scale for distance between, for example, histograms [10] by the use a multivariate test statistics. As a result,

- we can compare corresponding parts of two histograms,
- we can compare multidimensional histograms likewise as unidimensional histograms,
- we can compare two sets of several histograms simultaniously likewise as we compare a pair of histograms.

We are grateful to Vassili Kachanov for interest and support of this work. We thank Sergei Gleyzer, Yuri Gouz, Andrei Kataev, Nikolai Minaev and Vera Smirnova for helpful and constructive comments. The results of the work are obtained under support of Ministry of Education and Science RF (Agreement on October 17, 2014 N 14.610.21.0004, idintificator PNIER RFMEFI61014X0004).

References

[1] Linnemann J T 2003 *Proc. Conference on Staistical Problems in Particle Physics, Astrophysics and Cosmology (SLAC, Stanford)* Editors: L.Lyons, R.Mount, R.Reitmeyer, SLAC-R-703, p 35
[2] Zhu Y 2006 *High Ener. Phys. Nucl. Phys.* **30** 331 (*ePrint* arXiv:physics/0507145)
[3] Bityukov S, Krasnikov N, Nikitenko A, Smirnova V 2008 *Proc. of Science* PoS (ACAT08) 118
[4] Aubert B et al (BABAR Collaboration) 2008 *Phys.Rev.* D **78** 051102
[5] Bediaga I, Bigi I I, Gomes A, Guerrer G, Miranda J, dos Reis A C 2008 *Phys.Rev.* D **80** 096006
[6] Bityukov S, Krasnikov N, Nikitenko A, Smirnova V 2013 *Eur.Phys. J. Plus* 128:143
[7] Aaij R et al (LHCb Collaboration) 2013 *Phys. Lett.* B **726** 623
[8] Narsky I 2000 *Nucl.Instr.&Meth.*, A **450** 444
[9] Fisz M 1955 *Colloquium Mathematicum* **3** 199; also, Haight F A 1967 *Handbook of the Poisson distribution* (John Wiley & Sons, Inc.) (formula 6.4-1)
[10] Bityukov S I, Krasnikov N V, Maksimishkina A V, Smirnova V V 2016 Multidimensional test statistics and statistical comparison of histograms *Int. Journal of Economics and Statistics* **4** 98