On the notion "significance of the difference"

Sergey Bityukov
SRC Institute for High Energy Physics NRC "Kurchatov Institute"
Nikolai Krasnikov
Institute for Nuclear Research Russian Academy of Science
Nikita Okunev, Vera Taperechkina
Moscow Technological University (MIREA)

Introduction I

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].

It is shown that when comparing two independent samples obtained from one general population, the distribution of estimates of ``the significance of the difference'' for mean values, obtained for these samples, is close to the standard normal distribution.

Introduction II

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 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. 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 of one.

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) that appeared during the given time.

Introduction III

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 $\hat{\boldsymbol{b}}$ 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.

Classification I

In paper [3], a classification of significances in accordance with the scope of applicability was proposed. 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 $\boldsymbol{C}$.

(***A***) If $\boldsymbol{s}$ and $\boldsymbol{b}$ 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 $\boldsymbol{S}$ which is a parameter of experiment. $\boldsymbol{S}$ characterizes the quality of experiment

($\boldsymbol{S_{c12}}$ = 2 ($\sqrt{\boldsymbol{s}+\boldsymbol{b}}$ - $\sqrt{\boldsymbol{b}}$) [3] as example).

(***B***) 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, $\widehat{\boldsymbol{S}}_{\boldsymbol{cP}}$ (was proposed for using in HEP in ref. [7]). 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.

Classification II

(*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

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

where σ^2_{s+b} and σ^2_b 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

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

where K is a ratio of integrated luminosities of experiments.

It should be noted that in the general case, the number of signal events can be negative (for example, destructive interference). For this reason, this significance can be called the *significance of the difference*.

Asymptotical normality 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" for mean values, obtained for these samples, is close to the standard normal distribution ***N(0,1)***.

It is shown for several significances in paper [3] (significances $\boldsymbol{S_{c12}}$ and $\boldsymbol{\widehat{S}_{cP}}$) and in paper [6] (significance $\boldsymbol{S_d}$ (Eq. 2)) by Monte Carlo experiments. Fisz [8] shows that the significance $\boldsymbol{S_d}$ (Eq. 1) in case of Poisson distribution is asymptotically normal ***N(0,1)***.

Let us show the presence of this property for significance in common case ($\boldsymbol{S_d}$ (Eq. 2)).

Asymptotical normality II

Let we have two independent samples which are taken from the same stationary Poisson flow of events during time $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$, correspondingly. . We study subflow with parameter $\boldsymbol{\lambda}$ (which satisfy the certain conditions, for example, cuts) of this flow. Let us the subflow content only background events. Let there are chosen $\boldsymbol{\hat{n}_1}$ events in the first sample and $\boldsymbol{\hat{n}_2}$ events in the second sample.

To determine the significance of the difference we use the formula

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

where $\boldsymbol{K} = \frac{\boldsymbol{t_1}}{\boldsymbol{t_2}}$. Expected numbers of events under study in samples $\boldsymbol{\lambda t_1}$ and $\boldsymbol{\lambda t_2}$ are variances of corresponding Poisson distributions.

Asymptotical normality III

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 [8]),

i.e. $E(\hat{n}_i) = E(n_i) = \lambda t_i$, $i = 1,2$

(it is true also for statistically dual distributions [9]).

Then, due to the stationarity of the flow of events and the independence of samples, we have

$$E(S_d) = E\left(\frac{\hat{n}_1 - K\hat{n}_2}{\sqrt{\lambda t_1 + K^2 \lambda t_2}}\right) = \frac{E(\hat{n}_1) - KE(\hat{n}_2)}{\sqrt{\lambda t_1 + K^2 \lambda t_2}} = \frac{\lambda t_1 - K\lambda t_2}{\sqrt{\lambda t_1 + K^2 \lambda t_2}} = \mathbf{0}\,, \qquad (4)$$

$$D(S_d) = D\left(\frac{\hat{n}_1 - K\hat{n}_2}{\sqrt{\lambda t_1 + K^2 \lambda t_2}}\right) = \frac{D(\hat{n}_1) + D(K\hat{n}_2)}{\lambda t_1 + K^2 \lambda t_2} = \frac{\lambda t_1 + K^2 \lambda t_2}{\lambda t_1 + K^2 \lambda t_2} = \mathbf{1}\,. \qquad (5)$$

It means that asymptotical normality of Eq. 1 [8] take place in the case of Eq. 2.

Conclusions

This property allows to use significance $\boldsymbol{S_d}$ in many applications of data analysis.

Due to this property one can construct the scale for distance between, for example, histograms [10] (or dependencies) 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 simultaneously likewise as we compare a pair of histograms;

In principle, this “significance” is a value which is inverse to relative uncertainty of the difference between two measured values in gauss approximation.

References

[1] J.T. Linnemann, Measures of Significance in HEP and Astrophysics, Proceedings of the Conference on Staistical problems in Particle Physics, Astrophysics and Cosmology, SLAC, Stanford, 8-11 September 2003, Editors: L.Lyons, R.Mount, R.Reitmeyer, SLAC-R-703, pp.35-49, 2003

[2] Y. Zhu, On the Statistical Significance, High Ener. Phys. Nucl. Phys. 30 (2006) 331; arXiv:physics/0507145 [physics.data-an].

[3] S. Bityukov, N. Krasnikov, A. Nikitenko, V. Smirnova, Two approaches to Combining Significances, PoS (ACAT08) 118.

[4] B. Aubert, et al (BABAR Collaboration), Search for CP Violation in Neutral D Meson Cabibbo-suppressed Three-body Decays, Phys.Rev. D78 (2008) 051102.

[5] I. Bediaga, I.I. Bigi, A. Gomes, G. Guerrer, J. Miranda, A.C. dos Reis, On a CP Anisotropy Measurement in the Dalitz Plot, Phys.Rev. D80 (2009) 096006.

References

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

[7] I. Narsky, Estimation of upper limits using a Poisson statistic, Nucl.Instr.&Meth., A450, 444-455, 2000.

[8] M. Fisz, The limiting distribution of a function of two independent random variables and its statistical applications, Colloquium Mathematicum, 3 (1955) 199-202;
also, F.A. Haight, Handbook of the Poisson distribution, John Wiley & Sons, Inc., 1967 (formula 6.4-1).

[9] S. Bityukov, N. Krasnikov, S. Nadarajah, V. Smirnova, Statistically Dual Distributions and Estimation, Applied Mathematics, 5 (2014) 963-968. doi: 10.4236/am.2014.56091.

[10] 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.