On Statistical Significance of Signal[*]

ZHU Yong-Sheng[1)]

(Institute of High Energy Physics, CAS, Beijing 100049, China)

Abstract A definition for the statistical significance of a signal in an experiment is proposed by establishing a correlation between the observed p-value and the normal distribution integral probability, which is suitable for both counting experiment and continuous test statistics. The explicit expressions to calculate the statistical significance for both cases are given.

Key words statistical significance, p-value, normal probability, likelihood-ratio, Poisson distribution

1 Introduction

The statistical significance of a signal in an experiment of particle physics is to quantify the degree of confidence that the observation in the experiment either confirm or disprove a null hypothesis H_0, in favor of an alternative hypothesis H_1. Usually H_0 stands for the known or background processes, while the alternative hypothesis H_1 stands for a new or a signal process plus background processes with respective production cross section. This concept is very useful for usual measurements that one can have an intuitive estimation, to what extent one can believe the observed phenomena are due to backgrounds or a signal. It becomes crucial for the measurements which claim a new discovery or a new signal. As a convention in particle physics experiment, the "5σ" standard, namely the statistical significance $S \geqslant 5$ is required to define the sensitivity for discovery; while in the cases $S \geqslant 3$ ($S \geqslant 2$), one may claim that the observed signal has strong (weak) evidence.

However, as pointed out in Ref. [1], the concept of the statistical significance has not been employed consistently in the most important discoveries made over the last quarter century. Also, the definitions of the statistical significance in different measurements

differ from each other. Listed below are various definitions for the statistical significance in counting experiment (see, for example, Refs. [2—4]):

$$S_1 = (n-b)/\sqrt{b}, \tag{1}$$

$$S_2 = (n-b)/\sqrt{n}, \tag{2}$$

$$S_{12} = \sqrt{n} - \sqrt{b}, \tag{3}$$

$$S_{B1} = S_1 - k(\alpha)\sqrt{n/b}, \tag{4}$$

$$S_{B12} = 2S_{12} - k(\alpha), \tag{5}$$

$$\int_{-\infty}^{S_N} N(0,1)\mathrm{d}x = \sum_{i=0}^{n-1} \mathrm{e}^{-b}\frac{b^i}{i!}, \tag{6}$$

where n is the total number of the observed events, which is the Poisson variable with the expectation $s+b$, s is the expected number of signal events to be searched, while b is the known expected number of Poisson distributed background events. All numbers are counted in the "signal region" where the searched signal events are supposed to appear. In Eqs. (4) and (5), $k(\alpha)$ is a factor related to α that the corresponding statistical significance assumes $1-\alpha$ acceptance for positive decision about signal observation, and $k(0.5) = 0$, $k(0.25) = 0.66$, $k(0.1) = 1.28$, $k(0.05) = 1.64$ etc.[3]. In Eq. (6), $N(0,1)$ is a notation for the normal function with the expectation and variance equal to 0 and 1, respectively. On the

other hand, the measurements in particle physics often examine statistical variables that are continuous in nature. Actually, to identify a sample of events enriched in the signal process, it is often important to take into account the entire distribution of a given variable for a set of events, rather than just to count the events within a given signal region of values. In this situation, I. Nasky[4] gives a definition of the statistical significance via likelihood function

$$S_L = \sqrt{-2\ln L(b)/L(s+b)} \tag{7}$$

under the assumption that $-2\ln L(b)/L(s+b)$ distributes as χ^2 function with degree of freedom of 1.

Upon the above situation, it is clear that we desire to have a self-consistent definition for statistical significance, which can avoid the danger that the same S value in different measurements may imply virtually different statistical significance, and can be suitable for both counting experiment and continuous test statistics. In this letter we propose a definition of the statistical significance, which could be more close to the desired property stated above.

2 Definition of the statistical significance

The p-value is defined to quantify the level of agreement between the experimental data and a hypothesis[1, 5]. Assume an experiment makes a measurement for test statistic t being equal to t_{obs}, and t has a probability density function $g(t|H_0)$ if a null hypothesis H_0 is true. We futher assume that large t values correspond to poor agreement between the data and the null hypothesis H_0, then the p-value of an experiment would be

$$p(t_{\mathrm{obs}}) = P(t > t_{\mathrm{obs}}|H_0) = \int_{t_{\mathrm{obs}}}^{\infty} g(t|H_0)\mathrm{d}t. \tag{8}$$

A very small p-value tends to reject the null hypothesis H_0.

Since the p-value of an experiment provides a measure of the consistency between the H_0 hypothesis and the measurement, our definition for statistical significance S relates with the p-value in the form of

$$\int_{-S}^{S} N(0,1)\mathrm{d}x = 1 - p(t_{\mathrm{obs}}), \tag{9}$$

under the assumption that the null hypothesis H_0 represents that the observed events can be described merely by background processes. Because a small p-value means a small probability of H_0 being true, corresponds to a large probability of H_1 being true, one would get a large signal significance S for a small p-value, and vice versa. The left side of Eq. (9) represents the probability of the normal distribution in the region within $\pm S$ standard deviation ($\pm S\sigma$), therefore, this definition conforms itself to the meaning of that the statistical significance should have. In such a definition, some correlated S and p-values are listed in Table 1.

Table 1.   The statistical significance S and correlated p-value.

S	p-value
1	0.3173
2	0.0455
3	0.0027
4	6.3×10^{-5}
5	5.7×10^{-7}
6	2.0×10^{-9}

3 Statistical significance in counting experiment

A group of particle physics experiment involves the search for new phenomena or signal by observing a unique class of events that can-not be described by background processes. One can address this problem to that of a "counting experiment", where one identifies a class of events using well-defined criteria, counts up the number of observed events, and estimates the average rate of events contributed by various backgrounds in the signal region, where the signal events (if exist) will be clustered. Assume in an experiment, the number of signal events in the signal region is a Poisson variable with the expectation s, while the number of events from backgrounds is a Poisson variable with a known expectation b without error, then the observed number of events distributes as the Poisson variable with the expectation $s+b$. If the experiment observed n_{obs} events in the signal

region, then the p-value is

$$p(n_{\text{obs}}) = P(n > n_{\text{obs}}|H_0) =$$

$$\sum_{n=n_{\text{obs}}}^{\infty} \frac{b^n}{n!}\text{e}^{-b} = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{b^n}{n!}\text{e}^{-b} \ . \quad (10)$$

Substituting this relation to Eq. (9), one immediately has

$$\int_{-S}^{S} N(0,1)\text{d}x = \sum_{n=0}^{n_{\text{obs}}-1} \frac{b^n}{n!}\text{e}^{-b} \ . \quad (11)$$

Then, the signal statistical significance S can be easily determined. Comparing this equation with Eq. (6) given by Ref. [4], we notice the lower limit of the integral is different.

4　Statistical significance in continuous test statistics

The general problem in this situation can be addressed as follows. Suppose we identify a class of events using well-defined criteria, which are characterized by a set of N observations $X_1, X_2, \cdots, X_N$ for a random variable X. In addition, one has a hypothesis to test that predicts the probability density function of X, say $f(X|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)$ is a set of parameters which need to be estimated from the data. Then the problem is to define a statistic that gives a measure of the consistency between the distribution of data and the distribution given by the hypothesis.

To be concrete, we consider the random variable X is, say, an invariant mass, and the N observations $X_1, X_2, \cdots, X_N$ give an experimental distribution of X. Assuming parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k) \equiv (\boldsymbol{\theta}_s; \boldsymbol{\theta}_b)$, where $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_b$ represent the parameters related to signal (say, a resonance) and backgrounds contribution, respectively. We assume the null hypothesis H_0 stands for that the experimental distribution of X can be described merely by the background processes, while the alternative hypothesis H_1 stands for that the experimental distribution of X should be described by the backgrounds plus signal; namely, the null hypothesis H_0 specifies the fixed value(s) for a subset of parameters $\boldsymbol{\theta}_s$ (the number of

fixed parameter(s) is denoted as r), while the alternative hypothesis H_1 leaves the r parameter(s) free to take any value(s) other than those specified in H_0. Therefore, the parameters $\boldsymbol{\theta}$ are restricted to lie in a subspace ω of its total space Ω. On the basis of a data sample of size N from $f(X|\boldsymbol{\theta})$, we want to test the hypothesis H_0:$\boldsymbol{\theta}$ belongs to ω. Given the observations $X_1, X_2, \cdots, X_N$, the likelihood function is $L = \prod_{i=1}^{N} f(X_i|\boldsymbol{\theta})$. The maximum of this function over the total space Ω is denoted by $L(\hat{\Omega})$; while within the subspace ω the maximum of the likelihood function is denoted by $L(\hat{\omega})$, then we define the likelihood-ratio $\lambda \equiv L(\hat{\omega})/L(\hat{\Omega})$. It can be shown that for H_0 true, the statistic

$$t \equiv -2\ln\lambda \equiv 2(\ln L_{\max}(s+b) - \ln L_{\max}(b)) \quad (12)$$

is distributed as $\chi^2(r)$ when the sample size N is large[6]. In Eq. (12) we use $L_{\max}(s+b)$ and $L_{\max}(b)$ denoting $L(\hat{\Omega})$ and $L(\hat{\omega})$, respectively. If λ turns out to be in the neighborhood of 1, the null hypothesis H_0 is such that it renders $L(\hat{\omega})$ close to the maximum $L(\hat{\Omega})$, and hence H_0 will have a large probability of being true. On the other hand, a small value of λ will indicates that H_0 is unlikely. Therefore, the critical region of λ is in the neighborhood of 0, corresponding to a large value of statistic t. If the measured value of t in an experiment is t_{obs}, from Eq. (8) we have p-value

$$p(t_{\text{obs}}) = \int_{t_{\text{obs}}}^{\infty} \chi^2(t;r)\text{d}t. \quad (13)$$

Therefore, in terms of Eq. (9), we can calculate the signal significance according to the following expression:

$$\int_{-S}^{S} N(0,1)\text{d}x = 1 - p(t_{\text{obs}}) = \int_{0}^{t_{\text{obs}}} \chi^2(t;r)\text{d}t. \quad (14)$$

For the case of $r = 1$, we have

$$\int_{-S}^{S} N(0,1)\text{d}x = \int_{0}^{t_{\text{obs}}} \chi^2(t;1)\text{d}t = 2\int_{0}^{\sqrt{t_{\text{obs}}}} N(0,1)\text{d}x,$$

and immediately obtain

$$S = \sqrt{t_{\text{obs}}} = [2(\ln L_{\max}(s+b) - \ln L_{\max}(b))]^{1/2}, \quad (15)$$

which is identical to Eq. (7) given by Ref. [4].

5 Discussion and summary

In Section 2, the p-value defined by Eq. (8) is based on the assumption that large t values correspond to poor agreement between the null hypothesis H_0 and the observed data, namely, the critical region of statistic t for H_0 lies on the upper side of its distribution. If the situation is such that the critical region of statistic t lies on the lower side of its distribution, then Eq. (8) should be replaced by

$$p(t_{\mathrm{obs}}) = P(t < t_{\mathrm{obs}}|H_0) = \int_{-\infty}^{t_{\mathrm{obs}}} g(t|H_0)\mathrm{d}t, \qquad (16)$$

and the definition of statistical significance S expressed by Eq. (9) is still applicable. For the case that the critical region of statistic t for H_0 lies on both lower and upper tails of its distribution, and one determined from an experiment the observed t values in both sides: t_{obs}^{U} and t_{obs}^{L}, then Eq. (8) should be replaced by

$$p(t_{\mathrm{obs}}) = P(t < t_{\mathrm{obs}}^{L}|H_0) + P(t > t_{\mathrm{obs}}^{U}|H_0) =$$
$$\int_{-\infty}^{t_{\mathrm{obs}}^{L}} g(t|H_0)\mathrm{d}t + \int_{t_{\mathrm{obs}}^{U}}^{\infty} g(t|H_0)\mathrm{d}t. \quad (17)$$

In summary, we proposed a definition for the statistical significance by establishing a correlation between the normal distribution integral probability and the p-value observed in an experiment, which is suitable for both counting experiment and continuous test statistics. The explicit expressions to calculate the statistical significance for counting experiment and continuous test statistics in terms of the Poisson probability and likelihood-ratio are given.

References

1 Sinervo P K. Proc. of Conf., Advanced Statistical Techniques in Particle Physics. Durham, UK, 18-22 March, 2002, 64. hep-ex/0208005

2 Bityukov S I et al. Nucl. Instrum. and Methods, 2000, **A452**: 518

3 Bityukov S I et al. Proc. of Conf. Advanced Statistical Techniques in Particle Physics. Durham, UK, 18-22 March, 2002. 77

4 Narsky I. Nucl. Instrum. and Methods, 2000, **A450**: 444

5 Eidelman S et al(Particle Data Group). Phys. Lett., 2004, **B592**: 1

6 Eadie W T et al. Statistical Methods in Experimental Physics. Amsterdam: North-Holland Publishing Co., 1971. Section 10.5

关于信号的统计显著性[*]

朱永生[1)]

(中国科学院高能物理研究所　北京　100049)

摘要 基于实验测量的 p 值和正态函数积分概率之间的对应关系, 提出了实验测量中信号统计显著性的定义, 它对于"计数"实验和连续检验统计量这两种情形都适用, 并给出了计算信号显著性的相应公式.

关键词 统计显著性 p 值 正态概率 似然比 泊松分布