Astron. Astrophys. Suppl. Ser. **132**, 301–304 (1998)

A statistical method for testing assumed distributions of sources

Yi-Ping Qin[1], **G.Z. Xie**[1], **Xue-Tang Zheng**[2], **and Shi-Min Wu**[3]

[1] Yunnan Observatory, The Chinese Academy of Sciences, Kunming, Yunnan 650011, PR China
[2] Department of Physics, Nanjing University of Science and Technology, Nanjing, Jiangsu 210014, PR China
[3] Department of Astronomy, Beijing Normal University, Beijing 100875, PR China

Received November 14; accepted February 24, 1998

Abstract. For a known distribution of sources, the variance of the distribution function of samples can be obtained by applying mathematical statistics. A statistical method, called the 1σ distribution function deviation test, is defined and introduced to test this kind of distributions. We find that, when random variables vary continuously, a sample passing the test also passes any other statistical test which depends on the deviation of a statistic from its expected value at the 1σ confidence level. In particular, the sample passes the Kolmogorov-Smirnov test at the same confidence level. This new method is suitable for testing the distribution of gamma-ray bursts as well as the luminosity function of quasars.

An example of application of the new test is also presented in this paper.

Key words: galaxies: luminosity function, mass function — galaxies: statistic — methods: statistical

1. Introduction

In the study of distributions of sources, the shape of the distribution is often assumed to be known. Thus, one searches for the mean value of some statistics of the distribution. For example, in checking the uniformity of the distribution of gamma-ray bursts, one looks for the mean value of $V/V_{\max}$, $< V/V_{\max} >$, where V is the volume enclosed at the observed distance of a source and $V_{\max}$ is the volume enclosed at the maximum distance where the source would be detectable (see, e.g., Higdon & Schmidt 1990). Also, in checking the uniformity of the distribution of quasars, one applies the test of $< V/V_{\max} >$, or its variant $< V'/V'_{\max} >$, (Schmidt 1968; Mathez 1976; Avni & Bahcall 1980; Hartwick & Schade 1990). The mean value reflects a collective, rather than partial, behavior of the distribution. An unsatisfactory feature of the method is the fact that a given mean value may correspond to different distributions. Aside from the mean value method, the χ^2 test and the Kolmogorov-Smirnov (K-S) test have been frequently applied to test distributions (see, e.g., Boyle et al. 1987, 1988; Hartwick & Schade 1990; Warren et al. 1994; Pei 1995). The two tests correspond to different aspects of distributions. The K-S test concerns the maximum value of the deviation between the observed distribution function of a sample and a given cumulative distribution function, while the χ^2 test reflects the behavior of intervals of a sample compared with a given cumulative distribution function. The K-S test is more sensitive than the χ^2 test and is not affected by the artificial way of dividing intervals as the latter does.

In Sect. 2, we define and introduce a new statistical method for the test of known distributions. In Sect. 3, some properties of the new test are discussed and presented. An example of application of the new test is given in Sect. 4 and a summary of this paper is given in Sect. 5.

2. The test

We consider the cases for which distributions of sources are assumed to be known. In these cases, a probability $p\{\xi < x\}$ is well defined on every event $\{\xi < x\}$. This is the probability that the event is found in the range $\xi < x$, where ξ is the random variable. The cumulative distribution function of the random variable ξ for a given distribution is defined as

$$f(x) \equiv p\{\xi < x\}. \tag{1}$$

For a sample with size N, the frequency of events with $\{\xi < x\}$, $p^*\{\xi < x\}$, is used to define the distribution function

$$f_N(x) \equiv p^*\{\xi < x\}. \tag{2}$$

Let the sample be $S \equiv \{x_i | i \in I\}$, $I \equiv \{i | 1 \leq i \leq N\}$, where $x_i \leq x_{i+1}$, $\forall i, i+1 \in I$. According to the above definition, we have

$$f_N(x) = \begin{cases} 0 & (x \leq x_1) \\ \frac{i}{N} & (x_i < x \leq x_{i+1}) \, . \\ 1 & (x_N < x) \end{cases} \tag{3}$$

According to (1), giving a cumulative distribution function $f(x)$, the probability of event $\{\xi < x\}$ is known. This is $f(x)$ itself. For a sample, the number of events with $\{\xi < x\}$, N_x, follows the well-known binomial distribution, with

$$E\{N_x\} = Nf(x) \tag{4}$$

and

$$\mathrm{Var}\{N_x\} = Nf(x)[1 - f(x)], \tag{5}$$

where N is the size of the sample. According to (2), the distribution function of the sample is

$$f_N(x) = \frac{N_x}{N}. \tag{6}$$

It is clear that $f_N(x)$ is simply the percentage of the x_i's values less than x in the sample of size N.

Substituting (6) into (4) and (5) we have

$$E\{f_N(x)\} = \frac{E\{N_x\}}{N} = f(x) \tag{7}$$

and

$$\mathrm{Var}\{f_N(x)\} = \frac{\mathrm{Var}\{N_x\}}{N^2} = \frac{f(x)[1 - f(x)]}{N}. \tag{8}$$

Since the variance of the distribution function is known for any given value of the random variable, we are able to compare observational data with the given distribution at any data point. This leads to the following definition.

Definition 1. For a sample with size N, if the following condition

$$|f_N(x) - f(x)| < \sqrt{\frac{f(x)[1 - f(x)]}{N}} \tag{9}$$

is satisfied for any given x, the sample is said to pass the 1σ distribution function deviation test for the distribution of $f(x)$.

It is clear that this method concerns the individual, rather than the collective, behavior of samples. By comparing the distribution function of a sample with the given cumulative distribution function one can tell if the sample obeys the given distribution at the 1σ confidence level using Eq. (9).

3. Some properties

Any statistical test depends on the data made of random variables. Therefore, it is necessary to know the variance of these random variables.

Throughout this paper, we only consider the cases for which random variables vary continuously. In this kind of cases, the cumulative distribution function can be expressed as

$$f(x) = \int_{-\infty}^{x} \rho(x)\mathrm{d}x, \tag{10}$$

where $\rho(x)$ is the density function of the distribution (or the distribution density). $\rho(x)$ has the following properties:

1 $\rho(x) \geq 0$;
2 $\int_{-\infty}^{\infty} \rho(x)\mathrm{d}x = 1$;
3 $p\{a \leq \xi \leq b\} = f(b) - f(a) = \int_a^b \rho(x)\mathrm{d}x$.

According to (10) and property 1, we have

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \rho(x) \geq 0. \tag{11}$$

We can limit our discussion to the cases where $\rho(x) \neq 0$. The reason is that, when $\rho(x) = 0$, $\{\xi = x\}$ is an impossible event. A sample containing such events will not be adopted, or for such a sample, the distribution will not be applied. Therefore, in the concerned intervals, $f(x)$ must be a continuous and monotonic function of x. Thus, in these intervals, there must exist a continuous and monotonic inverse function

$$x = x(f) \tag{12}$$

and its derivative

$$\frac{\mathrm{d}x}{\mathrm{d}f} = \frac{1}{\frac{\mathrm{d}f}{\mathrm{d}x}} = \frac{1}{\rho(x)}. \tag{13}$$

According to the error propagation law, we have

$$\mathrm{Var}\{x\} = \left(\frac{\mathrm{d}x}{\mathrm{d}f}\right)^2 \mathrm{Var}\{f\} = \frac{1}{\rho^2(x)}\mathrm{Var}\{f\} \tag{14}$$

in these intervals.

The deviation of the distribution function $f_N(x)$ of a sample from the given cumulative distribution function $f(x)$, $|f_N(x) - f(x)|$, can be considered to be caused by the deviation of the random variable, x_i, $i \in I$, from its expected value, x_{0i}, $i \in I$, where $I \equiv \{i|1 \leq i \leq N\}$. Let

$$x_{0i} = x[f_N(x_i)], \qquad \forall i \in I. \tag{15}$$

Then

$$f(x_{0i}) = f_N(x_i), \qquad \forall i \in I. \tag{16}$$

According to the definition of the distribution function, we have

$$f_N(x_{0i}) = f_N(x_i), \qquad \forall i \in I. \tag{17}$$

Therefore

$$f_N(x_{0i}) = f(x_{0i}), \qquad \forall i \in I. \tag{18}$$

This shows that the value of the distribution function of sample $S_0 \equiv \{x_{0i}|i \in I\}$ at any data point x_{0i}, $f_N(x_{0i})$, meets exactly its expected value, $f(x_{0i})$, $i \in I$. The distribution of S_0 is exactly what would be expected from the given distribution without any deviation. Therefore, the deviation of x_i from x_{0i}, $i \in I$, leads to the deviation of $f_N(x)$ from $f(x)$. Thus, from (14) we have

$$\begin{aligned}\mathrm{Var}\{x_i\} &= \frac{1}{\rho^2(x_i)}\mathrm{Var}\{f_N(x_i)\} \\ &= \frac{f(x_i)[1 - f(x_i)]}{\rho^2(x_i)N}, \qquad \forall i \in I.\end{aligned} \tag{19}$$

The 1σ distribution function deviation test for the distribution of $f(x)$ leads to

$$|x_i - x_{0i}| < \frac{1}{\rho(x_i)}\sqrt{\frac{f(x_i)[1 - f(x_i)]}{N}}, \quad \forall i \in I, \tag{20}$$

which can be interpreted as the 1σ random variable deviation for the distribution.

One can verify that condition (20) can also be obtained by applying Eqs. (12) and (15) together with condition (9).

In the following, we present several statements concluded from Definition 1, which might be useful for statistical analysis.

Statement 1. In the cases for which random variables vary continuously, a sample passing the 1σ distribution function deviation test must also pass any other statistical test which depends on the deviation of a sample from its expected value at the 1σ confidence level.

Proof. Assuming that the random variables vary continuously, take sample $S \equiv \{x_i | i \in I\}$, $I \equiv \{i | 1 \leq i \leq N\}$. Let a statistical function T of the random variables be

$$T = T(x_1, x_2, ..., x_N). \tag{21}$$

Then

$$\begin{aligned}\mathrm{Var}\{T\} &= \sum_{i=1}^{N} \left(\frac{\partial T}{\partial x_i}\right)^2 \mathrm{Var}\{x_i\} \\ &= \sum_{i=1}^{N} \left(\frac{\partial T}{\partial x_i}\right)^2 \frac{f(x_i)[1-f(x_i)]}{\rho^2(x_i)N}.\end{aligned} \tag{22}$$

Assuming sample S pass the 1σ distribution function deviation test, condition (20) is then satisfied. Therefore,

$$|T(x_1, x_2, ..., x_N) - T(x_{01}, x_{02}, ..., x_{0N})| = \sqrt{\sum_{i=1}^{N} (\frac{\partial T}{\partial x_i})^2 (x_i - x_{0i})^2} < \sqrt{\mathrm{Var}\{T\}}. \tag{23}$$

This completes the proof.

Statement 2. If the random variables vary continuously, a sample passing the 1σ distribution function deviation test must also pass any mean-value test at the 1σ confidence level.

This statement is obvious according to Statement 1, as the mean value of any statistical function of a sample is also a statistical function of the random variables.

Statement 3. A sample passing the 1σ distribution function deviation test also passes the Kolmogorov-Smirnov test at the 1σ confidence level.

Proof. Let sample $S \equiv \{x_i | i \in I\}$, $I \equiv \{i | 1 \leq i \leq N\}$, pass the 1σ distribution function deviation test. According to (9), the following relation is satisfied:

$$\sqrt{N}\, |f_N(x) - f(x)| < \sqrt{f(x)[1 - f(x)]}. \tag{24}$$

However,

$$\max\left\{\sqrt{f(x)[1 - f(x)]}\right\} = 0.5, \tag{25}$$

since $0 \leq f(x) \leq 1$. Therefore,

$$\sqrt{N} \sup_{-\infty<x<\infty} |f_N(x) - f(x)| < 0.5. \tag{26}$$

When N is large enough, the distribution of $f_N(x)$ at any given x is near Gaussian. The confidence level of 1σ is 0.683. Let $L(\lambda) = 0.683$, where

$$L(\lambda) = 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2\lambda^2). \tag{27}$$

This gives $\lambda = 0.96$. From (26) we find that

$$\sqrt{N} \sup_{-\infty<x<\infty} |f_N(x) - f(x)| < 0.96. \tag{28}$$

Thus, the sample passes the Kolmogorov-Smirnov test at the 1σ confidence level.

A comparison of Eqs. (26) and (28) shows that if a sample passes the 1σ distribution function deviation test, then it passes it far better than the K-S test at the 1σ confidence level, since $0.5 \ll 0.96$.

4. An example of application

In this section, we illustrate the new test by an application to real data. The data employed are the V'/V'_m values from Table 5 of Schmidt (1968), where the data were believed to be uniformly distributed in the interval $[0, 1]$. Indeed, the mean value of the data is 0.50 (see Schmidt 1968), and one may verify that the sample passes the K-S test.

In this sample, there are 33 data in total. Because some data share the same value, we have only 28 different values. Let x denote the random variable V'/V'_m. For a uniform distribution from 0 to 1, the cumulative distribution function is

$$f(x) = x. \tag{29}$$

The distribution function of the sample can be calculated by Eq. (3). We can then build Table 1. In Table 1, Col. (1) gives the values of the random variable x. Column (2) presents the corresponding values of the cumulative distribution function $f(x)$, from Eq. (29). Column (3) gives the values of the distribution function $f_N(x)$, calculated by Eq. (3). Column (4) presents the allowed 1σ deviation of the assumed distribution, $\sqrt{\frac{f(x)[1-f(x)]}{N}}$, and Col. (5) presents the deviation of the sample from the distribution, $|f_N(x) - f(x)|$.

From Table 1 we find that there is one point, i.e. $x = 0.035$, where condition (9) is not satisfied. Therefore, the sample does not pass the 1σ distribution function deviation test for the assumed distribution according to definition 1.

5. Conclusions

In this paper, we define and introduce a statistical method, called the 1σ distribution function deviation test, to test assumed distributions of sources. An example of application of the new test is also given. The method is based on the fact that, for an assumed distribution of sources, the variance of the distribution function of samples is known. It is verified that, when the random variables vary continuously, a sample passing the test also passes any other statistical test which depends on the deviation of a statistical function from its expected value at the 1σ confidence level. Specifically, in these cases, a sample passing

Table 1. Data and deviations

x	$f(x)$	$f_N(x)$	$\sqrt{\frac{f(x)[1-f(x)]}{N}}$	$\lvert f_N(x) - f(x)\rvert$
(1)	(2)	(3)	(4)	(5)
0.01	0.01	0.0000	0.0173	0.0100
0.025	0.025	0.0303	0.0272	0.0053
0.035	0.035	0.0909	0.0320	0.0559
0.08	0.08	0.1212	0.0472	0.0412
0.13	0.13	0.1515	0.0585	0.0215
0.185	0.185	0.1818	0.0676	0.0032
0.255	0.255	0.2121	0.0759	0.0429
0.285	0.285	0.2424	0.0786	0.0426
0.295	0.295	0.3030	0.0794	0.0080
0.31	0.31	0.3333	0.0805	0.0233
0.35	0.35	0.3636	0.0830	0.0136
0.39	0.39	0.3939	0.0849	0.0039
0.42	0.42	0.4242	0.0859	0.0042
0.445	0.445	0.4545	0.0865	0.0095
0.465	0.465	0.4848	0.0868	0.0198
0.485	0.485	0.5152	0.0870	0.0302
0.535	0.535	0.5455	0.0868	0.0105
0.605	0.605	0.6061	0.0851	0.0011
0.645	0.645	0.6364	0.0833	0.0086
0.715	0.715	0.6970	0.0786	0.0180
0.78	0.78	0.7273	0.0721	0.0527
0.8	0.8	0.7576	0.0696	0.0424
0.815	0.815	0.8182	0.0676	0.0032
0.86	0.86	0.8485	0.0604	0.0115
0.905	0.905	0.8788	0.0510	0.0262
0.92	0.92	0.9091	0.0472	0.0109
0.935	0.935	0.9394	0.0429	0.0044
0.97	0.97	0.9697	0.0297	0.0003

the test must also pass any mean-value test within the same confidence level. In addition, the sample also passes the Kolmogorov-Smirnov test at the same confidence level. Since the mean value method and the K-S test are usually applied in the study of the distribution of sources in astronomy, we expect that this method will be applicable in testing the distribution of gamma-ray bursts as well as the luminosity function of quasars.

References

Avni Y., Bahcall J.N., 1980, ApJ 235, 694

Boyle B.J., Fong R., Shanks T., Peterson B.A., 1987, MNRAS 227, 717

Boyle B.J., Shanks T., Peterson B.A., 1988, MNRAS 235, 935

Hartwick F.D.A., Schade D., 1990, ARA&A 28, 437

Higdon J.C., Schmidt M., 1990, ApJ 355, 13

Mathez G., 1976, A&A 53, 15

Pei Y.C., 1995, ApJ 438, 623

Schmidt M., 1968, ApJ 151, 393

Warren S.J., Hewett P.C., Osmer P.S., 1994, ApJ 421, 412