EPJ Web of Conferences **226**, 03016 (2020) https://doi.org/10.1051/epjconf/202022603016
Mathematical Modeling and Computational Physics 2019

Nonparametric Tests for Purity of Low Statistics Data

Victor B. Zlokazov[1,⋆]

[1] *Laboratory of Information Technologies, Joint Institute for Nuclear Research, Joliot-Curie 6, Dubna 141980, Russia*

Abstract. The nonparametric methods are most suitable for tasks facing the uncertainty or complexity of models and the small statistics of the analyzed data. They are irreplaceable in those cases when high tempo of the analysis is required.

1 Introduction

Nonparametric methods for the analysis of experimental data are methods, which do not require too detailed parametric description of these data and, what is more important, a priori information about the values of these parameters. A general summary of the benefits of the non-parametric methods can be found in [1–3]. These methods are "short-cut statistics" [4], which means that they are easier for implementations and do not require a lot of input information, work faster and are more reliable, are robust, and insensitive to data statistics.

2 Ratio median/mean as test for purity

Let a distribution $A = t_i$, $i = 1, 2, \ldots, m$ (e.g., data of radioactive decay) be given. We consider the purity as a uniqueness of the distribution function for the given sample. We want to check whether it satisfies, e.g., a single exponential distribution $F(t) = 1 - \exp(-t/T)$, $t \in [0, \infty)$ with unknown parameter T, or a mixture of other distributions, possibly of the same type, but with different values T. We are not interested in the value of T, but if A is "pure", obtaining an estimate of T is a trivial statistical problem.

Thus, we set up a test – build a function $c(t_i)$, which has the following properties:

- its distribution density is significantly different from zero in some region R of variables t_i; the best case is when the mathematical expectation of function $c(t_i)$ is close to a point where this function has a bell shape;
- its density in the case of the single distribution is significantly different from the density function for a mixture of distributions just in the region R.

Here, a function M_m which is the ratio of the sample median m_d to the sample mean m_n is taken as such a test,

$$M_m = \frac{m_d}{m_n} \tag{1}$$

⋆e-mail: zlokazov@jinr.ru

and below we will look at its operation in the two important cases: that of the exponential density

$$f(t,T) = \frac{\exp(-t/T)}{T}, \qquad t \in [0,\infty)$$

and that of the normal one

$$f(t,a,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \cdot \exp\left[-\frac{(t-a)^2}{2\sigma^2}\right],$$

where T and a denote expectations and σ is the standard deviation (square root of the variance).

- The sample mean $m_n = \sum_{i=1}^{m} t_i/m$ is an unbiased and efficient estimate of the parameters T (exponential case) and a (normal one) provided A is "pure" and has the unbiased variance $m_n^2/(m-1)$ (exponential case) and $\sum_{i=1}^{m}(t_i - m_n)^2/(m-1)$ (normal case).
- The sample median m_d is defined according to the following algorithm:
  - arrange all the t_i in increasing order and let j denote the integer part of $m/2$;
  - if m is odd, then $m_d = t_{j+1}$; if m is even, then $m_d = (t_j + t_{j+1})/2$.

  The sample median is asymptotically unbiased and efficient when m grows.

Theoretically our test M_m has a distribution which is similar to the gamma one, getting more symmetric while $m \to \infty$ (in expo case) and similar to the Cauchy distribution (in case of the normal distribution).

In the pure case the expectations of the test M_m are

- (expo): $M_m = \ln(2)$. It follows from $1 - \exp(-m_d/T) = 0.5$ or $\exp(-m_d/T) = 0.5$.
- (normal): $M_m = 1$. It follows from $\mathrm{erf}\,(m_d) = 0.5$ which takes place at $m_d = a$.

3 The distributions of the M_m test

However, in the case of data distributions of a finite size m, the distribution functions for medians have no compact closed form so that we must restrict ourselves to Monte-Carlo simulation of M_m distribution for different m.

In the Figs. 1, 2 the sample distributions for both M_m are plotted on the basis of 10^6 event samples for $m = 30$. In [5] this m is called a good representative of the low data statistics.

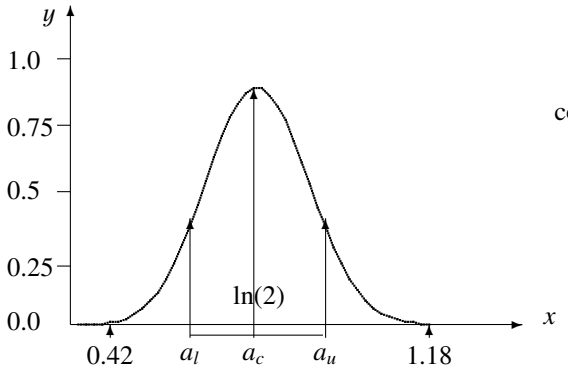
4 Confidence intervals for the M_m test

A test for exact data would point to the expectation $M_m = E$. However, for random data, M_m is random and will be scattered around E over a certain interval $[E-\delta_1, E+\delta_2]$ – the *confidence interval* (CI). Such an interval reflects the degree of our "confidence" to the ability of this test not to reject the right hypothesis and not to accept a false one.

As such a CI we should take the smallest possible interval $[E-\Delta_1, E+\Delta_2]$ covered by the maximum probability integral

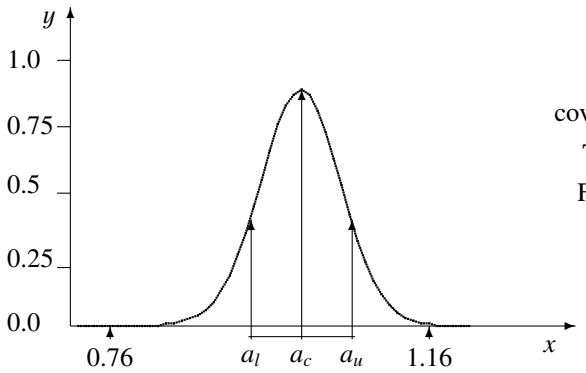
$$\int_{E-\Delta_1}^{E+\Delta_2} f(t)\,dt, \qquad f(t) \text{ is the density.}$$

Unfortunately, in the general case this variational problem has no solution. We must make use of some empirical compromise between the two requirements. In the world practice it is customary to take the 68 % probability integral and the corresponding interval as CI. Smaller values of this integral enhance the chance to commit the first error, the greater to make the second one.



Lines mark the confidence interval $[a_l,\ a_u]$,
covered by the probability integral of about 68 %.
The distribution has the maximum at a_c=ln(2)
For a sample size not larger than 30 events:
$a_l = 0.54$, $a_u = 0.80$.
If $m < 16$, $a_l = 0.46$, $a_u = 0.88$.

Figure 1. The distribution of M_m (expo case). The values along the x-abscissa denote ratios of the sample median to the sample mean; the values along the y-ordinates denote the probabilities of these ratios.



Lines mark the confidence interval $[a_l, a_u]$,
covered by the probability integral of about 68 %.
The distribution has the maximum at a_c=1.0
For the sample size not larger than 30 events:
$a_l = 0.947$ and $a_u = 1.050$.

Figure 2. The distribution of M_m (normal case). The values along the coordinate axes have the same meaning as in Fig. 1.

Given a data sample A, two types of hypotheses testing may be formulated:

- test whether the data does not contradict the hypothesis about its purity;
- test whether the data confirms the hypothesis H_0 about purity against some opposite hypothesis H_A;

In other words we are interested in getting answer to the question: is this data a sample from a single probability density $f(t)$ or (in a simple approach) from the sum of several densities $f_j(t)$ with different weights a_j,

$$\sum_{j=1}^{n} a_j \cdot f_j(t), \qquad j = 1, \ldots, n; \qquad \sum_{j=1}^{n} a_j = 1. \tag{2}$$

If the values t_i are produced by only one term of interest (let it be k-th density) then the data is "pure" and only one weight a_k will be nonzero. Otherwise, the sample is a mixture.

Let $[c_{11}, c_{12}]$ be the confidence interval (of the exponential or normal distribution) for m – the size of data A. Then the first problem is solved calculating M_m and checking whether it falls inside $[c_{11}, c_{12}]$; when it does, the data may be pure at the corresponding significance level.

For the second case we need two hypotheses:

1. H_0: the data is pure;
2. H_A the data possibly comes from a certain mixed density combination.

For each hypothesis the corresponding confidence intervals are built. Let them be $[c_{11}, c_{12}]$ for the hypothesis H_0 and $[c_{21}, c_{22}]$ for the hypothesis H_A. Usually these intervals overlap.

Suppose that on the coordinate axis they can be illustrated as shown in Fig. 3. Then:

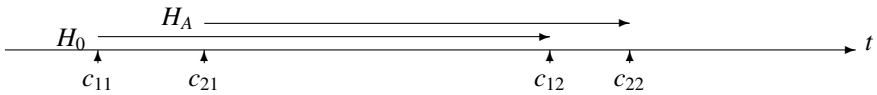


Figure 3. Two overlapping confidence intervals

- If M_m falls inside the range $[c_{11}, c_{21}]$, it means that the data does not contradict H_0 (the data is pure), but contradicts H_A (the data is a mixture); thus, confirms H_0.
- If M_m falls inside the range $[c_{12}, c_{22}]$, it means that the data does not contradict H_A (the data is a mixture), but contradicts H_0 (the data is pure); thus, confirms H_A.
- If M_m falls inside the range $[c_{21}, c_{12}]$, it means that the data does not contradict both H_0 and H_A. It is a failure of the test. One needs the data with a larger statistics to diminish both confidence intervals and, thus, the size of their overlap.
- The case $M_m < c_{11}$ or $M_m > c_{22}$ means also a test failure and one cannot make conclusions for a sample data with given statistics.

5 Conclusions

It has been shown that the proposed nonparametric "median/mean" criterion for testing a rather large class of data distributions for purity is suitable for use. This is especially important for the small data statistics and the lack of the a priori information about the parameters of the distribution function.

References

[1] V. Bagdonavicius, J. Kruopis, M.S. Nikulin, *Non-parametric tests for complete data*, (ISTE & Wiley, London & Hoboken, 2011)
[2] J.D. Gibbons, S. Chakraborti *Nonparametric Statistical Inference* (4th ed., CRC Press, 2003)
[3] M. Hollander, D.A. Wolfe, E. Chicken, *Nonparametric Statistical Methods* (John Wiley & Sons, 2014)
[4] J.E. Freund, *Mathematical Statistics*, (Prentice-Hall, Inc. Englewood Cliffs, N.J., 1962)
[5] S.S. Wilks, *Mathematical Statistics* (Princeton, N.J., Princeton University Press, 1947)