Computer Physics Communications 185 (2014) 3250–3258

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Detection of a small shift in a broad distribution

Bernd A. Berg

Department of Physics, Florida State University, Tallahassee, FL 32306-4350, USA

ARTICLE INFO

Article history:
Received 5 June 2012
Received in revised form
28 August 2014
Accepted 1 September 2014
Available online 15 September 2014

Keywords:
Monte Carlo methods in statistics
Monte Carlo
Statistics
Neutrino departure time distribution

ABSTRACT

Statistical methods for the extraction of a small shift in broad data distributions are examined by means of Monte Carlo simulations. This work was originally motivated by the CERN neutrino beam to Gran Sasso (CNGS) experiment for which the OPERA detector collaboration reported a time shift in a broad distribution with an accuracy of ±7.8 ns, while the fluctuation of the average time turns with ±23.8 ns out to be much larger. Although the physical result of a big shift has been withdrawn, statistical methods that make an identification in a broad distribution with such a small error possible remain of interest.

1. Introduction

In highly publicized CERN announcements [1] it was claimed that neutrinos from the CNGS arrived at Gran Sasso

$$\delta t = \left[-57.8 \pm 7.8\,(\text{stat.})\,{}^{+8.3}_{-5.9}\,(\text{sys.})\right]\,\text{ns} \tag{1}$$

too early, violating the $\delta t = 0$ limit set by the speed of light. Meanwhile, initially overlooked systematic errors [2] have wiped out the estimate of a large shift. But the estimate of the statistical error remains of interest as it exemplifies the extraction of a small shift from a broad distribution. The purpose of this article is to shed light on subtleties of an analysis, which leads to the statistical part of the estimate (1).

The CNGS sample of 15 223 neutrinos was produced in extractions that last about 10 500 ns each. Two different types of extractions were used leading to probability densities (PD)

$$p_k(t), \quad k = 1, 2 \tag{2}$$

for neutrinos departure times, which are reproduced here in Fig. 1. The PD used in our paper have been discretized in intervals of 1 ns and can be downloaded from the author's website [3].

One can now perform a statistical bootstrap [4] analysis by Monte Carlo (MC) generation of departure times with the PD of Fig. 1. This is already remarked in [1], where the application remains limited to testing of their maximum likelihood procedure on a sample of 100 MC data sets. As the MC generation of departure times can be repeated almost arbitrarily often with distinct random numbers, one can analyze and verify statistical methods that one wants to apply to discover a shift in the data.

For the uniform PD over 10 500 ns, it has been noted [5] that with $n = 16\,111$ events the variance of the departure time average $\overline{t}$ is approximately $\triangle\overline{t} = 24$ ns, i.e., much larger than the statistical error bar in Eq. (1). It will be discussed in this paper that the time shift δt (1) defined in [1] behaves indeed differently than a statistical fluctuation of the time average

$$\delta\overline{t} = \frac{n_1\,\delta\overline{t}_1 + n_2\,\delta\overline{t}_2}{n_1 + n_2}, \qquad \delta\overline{t}_i = \overline{t}_i - \widehat{t}_i, \quad i = 1, 2. \tag{3}$$

Here $\overline{t}_i$ are the measured departure time averages, $\widehat{t}_i$ are the mean departure times obtained from the underlying PD, and n_i are the numbers of events in each extraction. The distinction between δt (1) and $\delta\overline{t}$ (3) is made by an overline on t or not. Obviously,

$$\langle\delta t\rangle = \langle\delta\overline{t}\rangle \tag{4}$$

holds for the expectation values, but their error bars behave differently.

To set the groundwork, it is shown in Section 2 for the uniform distribution that a shift $\delta t = -57.8$ ns can be identified with certainty (probability to miss it $< 10^{-36}$) when there are 15 223 events and the departure time range is 10 500 ns. In Section 3 the MC generation of departure times is described. Section 4 gives examples of descriptive histograms from MC data. Suggested by the uniform distribution, the front tails of the distributions are of particular interest. For their study the cumulative distribution function (CDF) is better suited than a histogram, because it allows easily to focus on outliers. This is investigated in Section 5. To estimate the shift value

E-mail address: berg@hep.fsu.edu.

http://dx.doi.org/10.1016/j.cpc.2014.09.001
0010-4655/

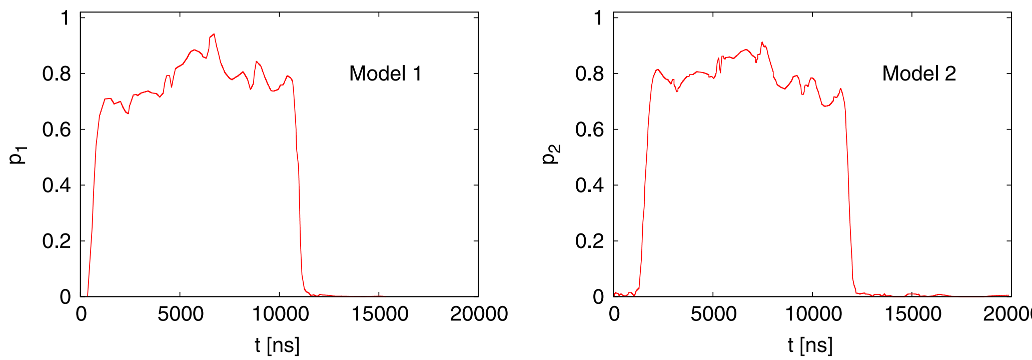

Fig. 1. Departure time probability densities modeled after Fig. 11 of Ref. [1].

δt, the maximum likelihood method is used in [1]. In Section 6 features of this method are calculated by applying it to a large number of MC generated departure time samples.

Independently of the special example, the approaches discussed in Sections 2–6 are of interest, because they address the general problem of extracting a precise estimate of a shift from a broad distribution. Summary and conclusions follow in Section 7.

2. Uniform distribution

Using the uniform PD over a time window of 10 500 ns, the standard deviation of the average

$$\bar{t} = \frac{1}{n}\sum_{j=1}^{n} t^j \tag{5}$$

is for $n = 15\,223$ events much larger than the statistical error bar quoted in Eq. (1), namely approximately

$$\triangle\bar{t} = 25 \text{ ns}. \tag{6}$$

How can this be? That the average (5) fluctuates with the variance (6) is unavoidable. However, the effect we are after is a systematic shift of each departure time by an amount $\delta t = -57.8$ ns. Again for the uniform distribution, drawn in Fig. 2, it is easily illustrated that this can very well be identified. Events indicated on the left of the figure are impossible unless there is a shift. Now, with a shift of -57.8 ns the probability to find a particular event to the left of the uniform PD is given by

$$p = 57.8/10\,500 = 0.005505\ldots \tag{7}$$

and the probability to find none is

$$(1-p)^{15\,672} = 10^{-36.5}. \tag{8}$$

The distance of the smallest time from the left edge of the uniform PD is a lower bound on δt and a direct estimate for the time shift (1) is

$$\delta t = n_{\text{left}}\; 10\,500 \text{ ns}/15\,223, \tag{9}$$

where n_{left} is the number of events observed on the left outside of the uniform PD. Confidence limits can be established from the binomial distribution.

For the tiny range of 57.8 ns indicated by the somewhat thicker line on the right side of the uniform PD in Fig. 2, the situation is the other way round. It has to be empty when there is a shift by $\delta t = -57.8$ ns. The probability that this happens by chance when there is in fact no shift is also given by (9). When δt is not known the distance of the largest measured time from the right edge of the uniform PD is an upper bound on δt.

We do not pursue the uniform PD any further, because we are interested in the more complicated case of the less sharp PD of Fig. 1.

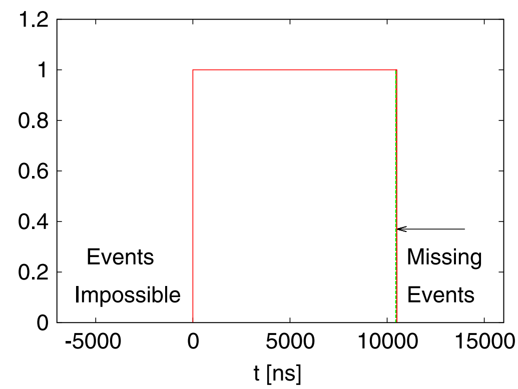

Fig. 2. Uniform distribution: Impossible (left) and missing events (in the enlarged thickness of the right border).

3. MC generation of departure times

As mentioned in the introduction, the PD of Fig. 1 have been discretized in 1 ns intervals and are available on the Web [3]. The resolution of 1 ns allows for easy MC generation of departure times and is sufficient for the intended accuracy of the estimate of a shift. In the following our thus defined models are labeled by $k = 1, 2$. The probabilities as function of time t are defined by

$$p_k(t) = p_k(i^t) \quad \text{for } i^t \le t < i^t + 1, \tag{10}$$

where i^t are integers in ns units. As it is convenient for the MC generation of departure times, the normalization for the discretized PD is (distinct from Fig. 1) chosen so that

$$p_k^{\max} = \max_i\,[p_k(i)] = 1 \tag{11}$$

holds. Proper normalizations $\sum_i p_k(i)\,\triangle t_k = 1$ could still be achieved by choosing instead of ns some unconventional unit for $\triangle t_k$. For the generation of correctly distributed random times this is irrelevant. For a short time range model 1 probabilities $p_1(t)$ are enlarged in Fig. 3.

After discretization the smallest $i_k^{\min}$ and largest $i_k^{\max}$ times with non-zero $p_k(i^t)$ values are

$$i_1^{\min} = 359 \quad \text{and} \quad i_1^{\max} = 15\,368,$$
$$i_2^{\min} = 12 \quad \text{and} \quad i_2^{\max} = 19\,877.$$

In particular for the large i^t values, these ranges include a number of zero probabilities. MC generated departure times t_k^j ($j = 1, \ldots, n$) with n the number of data are obtained from uniformly distributed random numbers t in a range enclosing $\left(i_k^{\min}, i_k^{\max}+1\right)$, here chosen to be $(1, 22\,000)$: for model k a proposed random time t is accepted with probability $p_k(i^t)$ for $i^t \le t < i^t + 1$. If t is rejected, the procedure is repeated until a value gets accepted,