Computer Physics Communications 200 (2016) 168–175

Contents lists available at ScienceDirect

Computer Physics Communications

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

Breakdown of statistical inference from some random experiments

Marian Kupczynski [a], Hans De Raedt [b,*]

[a] Département de l'Informatique, Université du Québec en Outaouais (UQO), Case postale 1250, succursale Hull, Gatineau. Quebec, J8X 3X 7, Canada

[b] Zernike Institute for Advanced Materials, University of Groningen, Nijenborgh 4, NL-9747 AG Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 16 September 2015
Received in revised form
20 November 2015
Accepted 21 November 2015
Available online 27 November 2015

Keywords:
Finite statistics
Statistical inference
Monte Carlo simulations
Internal protocols
Breakdown of significance tests
Sample inhomogeneity
Chi-square compatibility tests
Chi-square frequency histograms

ABSTRACT

Many experiments can be interpreted in terms of random processes operating according to some internal protocols. When experiments are costly or cannot be repeated only one or a few finite samples are available. In this paper we study data generated by pseudo-random computer experiments operating according to particular internal protocols. We show that the standard statistical analysis performed on a sample, containing 10^5 data points or more, may sometimes be highly misleading and statistical errors largely underestimated. Our results confirm in a dramatic way the dangers of standard asymptotic statistical inference if a sample is not homogeneous. We demonstrate that analyzing various subdivisions of samples by multiple chi-square tests and chi-square frequency graphs is very effective in detecting sample inhomogeneity. Therefore to assure correctness of the statistical inference the above mentioned chi-square tests and other non-parametric sample homogeneity tests should be incorporated in any statistical analysis of experimental data. If such tests are not performed the reported conclusions and estimates of the errors cannot be trusted.

1. Introduction

Outcomes of experiments or surveys in various domains of science are usually interpreted as observed values of one or more random variables obeying some, in general, multivariate probability distribution. Gathered data are often assumed to be simple random samples. A random sample is simple if it is homogeneous and all trials are independent.

The dangers of statistical inference based on finite samples are well known to statisticians but many experimentalists seem to be unaware of them. Let us cite here [1]: “incorrect assumptions of ‘simple’ random sampling can invalidate statistical inference”.

Computer packages for statistical analysis produce descriptive statistics and outcomes of various significance tests. However these packages cannot replace statistical thinking and mistaken conclusions are often drawn in a variety of studies because the researchers do not appreciate the significance of the assumptions about the probability distribution underlying a model and for other reasons [2].

Particular caution is needed in the case where only one large sample of data is available and we want to make a sound statistical inference based on it, as in for example, the data obtained in the experiments of Christensen et al. [3] and Giustina et al. [4]. One may not simply assume that the experimental data are ‘simple’ random samples without verifying it.

Many experimentalists believe, when a sample size is 10^4 or larger, that a sample average and a sample mean error give reliable information about studied statistical population even if a studied sample is not a perfect *simple random sample*. In this paper we show that such belief is unjustified and a careful study of sample homogeneity is always necessary. Some experimental devices, operating according to specific internal protocols may produce strange, but legitimate, outcomes which usually would be considered as outliers and rejected.

In order to explore possible anomalies in large finite samples, we study several pseudo-random computer experiments generating time series of discrete data according to different internal protocols. We use the term “internal” to indicate that the details of the protocol are inaccessible to any person analyzing the data, as in real-life applications. We demonstrate that standard statistical inference of one or a few of large samples (containing as much as 10^7 data points) generated by some of these protocols in terms of standard errors and various confidence intervals can be highly misleading.

[*] Corresponding author.
E-mail address: h.a.de.raedt@rug.nl (H. De Raedt).

http://dx.doi.org/10.1016/j.cpc.2015.11.010

By subdividing our samples into 100 bins (each bin containing 10^5 data items) and by performing 4950 chi-square bin-to-bin compatibility tests we demonstrate that samples produced by some of our computer experiments are not homogeneous, explaining the invalid conclusion based on the performed significance tests. For the samples which are homogeneous we obtain close to perfect agreement with a corresponding probabilistic model.

In our paper we not only demonstrate the dramatic consequences of sample inhomogeneity but we suggest which preliminary supplementary statistical tests of the data should be performed in order to assure a sound statistical inference. These tests detected the anomalies in our computer generated samples without making use of any knowledge about a particular protocol.

2. Standard statistical inferences

Let us assume that A can take k different values: $a_1, a_2, \ldots, a_k$. In a long run of the experiment we obtain a random sample $S = \{x_1, x_2, \ldots, x_N\}$ of size N which according to standard sampling methods is interpreted as an observation of a multivariate random variable $\{A_1, A_2, \ldots, A_N\}$ where A_i are independent and identically distributed random variables (i.i.d.): $A_i \sim D$.

The empirical frequency distributions of various outcomes $f_i = \#(x_j = a_i)/N$ are found and believed to approach the probabilities provided by the theory. Furthermore the probability distribution of the variable $\bar{A} = \frac{1}{N}\sum_{i=1}^{N} A_i$ is, due to the central limit theorem (CLT), believed to be well approximated by a normal distribution $N(\mu_{\bar{A}}, \sigma_{\bar{A}}^2)$ with $\mu_{\bar{A}} = \mu_A$ and $\sigma_{\bar{A}} = \sigma_A/\sqrt{N}$ where μ_A and σ_A are the mean and the standard deviation of the random variable A. In spite of the fact that CLT is valid when N tends to infinity it is often assumed that already for $N \geq 30$ the normal distribution provides a reasonable approximation and that the unknown variance σ_A^2 can be replaced by a value of its unbiased estimator s^2:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}. \tag{1}$$

A sample mean $\bar{x}$ is considered as a good estimate of $\langle A \rangle = \mu_A$ and a standard error of the mean SEM $= s/\sqrt{N}$ as a good estimate of $\sigma_{\bar{A}} = \sigma_A/\sqrt{N}$.

As N increases the confidence in the validity of the approximation by a normal distribution is increasing and the errors become smaller and smaller. The most exact probabilistic statement, if the normal-distribution approximation is valid, can be expressed in terms of the confidence intervals I_α:

$$I_\alpha = \left[\bar{x} - z_{\alpha/2}s/\sqrt{N}, \bar{x} + z_{\alpha/2}s/\sqrt{N}\right] \tag{2}$$

saying that the probability that the interval I_α covers the unknown value μ_A is $(1-\alpha)$.

If the asymptotic normality of the distribution is not assumed, the Chebyshev's inequality can be used and the confidence interval (2) is replaced by

$$I_c = \left[\bar{x} - cs/\sqrt{N}, \bar{x} + cs/\sqrt{N}\right] \tag{3}$$

and the probability that the interval I_c covers the unknown value μ_A is $(1 - \frac{1}{c^2})$.

Of course the estimation of SEM $= s/\sqrt{N}$ is valid if the variables A_i are independent and identically distributed random variables (i.i.d.): $A_i \sim D$. But this has to be carefully checked and not taken for granted. The Chebychev's inequality for the finite samples is valid under the supplementary assumptions [5,6].

3. Experiments and invisible internal protocols

Let us imagine a following experiment. A signal is entering a measuring device (considered to be a black box) and from time to time some discrete outcomes are produced and a sample S is obtained. If the outcomes seem to be randomly distributed we could assume a following probabilistic model:

- a signal is described by a probability distribution $p_1(m)$
- a state of the device at the moment of a measurement is described by a probability distribution $p_2(n)$
- the output of the device is one of discrete values $A(m, n)$.

If this simple probabilistic model is assumed then the expectation value:

$$\langle A \rangle = \sum_{m,n} A(m, n)p_1(m)p_2(n). \tag{4}$$

The probability distribution $p(A(m, n) = a)$ and the standard deviation σ_A are easily found and compared with experimental data.

As we mentioned above we do not know how our device produces successive outcomes. Therefore we perform several Monte Carlo simulations using various possible internal protocols and we compare the properties of the finite samples generated by these protocols.

We define three different protocols:

Protocol $1 = (N_1, 1, m)$:

- generate one value of n and one value of m using $p_1(m)$ and $p_2(n)$
- evaluate $A(m, n)$ and output this value
- repeat the process N_1 times in order to create a sample of size $N = N_1$.