Measurement 65 (2015) 94–111

Contents lists available at ScienceDirect

Measurement

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

Review

A qualitative study of probability density visualization techniques in measurements

Lee Gonzales-Fuentes [a,*], Kurt Barbé [a], Lee Barford [b], Lieve Lauwers [a], Lieven Philips [b]

[a] Department of Mathematics (DWIS), Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium
[b] Measurement Research Laboratory, Keysight Technologies, 561 Keystone Ave. MS 434, Reno, 89503 NV, USA

ARTICLE INFO

Article history:
Received 24 February 2014
Received in revised form 4 November 2014
Accepted 11 December 2014
Available online 7 January 2015

Keywords:
Density estimation
Engineer
Histogram
Kernel density estimation
Measurement instrument
Nonparametric
Orthogonal series estimation
Polynomial function
Probability density function
Uncertainty characterization

ABSTRACT

Engineers find interpreting plots of a measured physical variable more straightforward than doing a formal statistical analysis. The default choice to display the data behavior is the histogram. The histogram's performance has proved to be sufficient. However, histograms have a number of limitations including sensitivity to the binwidth and a non-physical roughness. Over the past years, statisticians have developed different techniques to address these problems. These techniques provide a much clearer visualization of the probability density and a more accurate estimation of the statistical properties of the measured data. Despite their increasing use in other fields, these techniques are rarely used in the measurement community. For instance, most measurement instruments provide histograms only. This review article revisits these techniques from an engineer viewpoint to encourage its use. Different examples that include known and unknown densities result in practical guidelines that help the measurement engineer to visualize the probability content.

Contents

* Corresponding author. Tel.: +32 2 629 29 53; fax: +32 2 629 28 50.
E-mail address: legonzal@vub.ac.be (L. Gonzales-Fuentes).

http://dx.doi.org/10.1016/j.measurement.2014.12.022
0263-2241/

Table 1
Evolution of the statistical density estimation techniques.

Year	Author	Contribution
1632	Galileo	Histogram-like diagrams [6]
1662	Graunt	Binned frequency curve [9,10]
1891	Pearson	Data representation method using columns [6]
1894	Pearson	'Histogram' term [11]
1932	Fisher	Histogram as statistical method for research
1951	Fix and Hodges	Nonparametric density estimation [12]
1956	Rosenblatt	Weight function for nonparametric density estimation [13]
1962	Parzen	Statistical analysis of Parzen's method [14]
	Cencov	Introduction of orthogonal Series for density estimation [4]
1964	Cacoullos	Parzen's method for multivariate data
1966	Cacoullos	'Kernel function' term [4]
1965	Loftsgaarden and Quesenberry	Kernel density estimation for multivariate data [15]
1969	Watson	Density estimation by orthogonal series [4]

1. Introduction

Many problems in the scientific and engineering fields are inherent to variability in real data or measurement uncertainty. Discovering and understanding the sources of this randomness can help to make any scientific process more robust. Hence, describing this randomness with only fixed numbers or simple plots can sometimes be inappropriate and uninformative. Statisticians developed tools to quantify and visualize this randomness in a probabilistic manner so that more information on the behavior of the data can be obtained. Over the past years, advances on computer science have revolutionized this practice. This allowed most experimental researchers and engineers interpreting graphs rather than doing formal statistical analysis [1,2].

The probability density function characterizes the behavior of a particular random variable by describing the values that the variable can take on in a specific interval. Density estimation is the process of building an approximation $\hat{f}(x)$ of the unobservable probability density $f(x)$ just using n observations of X i.e. $X_1, \ldots, X_n$, where X is a continuous random variable [2,3]. A continuous probability density function f satisfies [4]

$$(f(x) \geqslant 0, \quad \int_R f(x)dx = 1, \tag{1}$$

and so should its density estimate $\hat{f}$. A perfect knowledge of f would allow some statistical properties and applications to be determined [4–6]. The choice of the estimation method depends on the engineer's objective. For instance, if some specific derivations are required, the statistical technique should be *parametric*: a specific density function (Normal, Exponential, Poisson, Uniform, etc.) should be assumed. If the type of the underlying density function, presence of tails or skewness needs to be determined, the methodology should be *nonparametric*: no specific density function is assumed. Parametric methods are known for being powerful for unimodal distributions [7]. However, in practice most densities are multimodal and high-dimensional [8].

The histogram is the most used nonparametric density estimation technique. It appeared in the 17th century as a tool that grouped and tabulated data into bins which form a frequency curve and became an accepted statistical data representation method by the 20th century, see Table 1. Since then, the histogram has remained as an important visualization tool [9]. Nowadays, many software packages and measurement instruments generate histograms to estimate the distribution of repeated measurements. In order to analyze if this technique does a proper job, we will assess its performance on a known density function.

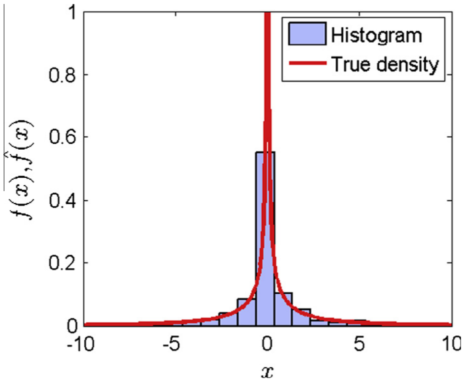

Fig. 1. Probability density function of cubic power of a Gaussian distributed random variable X.

An interesting example is the cubic power of a Gaussian distributed random variable X, which presents a symmetrical and highly-peaked distribution. The true probability density function of X^3 is

$$f_{X^3}(x) = \frac{1}{3|x|^{2/3}} f_X\left(|x|^{1/3}\right)$$

where $f_X(x)$ represents the true density function of the Gaussian random variable X. Using a data set of size $n = 1000$, the true density is computed and shown as a red[1] curve in Fig. 1. The histogram is plotted when the data is split into 63 bins, represented as blue bars, which results in a binwidth $h = 0.99$. The plot shows the histogram in the interval $[-10, 10]$.

Does the histogram resemble the true density? Not fully. The shape of the histogram is mainly ragged, despite

[1] For interpretation of color in Figs. 1 and 3, the reader is referred to the web version of this article.