Histogram, an Ancient Tool and the Art of Forecasting

Katta G. Murty
Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109-2117, USA
Phone: 734-763-3513, fax: 734-764-3451
e-mail:katta_murty@umich.edu

March 2002

Abstract

One of the motivating factors behind the development of the theories of probability and statistics is to obtain good forecasts for future values of random variables that appear in many applications. Managers in industry and business depend heavily on good forecasts of product demand. History shows many examples of firms benefiting from accurate forecasts and paying the price for poor forecasting, and we see many managers complaining about the quality of the forecasts they are getting.

All existing forcasting techniques are based on developing a point estimate of the expected value. Questions about the updated distribution of the random variable under study are rarely raised, but if it arises, it is commonly assumed to be the normal distribution with the forecasted value as the expectation and most often the same variance as before.

In this article, we propose a simple but different technique for actually updating the whole distribution of a random variable, not just its expected value; based on the ancient tool of the histogram. This technique is likely to be useful in situations involving random variables whose distributions undergo frequent changes, for which the normality assumption may be inappropriate, and there is adequate data on the values of the random variable.

Key words Random variables, decision making, probability distribution, histogram, expected value, forecasting, normal distribution, exponential smoothing.

1 Introduction

Forecasting has always been an extremely important activity. The theories of probability and statistics were developed mainly to provide scientific tools for forecasting, and now forecasting is a very important chapter in statistical theory and it is a big business. All planning in industry, business, and governments is based on some type of forecasts.

Forecasting is characterizing the future value of a random variable (RV). The probability distributions of many RVs encountered in industry, business and government are subject to changes over time, that's why forecasting becomes necessary. It is well recognized that the value of an RV follows a probability distribution, so the fundamental goal of forecasting is actually to update its probability distribution based on information contained in its present and past values. However, most people who depend on and use forecasts are not well versed in the subtle meanings of statistical terminology, so they misunderstand the purpose of forecasting to be that of generating a single numerical value. This misunderstanding is also supported by existing forecasting methods, because all of them only provide an estimate of the expectation of the RV in the next period [1, 3]. This misunderstanding is the main reason behind the frequent complaints by decision makers that the forecasts they are getting are unreliable.

Forecasting methods implicitly assume that the user knows the functional form of the probability distribution of the RV under study, and that the only change that may take place in the next period is a change in its expected value. Their strategy amounts to instructing the user to treat the updated probability distribution of the RV to be the one with the new expected value substituted in its known functional form; and use this updated probability distribution in making any planning decisions. This strategy is useful only when changes in the probability distribution of the RV can be captured by a single parameter, the expectation. These methods seem inadequate to capture all the dynamic changes occurring in the shapes of probability distributions of RVs in a variety of applications.

A better strategy is to actually update the complete probability distribution of the RV and present that as the forecast. This eliminates errors due to normality assumptions, and also forces the users to adopt the whole updated probability distribution in their calculation for decision making rather than the expected value only. In this article we present such a method for updating the whole probability distribution of RVs.

2 The Method

Empirical Distributions and Probability Density Functions

The concept of the probability distribution of an RV evolved from the ancient practice of drawing histograms for the observed values of the RV. The observed range of variation of the RV is usually divided into a convenient number of value intervals (in practice about 10 to 25) of equal length, and the relative frequency of each interval is defined to be the proportion of observed values of the RV that lie in that interval. The chart obtained by marking the value intervals on the horizontal axis, and erecting a rectangle on each interval with its height along the vertical axis equal to the relative frequency is known as the relative frequency histogram of the RV, or its discretized probability distribution. The relative frequency in each value interval I_i is the estimate of the probability p_i that the RV lies in that interval, see Figure 1.

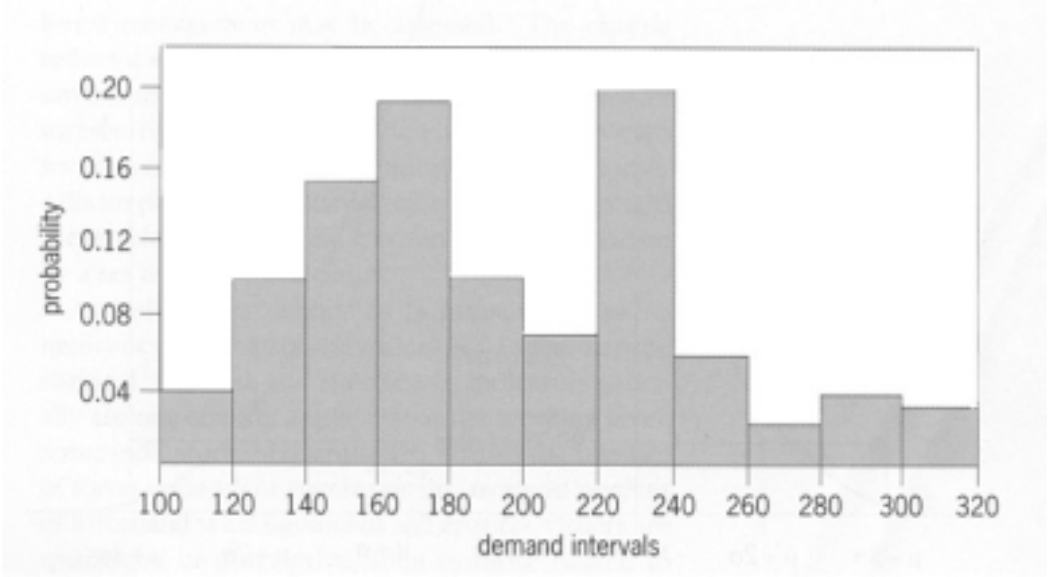

Figure 1: Relative frequency histogram for daily demand for a major component at a PC assembling plant in California.

Let $I_1, \ldots, I_n$ be the value intervals with $u_1, \ldots, u_n$ as their midpoints, and $p = (p_1, \ldots, p_n)$, the probability vector in the empirical distribution of the RV. Let

$$\bar{\mu} = \sum_{i=1}^{n} u_i p_i, \qquad \bar{\sigma} = \sqrt{\sum_{i=1}^{n} p_i (u_i - \mu)^2}$$

Then $\bar{\mu}, \bar{\sigma}$ are estimates of the expected value μ, standard deviation σ of the RV respectively.

We will use the phrase empirical distribution to denote such a discretized probability distribution of an RV, obtained either through drawing the histogram, or by updating a previously known discretized probability distribution based on recent data.

When mathematicians began studying RVs from the 16th century onwards, they found it convenient to represent the probability distribution of the RV by the probability density function which is the mathematical formula for the curve defined by the upper boundary of the relative frequency histogram in the limit as the value interval length is made to approach 0, and the number of observed values of the RV goes to infinity. So the probability density function provides a mathmatical formula for the height along the vertical axis of this curve as a function of the variable corresponding to the horizontal axis. Because it is a mathematically stated function, the probability density function lends itself much more nicely into mathematical derivations than the somewhat crude relative frequency histogram.

In course of time it has become a common practice to assume that the probability distributions of most RVs encountered in applications can be approximated by a particular probability density function called the normal distribution, see Figure 2.

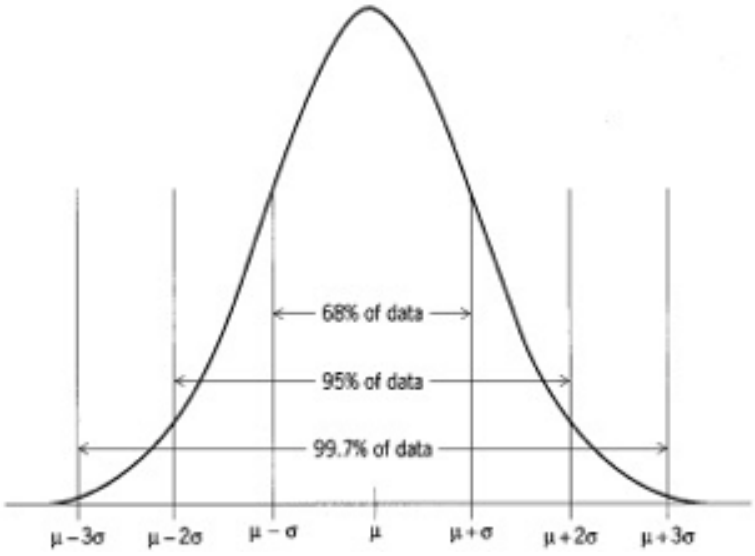

Figure 2: Normal distribution with mean μ and standard deviation σ. The interval $\mu \pm 3\sigma$ is associated with a probability of 0.997 in it.

The normal distribution is completely specified by two parameters, the mean μ and the standard deviation σ. It is perfectly symmetric around the mean, and as shown in Figure 2, the probabilities corresponding to the intervals $[\mu-\sigma, \mu+\sigma], [\mu-2\sigma, \mu+2\sigma], [\mu-3\sigma, \mu+3\sigma]$ are 0.68, 0.95. 0.997 respectively.

Now a days the most commonly used probability distribution in models for decision making is the normal distribution, even though histograms for the RVs in some applications indicate that their distributions are not symmetric around the mean. Of course in some areas other distributions are used, for example the Weibull distribution in reliability studies, etc.

One of the theoretical advantages that the normality assumption confers is that when the probability distribution of the RV changes, one has to change only the values of the mean and the standard deviation in the models. In practice, almost always it is only the value of the mean that is changed; the standard deviation is usually assumed to remain unchanged.

Why Update the Probability Distribution?

The probability distributions of RVs may change with time. For example in supply chain management, the important RVs are daily or weekly demands of various items (raw materials, components, finished goods etc.) that companies either buy from suppliers are sell to their customers. The important factor in this area today is the highly competitive environment and the rapid rate of technological change that is shortening product life cycles, and causing probability distributions of demand variables to change frequently. The results obtained and the decisions reached do indeed depend on the probability distributions used in the decision making model, and unles changes occurring are captured in the models, the conclusions obtained by them will not be accurate. That's why we need to periodically update the probability distributions of the RVs based on recent data.

It is rare to see empirical distributions used in decision making models these days. Almost everyone uses mathematically defined density functions characterized by a small number of parameters (typically two or less) to represent probability distributions. In these decision making models, the only freedom we have in incorporating changes is to change the values of those parameters. This may be inadequate to capture all dynamic changes occurring in the shapes of probability distributions from time to time.

Representing Probability Distributions by the Empirical Distributions Makes All Changes Possible

We will now see that representing the probability distributions of RVs by

their empirical distributions gives us unlimited freedom in making any type of change including changes in shape.

Let $I_1, \ldots, I_n$ be the value intervals, and $p_1, \ldots, p_n$ the probabilities associated with them in the present empirical distribution of an RV. In updating this distribution, we have the freedom to change the values of all the p_i, this makes it possible to capture any change in the shape of the distribution.

Changes, if any, will reflect in recent observations on the RV. Following table gives the present empirical distribution, histogram based on most recent observations on the RV (for example most recent k observations where k could be about 50), and the algebraic symbols representing the probabilities in the updated empirical distribution.

Value interval	Probability vector in the		
	Present empirical distribution	Recent histogram	Updated empirical distribution
I_1	p_1	f_1	x_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_n	p_n	f_n	x_n

$f = (f_1, \ldots, f_n)$ represents the estimate of the probability vector in the recent histogram, but it is based on too few observations. $p = (p_1, \ldots, p_n)$ is the probability vector in the empirical distribution at the previous updating. $x = (x_1, \ldots, x_n)$, the updated probability vector, should be obtained by incorporating the changing trend reflected in f into p. The weighted least squares method provides the following model [2] to compute x from p and f.

$$\begin{aligned} \text{minimize} & \beta \sum_{i=1}^{n} (p_i - x_i)^2 + (1-\beta) \sum_{i=1}^{n} (f_i - x_i)^2 \\ \text{subject to} & \sum_{i=1}^{n} x_i = 1 \\ & x_i \geq 0, \quad i = 1, \ldots, n \end{aligned} \tag{1}$$

where β is a weight between 0 and 1. x is taken as the optimum solution of this convex quadratic program. $\beta = 0.8$ or 0.9 works well, the reason for choosing the weight for the second term in the objective function to be small is because the vector f is based on only a small number of observations. Since the quadratic model minimizes the weighted sum of squared forecast errors over all value intervals, when used periodically, it has the effect of tracking gradual changes in the probability distribution of the RV.

The above quadratic program has a unique optimum solution given by the following explicit formula.

$$x = \beta p + (1 - \beta) f \tag{2}$$

So we take the updated empirical distribution to be the one with the probability vector given by (2).

The formula (2) for updating the probability vector in the above formula is exactly analogous to the formula for forcasting the expected value of an RV using the latest observation in exponential smoothing [3]. Hence the above formula can be thought of as the extension of the exponential smoothing method to update the probability vector in the empirical distribution of an RV .

When there is a significant increase or decrease in the mean value of the RV, new value intervals may have to be opened up at the left or right end. In this case the probabilities associated with value intervals at the other end may become very close to 0, and these intervals may have to be dropped from further consideration at that time.

3 Conclusion

We have shown that representing the probability distribution of an RV by ite empirical distribution lends itself very nicely to updating its complete probability distribution through an exponential smoothing like formula. This updating formula captures all the changes occurring in the probability distribution.

Forecasting is an essential ingradient for scientific decision making. In existing theory forecasting is based on the assumption that the RV follows the probability distribution defined by a mathematical density function that depends on only one parameter, the expectation; and updating the expectation periodically. This strategy is fine if there is knowledge about the RV that justifies this assumption as reasonable, or if there is not enough data to draw a histogram for the RV. Otherwise, the strategy of forecasting the empirical distribution of the RV based on periodic use of the exponential smoothing updating formula (2) described above is a much better alternative.

This shows that the histogram, an ancient tool that gets no respect today, is as important as all the mathematical density functions in statistics literature.

4 Acknowledgement

I considered both models (1), (2) for updating the probability vector in the empirical distribution, but thought that the results from (1) would be better. I thank Andrew Lim for pointing out that both (1) and (2) are the same.

5 References

[1] G. E. P. Box and G. M. Jenkins, Time Series Analysis, Forecasting and Control, Holden Day, San Francisco, 1970.
[2] K. G. Murty, "Supply Chain Management in the Computer Industry", Technical report, IOE Dept., University of Michigan, Ann Arbor, 1999.
[3] S. Nahmias, Production and Operations Analysis, Irwin, Boston, 1993.