The Power Binomial Distribution: a flexible (two-parameter) finite probability distribution

N.I. Fisher
School of Mathematics & Statistics F07
University of Sydney, NSW 2006, Australia
&
ValueMetrics Australia
nif@valuemetrics.com.au

March 2002

Abstract

This note presents a two-parameter generalisation of the Binomial distribution that provides flexibility in simulating discrete finite distributions.

The two parameters may be interpreted directly as location and dispersion, making the Power Binomial family suitable for simulating the sorts of samples that arise in certain forms of customer satisfaction surveys.

KEY WORDS: Discrete distribution; Power Binomial; Customer Value survey

1 Introduction

Statistical data arising from surveys of customer satisfaction or staff satisfaction are often recorded on a simple discrete scale. For example, the methodology of Customer Value Analysis (e.g. Gale [2]; Kordupleski *et al.* [4]; Laitamäki & Kordupleski [5]) recommends that *Satisfaction* be recorded in the range 1, 2, ..., 10, where 1 = *Poor* and 10 = *Excellent*. It is desirable to have a simple flexible statistical distribution that can serve as a plausible model for such data, particularly in the development and evaluation of new methods.

For a given market segment, Customer Value data tend to be unimodal, so one might hope that the (one-parameter) Binomial $B(m, p)$ distribution would be suitable. However, it is not sufficiently flexible to reflect the sorts of skewness typically observed in practice. Clark, Cleveland, Denby & Liu [1] reported the results of a major study of Customer Value survey data. They noted that the mode of the data in homogeneous groups appeared to be around 7 or 8, resulting in data distributions skewed to the left.

The purpose of this note is to present a generalised version of the Binomial distribution that provides such flexibility, and with parameters having a direct interpretation in terms of location and dispersion. For this reason, the proposed class of distributions is quite different from the Katz family of distributions (*e.g.* Johnson *et al.* [3]) and from the Beta-Binomial distribution, and make it more amenable to simulating plausible customer satisfaction data, the primary focus of this note. On the other hand, whilst the link between parameters and moments is somewhat opaque for the Beta-Binomial distribution, that distribution is to be preferred for purposes of statistical modelling and inference (*e.g.* Griffiths 1973); data-based computations and inference are tedious with the Power Binomial distribution.

In its basic form, the Power Binomial family is indexed by two parameters; in some applications a simply-derived three-parameter version might also be of use .

2 Definition and properties

Let X be a random variable with a Binomial $B(m,p)$ distribution, where $0 < p < 1$ and m is a positive integer. Define

$$a_r \equiv \text{Prob}(X = r) = \tbinom{m}{r} p^r (1-p)^{m-r}, r = 0, ..., m,$$

and, for real-valued α, set

$$\pi_r = a_r^\alpha / A(m, \alpha)$$

where

$$A(m, \alpha) = \sum_0^m a_r^\alpha$$

Let Y be a random variable such that $Prob(Y = r) = \pi_r, r = 0, ..., m$. Then Y will be said to follow the Power Binomial $PB(m, p, \alpha)$ distribution. When $\alpha = 0$ the distribution is just the discrete uniform distribution on $\{0, ..., m\}$.

The variety of shapes afforded by the extra parameter is exhibited in Figures 1 and 2.

Figure 1 shows the sorts of shapes that arise when α is positive, for two different values of the range variable m. Figure 1a shows results for the Customer Survey value $m = 9$ (which yields a distribution concentrated on 10 points), with $0.2 \leq \alpha \leq 3$ and $0.05 \leq p \leq 0.95$. The distributional shapes for $p < \frac{1}{2}$ are, of course, the mirror image of those for $p > \frac{1}{2}$; however, the partial duplication is useful for studying larger values of m. The distributions are all unimodal, with some of them not atypical of the survey data they are intended to model. The basic Binomial parameter p is the primary location parameter, and α moderates the dispersion of the distribution. The value $\alpha = 1$ corresponds to the Binomial distribution. Crudely speaking, the mean of the distribution will be approximately p (this being exactly true for $\alpha = 1$), more concentrated for $\alpha > 1$ and less for $\alpha < 1$, Figure 1b exhibits the results for a rather larger range, $m = 29$, together with values of α and p that show the possibilities here.

When $\alpha < 0$, the resulting distributions are mainly U–shaped. Figure 2 shows examples for the same values of m, 9 and 29.

In summary, except for J–shaped distributions that occur for p near 0 or 1, the shape of a Power Binomial distribution for positive dispersion parameter is unimodal, with the probability function falling away on each side of the modal value. Models with the desired skewness to the left are provided by values of $p > \frac{1}{2}$ and $\alpha > 0$.

3 Model fitting

Estimates of α and p are not available in closed form. However, approximate starting values for an iterative fitting procedure such as Minimum Chi-squared can be obtained by using the fact that

$$\frac{\pi_r}{\pi_{r-1}} = (\frac{m-r+1}{r}\frac{p}{1-p})^{\alpha},\ r = 1, ..., m,$$

so that

$$\frac{\pi_r}{\pi_{r-1}}/\frac{\pi_{r-1}}{\pi_{r-2}} = (\frac{m-r+1}{r}\frac{r-1}{m-r+2})^{\alpha}$$

Thus, for any $r = 1, ..., m$,

$$\alpha = \frac{\ln(\pi_r\pi_{r-2}/\pi_{r-1}^2)}{\ln\{m-r+1)(r-1)/[r(m-r+2)]\}} \text{ and } p = \frac{r(\frac{\pi_r}{\pi_{r-1}})^{1/\alpha}}{m-r+1+r(\frac{\pi_r}{\pi_{r-1}})^{1/\alpha}}.$$

The model has been shown to provide a satisfactory fit, for a range of (confidential) commercial Customer Value data sets.

4 Comparisons and extensions

The Katz family of distributions is also indexed by two parameters, θ_1 and θ_2 say, where $\theta_1 > 0$ and $\theta_2 < 1$. Its probabilities $\varpi_0, \varpi_1, ...$are defined recursively by

$$\varpi_{j+1} = \frac{\theta_1+\theta_2 j}{1+j}\varpi_j, \qquad j = 0, 1, 2, ...$$

where

$$\text{if } \ \theta_1 + \theta_2 j < 0, \text{ then } \varpi_{j+i} = 0, \text{ all } i > 0$$

In effect, this latter condition defines the range of the variate, effectively accounting for one of the two parameters. The other parametric degree of freedom is then used to locate the distribution. For this reason, the Katz family is not suitable for our purposes.

Reference was made in the Introduction to a 3-parameter version of the Power Binomial family. From Figure 1*b*, it is clear that for some values of α and p and for larger values of m, the resulting shapes are concentrated on a small subset of the range of possible values. This suggests that m could also be treated as a parameter if one is seeking to model discrete data that are concentrated on an interval somewhat removed from the origin.

Finally, by analogy with the Poisson limit for the Binomial distribution, a two-parameter 'Power Poisson' distribution can be produced, although the value of such a creation is not evident: there are other more tractable discrete two-parameter families concentrated on $\{n : n \geq 0\}$.

References

[1] Clark L., Cleveland W.S., Denby L. & Liu C. Modelling of Customer Polling Data.1997: In *Proceedings of the 4th Workshop on Case Studies in Bayesian Statistics*, R. Kass (ed.). Carnegie Mellon University, Pittsburgh.

[2] Gale B.T. with Wood R.C. *Managing Customer Value.* The Free Press, 1994; New York.

[3] Griffiths, D.A. Maximum likelihood estimation for the Beta-Binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics.* 1973; **29**: 637-648.

[4] Johnson N.L., Kotz S. & Kemp A. *Discrete Distributions.* Second edition. New York, 1992; John Wiley & Sons.

[5] Kordupleski R.E., Rust R.T. & Zahorik A.J. *W*hy Improving Quality Doesn't Improve Quality (Or Whatever Happened to Marketing?). *California Management Review.* 1993; **35** (Spring): 82-95.

[6] Laitamäki J. & Kordupleski R.E. Building and deploying profitable growth strategies based on the waterfall of Customer Value Added. *European Management Journal.*1997; **15**(2): 158-166.

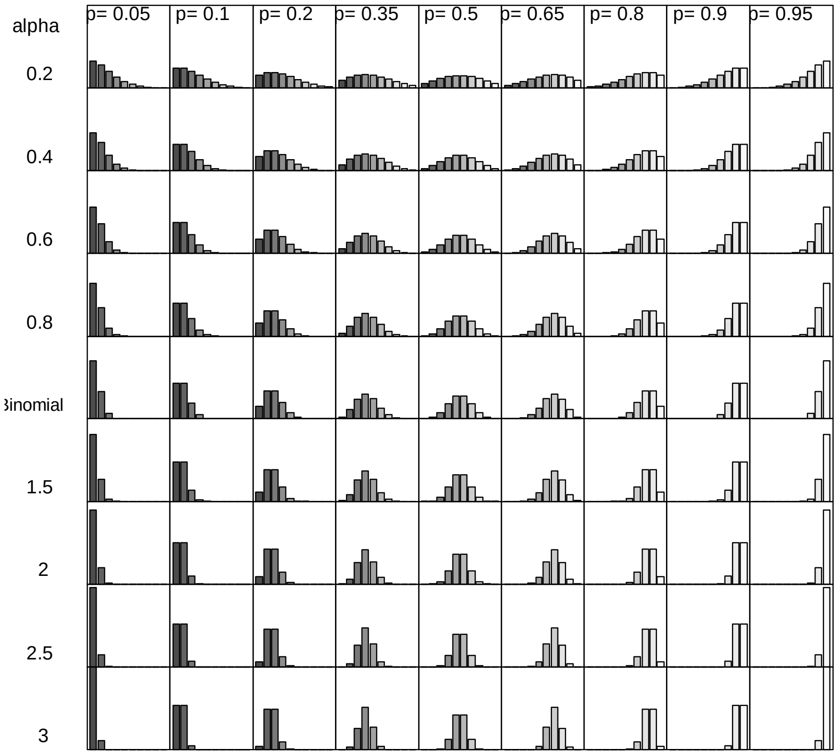

Figure 1: Figure 1a. Shapes of the Power Binomial $PB(m, p, \alpha)$ distribution, for $m = 9$ and positive values of the dispersion parameter α. $\alpha = 1$ corresponds to the Binomial distribution.

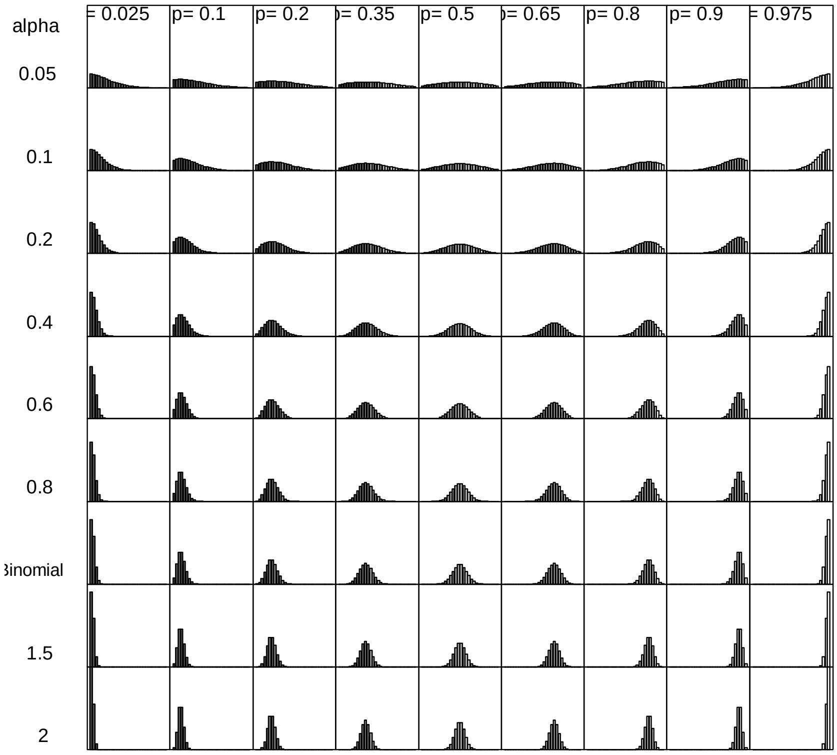

Figure 2: Figure 1b. Shapes of the Power Binomial $PB(m, p, \alpha)$ distribution, for $m = 29$ and positive values of the dispersion parameter α. $\alpha = 1$ corresponds to the Binomial distribution.

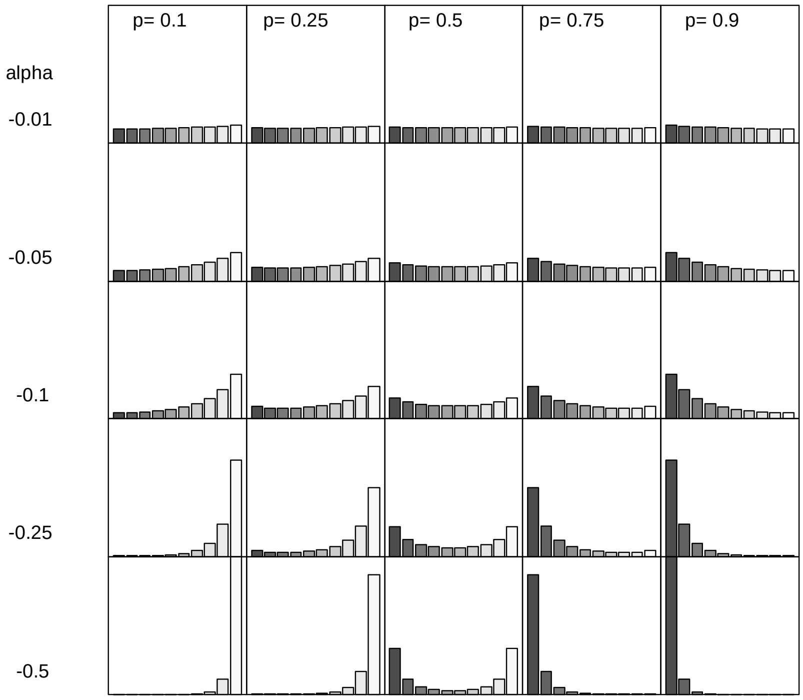

Figure 3: Figure $2a$. Shapes of the Power Binomial $PB(m, p, \alpha)$ distribution, for $m = 9$ and negative values of the dispersion parameter α.

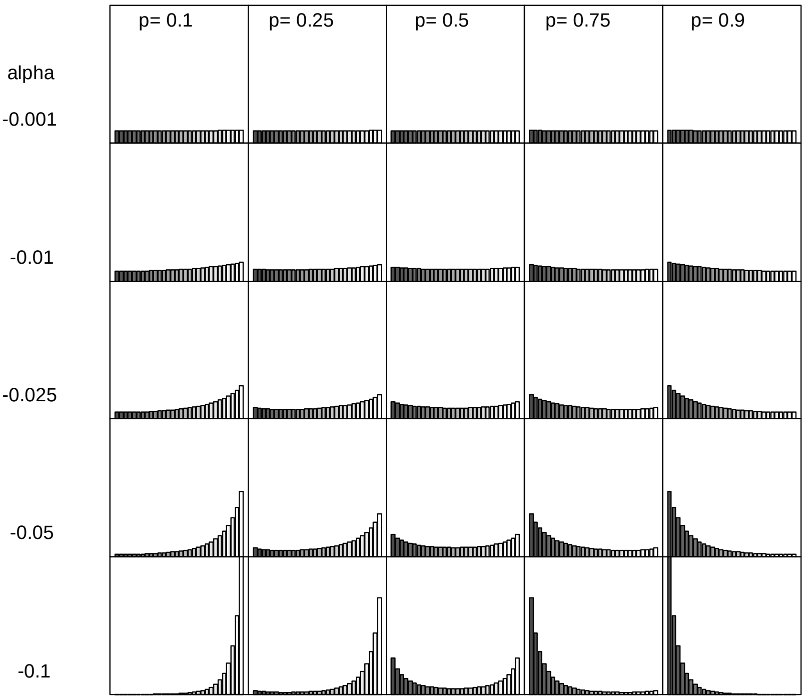

Figure 4: Figure 2*b*. Shapes of the Power Binomial $PB(m, p, \alpha)$ distribution, for $m = 29$ and negative values of the dispersion parameter α.