Exploring the Equivalence of Two Common Mixture Models for Duration Data

Peter S. Fader
The Wharton School, University of Pennsylvania

Bruce G. S. Hardie
London Business School

Daniel McCarthy
Goizueta Business School, Emory University

Ramnath Vaidyanathan
DataCamp Inc.

October 2018

Abstract

The beta-geometric (BG) distribution and the Pareto distribution of the second kind (P(II)) are two basic models for duration-time data that share some underlying characteristics (i.e., continuous mixtures of memoryless distributions), but differ in two important respects: first, the BG is the natural model to use when the event of interest occurs in discrete time, while the P(II) is the right choice for a continuous-time setting. Second, the underlying mixing distributions (the beta and gamma for the BG and P(II), respectively, are very different — and often believed to be non-comparable with each other. Despite these and other key differences, the two models are strikingly similar in terms of their fit and predictive performance as well as their parameter estimates. We explore this equivalence, both empirically and analytically, and discuss the implications from both a substantive and methodological standpoint.

Keywords: beta-geometric; Pareto of the second kind; Grassia(II)

1 Introduction

For the past 20+ years, the first two authors have been teaching probability modeling courses that focus on the development of data-based models that are used to characterize and predict behavior in many areas of business (and, more generally, the social sciences). The initial focus is on models for three fundamental behavioral processes — counting ("how many"), timing ("when / how long"), and "choice" ("whether / which") — drawing on the rich tradition of (continuous) mixture models.

The first timing model considered is the beta-geometric distribution as applied to discrete-time duration data. This is based on the assumption that individual duration times are characterized by a geometric distribution with parameter θ, and the differences in θ across individuals are captured by a beta distribution with shape parameters γ and δ. The probability mass function (pmf) and cdf of this (continuous) mixture model are:

$$\begin{aligned} P(T = t \,|\, \gamma, \delta) &= \int_0^1 \theta(1-\theta)^{t-1} \frac{\theta^{\gamma-1}(1-\theta)^{\delta-1}}{B(\gamma,\delta)}\, d\theta \\ &= \frac{B(\gamma+1, \delta+t-1)}{B(\gamma,\delta)}, \quad t = 1, 2, \ldots \qquad (1) \\ F(t \,|\, \gamma, \delta) &= \int_0^1 \left\{1-(1-\theta)^t\right\} \frac{\theta^{\gamma-1}(1-\theta)^{\delta-1}}{B(\gamma,\delta)}\, d\theta \\ &= 1 - \frac{B(\gamma, \delta+t)}{B(\gamma,\delta)}, \quad t = 0, 1, 2, \ldots \qquad (2) \end{aligned}$$

The BG distribution has been used to model a number of duration-time phenomena, including the time required to achieve pregnancy (Potter and Parker 1964), the length of stay in a psychiatric hospital (Kaplan 1982), and the length of a customer's relationship with a firm (Fader and Hardie 2007).

One homework exercise associated with this session poses the following (real) problem. Ace Snackfoods, Inc. has developed a new shelf-stable juice product called Kiwi Bubbles. (Both the company and brand names are masked.) Before deciding whether or not to undertake a nationwide launch of the new product, the marketing manager of Kiwi Bubbles has decided to conduct a year-long test market in a limited geographic area with a view to getting a clearer picture of the product's potential. Purchasing of the new product is being monitored via a market

research panel that tracks the purchasing of a demographically representative sample of 1499 households across supermarkets, convenience stores, warehouse clubs and mass merchandisers in the test area. The product has now been under test for 24 weeks. On hand is a dataset documenting the cumulative number of households that have made a trial purchase (i.e., first-ever purchase of the new product) by the end of each week — see Table 1.

Week	# Households	Week	# Households
1	8	13	68
2	14	14	72
3	16	15	75
4	32	16	81
5	40	17	90
6	47	18	94
7	50	19	96
8	52	20	96
9	57	21	96
10	60	22	97
11	65	23	97
12	67	24	101

Table 1: Cumulative number of households that have made a trial purchase by the end of weeks 1–24.

The marketing manager for Kiwi Bubbles would like a forecast of the product's year-end performance in the test market. As a first step, she wants a forecast of the number of households that will have made a trial purchase by week 52.[1] Looking closely at Table 1, we see that eight households made their trial purchase of Kiwi Bubbles during its first week on the market. Another six made their trial purchase of Kiwi Bubbles during its second week on the market. And so on. At the end of the 24-week test period, 1398 households have yet to make a trial purchase.

In order to generate the forecast of interest to the manager, we need a statistical model that characterizes the duration of the time from the new product's launch to its purchase by the households in the market. Solving this problem first sees students fitting the BG to the data in Table 1. The maximum likelihood estimates of the model parameters are obtained by

[1] Ultimately, she will want a forecast of both trial and repeat purchases of the new product to come up with an estimate of total sales. See Fader et al. (2014) for a discussion of the historical development of models of repeat-buying for new products, and Fader et al. (2004) for an integrated model of new product purchasing. See Urban and Hauser (1993) for a discussion of the use of test markets by companies as part of their new product development and commercialization processes.

maximizing the the following log-likelihood function:

$$LL(\gamma, \delta) = \left\{ \sum_{t=1}^{24} n_t \ln \left[P(T = t \,|\, \gamma, \delta) \right] \right\} + \left(1499 - \sum_{t=1}^{24} n_t \right) \ln \left[1 - F(24 \,|\, \gamma, \delta) \right], \qquad (3)$$

where n_t is the number of households that made a trial purchase in week t ($n_1 = 8$, $n_2 = 6$, etc.). The maximum value of the log-likelihood function is $LL = -681.4$, which occurs at $\hat{\gamma} = 0.050$ and $\hat{\delta} = 8.434$. The expected number of households that will have made a trial purchase by (the end of) week t is simply $1499 \times F(t \,|\, \gamma, \delta))$.

A later class session starts by looking at this dataset and noting that the trial purchasing process is not discrete; the underlying process operates in continuous time — the product can be purchased at any time of the day or night, seven days a week — but the data have been discretized in the reporting process. The development of a model for continuous-time duration data starts by identifying a suitable replacement for the geometeric distribution and an appropriate mixing distribution. A natural starting point to assume that individual duration times are characterized by an exponential distribution with parameter λ. We then assume that differences in λ across individuals are captured by a gamma distribution with shape parameter r and inverse-scale parameter α.[2] The resulting (continuous) mixture model has the following pdf and cdf:

$$\begin{aligned} f(t \,|\, r, \alpha) &= \int_0^\infty \lambda e^{-\lambda t} \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)} \, d\lambda \\ &= \frac{r}{\alpha} \left(\frac{\alpha}{\alpha + t} \right)^{r+1}, \; t \geq 0 \,, \qquad (4) \\ F(t \,|\, r, \alpha) &= \int_0^\infty \left\{ 1 - e^{-\lambda t} \right\} \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)} \, d\lambda \\ &= 1 - \left(\frac{\alpha}{\alpha + t} \right)^{r}, \; t \geq 0 \,. \qquad (5) \end{aligned}$$

This gamma mixture of exponentials is also known as the Pareto distribution of the second kind, hereafter P(II), and the Lomax distribution. It has been used to model many duration-time phenomena, including the duration of jobs, strikes, and wars (Morrison and Schmittlein 1980), business failures (Lomax 1954), new product trial purchasing (Anscombe 1961, Hardie et al. 1998), and the expulsion of intrauterine contraceptive devices (Aalen 1987).

[2] In many references, α is called a scale parameter; in R, it is called a rate parameter.

Fitting this model to the data in Table 1 sees us maximizing the following log-likelihood function, which takes into account the interval-censored nature of the data:

$$LL(r,\alpha) = \left\{\sum_{t=1}^{24} n_t \ln\left[\int_{t-1}^{t} f(u\,|\,r,\alpha)du\right]\right\} + \left(1499 - \sum_{t=1}^{24} n_t\right)\ln\left[1 - F(24\,|\,r,\alpha)\right]$$
$$= \left\{\sum_{t=1}^{24} n_t \ln\left[F(t\,|\,r,\alpha) - F(t-1\,|\,r,\alpha)\right]\right\} + \left(1499 - \sum_{t=1}^{24} n_t\right)\ln\left[1 - F(24\,|\,r,\alpha)\right]. \quad (6)$$

The maximum value of the log-likelihood function is $LL = -681.4$, which occurs at $\hat{r} = 0.050$ and $\hat{\alpha} = 7.973$. The expected number of households that will have made a trial purchase by (the end of) week t is simply $1499 \times F(t\,|\,r,\alpha)$.

Given the nature of the problem being solved via these probability models, it is natural to look at the associated forecasts of trial purchasing. For both models, the estimates of the expected number of households that will have made a trial purchase by (the end of) week t are compared with the actual cumulative number of households making a trial purchase in Figure 1. (Weeks 1–24 are the numbers from Table 1 as used for model calibration; weeks 25–52 are a longitudinal validation period.) The first thing we note is that the model-based projections of cumulative trial generated from such simple models of household buying behavior are quite accurate. Looking at it more closely, we see that it is not possible to distinguish visually between the two model-based predictions. Over this 52-week period, the largest difference between the two cdfs is $7.76E{-}06$.[3]

Stepping back and comparing the estimation results, we note is that the two log-likelihood functions take on the same value.[4] This follows from the observation that the two model-based predictions are the same. But why are the predictions from these two different models the same on the first place? Is this a chance happening? Looking at the two sets of parameter estimates, we see that $\hat{\gamma} = \hat{r}$ (at least to three decimal places) and that $\hat{\delta}$ is surprisingly similar to $\hat{\alpha}$. A priori there does not appear to be any reason why this should be the case. Why would the shape parameter of the gamma distribution underlying the P(II) model equal the first shape parameter of the beta distribution underlying the BG model? And why would the inverse-scale

[3]The greatest absolute difference between the two cdfs over an indefinitely long time horizon is less than $1.9E{-}04$, which occurs at $t \approx 2.98E{+}10$ where $F(t) \approx 0.6696$.

[4]Strictly speaking, they differ by less than 0.0001%.

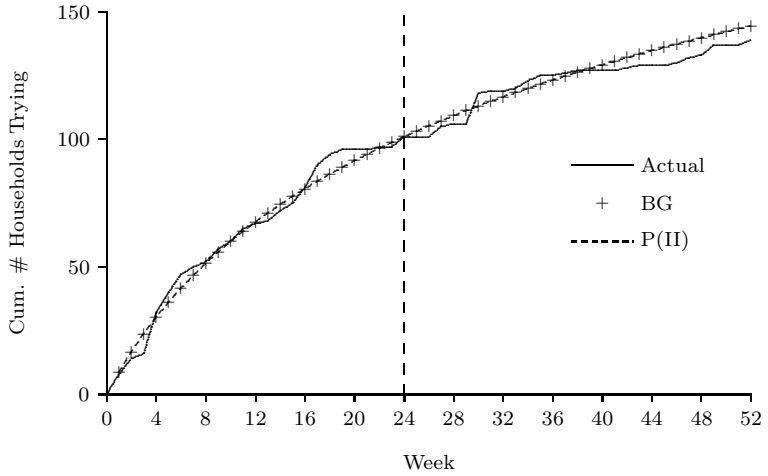


Figure 1: Comparing the estimates of expected cumulative trial generated by the BG and P(II) models with the actual numbers.

parameter of the gamma distribution be roughly equal to the second shape parameter of the beta distribution? We explore the phenomenon of similar model fit in Section 2 and the near equivalence of the parameters in Sections 3 and 4.[5]

2 Similarity of Fit

Let us return to the BG model. Given the assumption of geometrically distributed duration times at the level of the individual, the beta distribution is the obvious choice for modeling heterogeneity in θ. A less-obvious alternative would be to use the transformation $\theta = 1-\exp(-\lambda)$ where differences in λ across individuals are characterized by the gamma distribution with parameters r and α. This is equivalent to assuming that heterogeneity in θ is captured by a Grassia(II) distribution (Grassia 1977). (The Grassia(I) distribution, also known as the log-gamma distribution (Consul and Jain 1971) and the unit-gamma distribution (Ratnaparkhi and Mosimann 1990), follows from the transformation $\theta = \exp(-\lambda)$.)

[5]The Merriam-Webster online dictionary defines *equivalence* is "the state or property of being equivalent," with *equivalent* defined as "equal in [...] value."

The resulting mixture model has the following pmf and cdf:

$$\begin{aligned} P(T=t\,|\,r,\alpha) &= \int_0^\infty (1-e^{-\lambda})\big(e^{-\lambda}\big)^{t-1}\frac{\alpha^r\lambda^{r-1}e^{-\alpha\lambda}}{\Gamma(r)}\,d\lambda \\ &= \left(\frac{\alpha}{\alpha+t-1}\right)^r - \left(\frac{\alpha}{\alpha+t}\right)^r, \quad t=1,2,\ldots \qquad (7) \\ F(t\,|\,r,\alpha) &= \int_0^\infty \Big\{1-\big(e^{-\lambda}\big)^t\Big\}\frac{\alpha^r\lambda^{r-1}e^{-\alpha\lambda}}{\Gamma(r)}\,d\lambda \\ &= 1-\left(\frac{\alpha}{\alpha+t}\right)^r, \quad t=0,1,2,\ldots \qquad (8) \end{aligned}$$

In other words, the Grassia(II)-geometric (G_2G) distribution is simply a discretized P(II) distribution. (Substituting (7) and (8) for their BG equivalents in (3) gives us (6).)

Griffiths and Schafer (1981) examine the "closeness" of the beta and Grassia (I and II) distributions using the Kolmogorov-Smirnov (K-S) distance measure and conclude that "Grassia's two distributions mimic the Beta in behaviour and that properties of all three for given mean and variance are sufficiently similar as to be regarded as practically identical. Thus, in applications, choice between them could rest entirely on mathematical convenience" (p. 245).[6]

Given the "closeness" of the beta and Grassia (II) distributions, the "closeness" of the BG and P(II) distributions — as reflected in the equal maximum values of the log-likelihood functions and Figure 1 — should come as no surprise. But this does not explain why the parameter estimates should be so close.

3 The P(II) as a Limit of the BG

The data presented in Table 1 are reported at a weekly level and the underlying unit of time chosen when specifying the models was a week. But what happens if we change the underlying time scale? Suppose it were a day (in which case the ends of the weeks correspond to $t = 7, 14, \ldots$), or an hour (in which case the ends of the weeks correspond to $t = 168, 336, \ldots$)? What happens to the associated parameter estimates?

We re-fit the BG and P(II) models to the data presented in Table 1, changing the underlying

[6] As a mixing distribution for the geometric distribution, the Grassia(I) distribution is not as "mathematically convenient" as the Grassia(II) distribution in that the resulting expression for the cdf of the Grassia(I)-geometric distribution, also known as the geometric-gamma distribution (Dubey 1966), is more complex.

time scale in the manner suggested above. The associated parameter estimates are reported in Table 2.

		BG			P(II)		
Time unit	n	γ	δ	δ/n	r	α	α/n
Week	1	0.050	8.434	8.434	0.050	7.973	7.973
Day	7	0.050	56.287	8.041	0.050	55.813	7.973
Hour	168	0.050	1339.745	7.975	0.050	1339.517	7.973

Table 2: Examining the impact of changing the underlying time scale on the parameter estimates of the BG and P(II) models.

In the case of the P(II) model, the parameter estimates behave as would be expected. The parameters r and α of the underlying gamma distribution are called the shape and inverse-scale parameters, respectively. As we rescale time, r remains constant and, as expected from examining (5), α is simply rescaled by the number of underlying time units in a week (n).

While not obviously expected given (2), we observe a similar pattern in the parameters of the BG model: γ remains constant (and equal to r) and δ is approximately rescaled by n. Looking at δ/n more closely, we note that it appears to be converging towards the $n = 1$ estimate of α from the P(II) model as n increases.

Is this pattern to be expected? The answer is yes, as the P(II) distribution can be derived as a limiting form of the BG distribution, as developed in the following proposition.

Proposition: *Suppose our (discrete) time unit is divided into n equal subintervals. Let the random variable T_n denote the number of subintervals to the time at which the event of interest occurs, with $F_n(t) = P(T_n \leq nt)$. If T_n is distributed beta-geometric with parameters $(r, n\alpha)$, then*

$$\lim_{n\to\infty} F_n(t) = 1 - \left(\frac{\alpha}{\alpha + t}\right)^r ,$$

which is the cdf of the P(II) distribution with parameters r and α.

Proof: We start by noting the following limit operation on the ratio of gamma functions:

$$\begin{aligned}\lim_{n\to\infty}\frac{\Gamma(cn+a)}{\Gamma(cn+b)}n^{b-a} &= \lim_{y\to\infty}\frac{\Gamma(y+a)}{\Gamma(y+b)}\Big(\frac{y}{c}\Big)^{b-a}\\ &= c^{a-b}\lim_{y\to\infty}\frac{\Gamma(y+a)}{\Gamma(y+b)}y^{b-a}\end{aligned}$$

which, noting Abramowitz and Stegun (1972), equation 6.1.46,

$$= c^{a-b}\,. \tag{9}$$

Now,

$$F_n(t) = P(T_n \le nt) = 1 - \frac{B(r, n\alpha + nt)}{B(r, n\alpha)}\,.$$

Multiplying the ratio of beta functions by n^r/n^r, expressing the beta functions in term of gamma functions, and rearranging terms give us

$$F_n(t) = 1 - \frac{\Gamma(r+n\alpha)}{\Gamma(n\alpha)n^r}\frac{\Gamma(n(\alpha+t))n^r}{\Gamma(r+n(\alpha+t))}\,.$$

It follows from (9) that

$$\lim_{n\to\infty}\frac{\Gamma(r+n\alpha)}{\Gamma(n\alpha)n^r} = \alpha^r \text{ and } \lim_{n\to\infty}\frac{\Gamma(n(\alpha+t))n^r}{\Gamma(r+n(\alpha+t))} = (\alpha+t)^{-r}\,,$$

and therefore

$$\lim_{n\to\infty} F_n(t) = 1 - \left(\frac{\alpha}{\alpha+t}\right)^r\,.$$

To the best of our knowledge, this result is previously undocumented. However, at an intuitive level, it should not be too surprising given the basic result that the exponential distribution can be viewed as the limiting form of the geometric distribution.

What we take away from this is that the approximate equivalence of the model parameters examined in the previous section becomes strict equivalence (i.e., $\gamma = r$ and $\delta = \alpha$) as the underlying time scale on which event times are recorded gets smaller and smaller. But this still does not explain the near equivalence of the model parameters on the original time scale.

4 Similarity of the Parameter Estimates

In order to explore the similarity of the parameter estimates from the two models, let us equate the cdfs of the BG and P(II) distributions at $t = 1$ and $t = 2$. Expressing the beta functions in terms of gamma functions and simplifying, we get

$$\frac{\delta}{\gamma + \delta} = \left(\frac{\alpha}{\alpha + 1}\right)^r \tag{10}$$

$$\left(\frac{\delta}{\gamma + \delta}\right)\left(\frac{\delta + 1}{\gamma + \delta + 1}\right) = \left(\frac{\alpha}{\alpha + 2}\right)^r \tag{11}$$

It follows from (10) that

$$\delta = \frac{\gamma}{\left(1 + \frac{1}{\alpha}\right)^r - 1}\,. \tag{12}$$

Substituting (10) and (12) in (11) and solving for γ, we get

$$\gamma = \frac{\left[\left(\frac{\alpha}{\alpha + 1}\right)^r - 1\right]\left[\left(\frac{\alpha + 1}{\alpha + 2}\right)^r - 1\right]}{\left(\frac{\alpha + 1}{\alpha + 2}\right)^r - \left(\frac{\alpha}{\alpha + 1}\right)^r}\,. \tag{13}$$

For small r, we can approximate (13) by its second-order Taylor-series approximation about $r = 0$,

$$\gamma \approx r \left\{ \frac{\ln\left(\frac{\alpha}{\alpha + 1}\right) \ln\left(\frac{\alpha + 1}{\alpha + 2}\right)}{\ln\left(\frac{\alpha + 1}{\alpha + 2}\right) - \ln\left(\frac{\alpha}{\alpha + 1}\right)} \right\}\,. \tag{14}$$

The bracketed term in (14) is only a function of α. It rapidly converges to 1.0, equaling 0.95 when $\alpha = 0.442$ and 0.99 when $\alpha = 1.95$. As long as α is sufficiently large, γ is for all intents and purposes equal to r. For the new-product trial dataset, $\hat{\alpha} = 7.973$, which puts γ and r within 99.9% of each other, explaining the observed equivalence to three decimal places of the two parameter estimates.

Let us turn to the relationship between δ and α. Recall (12). If we assume $\gamma \approx r$, we have

$$\delta = \frac{r}{\left(1 + \frac{1}{\alpha}\right)^r - 1}\,. \tag{15}$$

Let us consider the ratio of δ and α. We take a second-order Taylor-series approximation of (15)

about $r = 0$ and divide through by α, giving us

$$\delta/\alpha \approx \frac{1}{\alpha \ln\left(\frac{\alpha+1}{\alpha}\right)} - \frac{r}{2\alpha} + \ln\left(\frac{\alpha+1}{\alpha}\right)\frac{r^2}{12\alpha}\,. \tag{16}$$

If $\alpha \gg r$, then $r/\alpha \approx 0$ and, because we have assumed r is small, $r^2/\alpha \approx 0$. This implies $\delta/\alpha \approx 1/(\alpha \ln((\alpha+1)/\alpha))$, and δ/α is once again only a function of α, rapidly approaching 1.0 as α gets large. (We observe this in the parameter estimates reported in Table 2.) For the new-product trial dataset, $\hat{\alpha} = 7.973$, which would suggest that $\hat{\delta}$ is approximately 6.1% larger than $\hat{\alpha}$; this is within striking distance of the 5.8% we actually observe.

We complement this analytical examination of parameter equivalence with a numerical examination in which the BG model is fitted to data generated from the P(II) model, and the maximum likelihood estimates of γ and δ are compared to the true values of r and α, respectively.

In order to identify the parameter space to explore, we plot in Figure 2 the median time-to-event-occurrence under the P(II) distribution as a function of r and α. (The median of the P(II) is given by $\alpha(2^{1/r} - 1)$.)

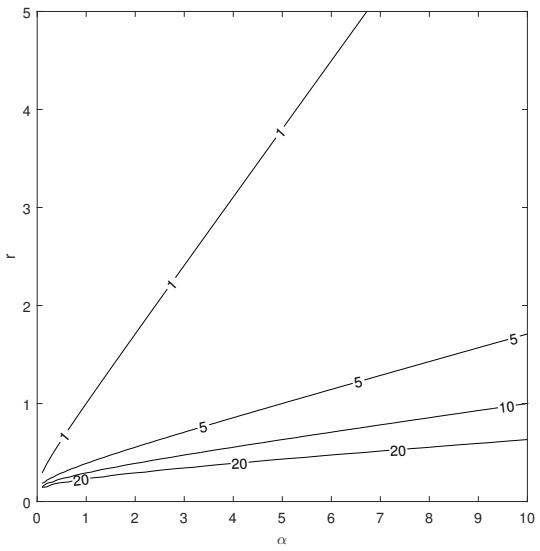


Figure 2: Median lifetime under the P(II) distribution as a function of r and α.

We take the view that settings with very low median lifetimes are rarely encountered, and

therefore choose the parameter space to avoid those areas where the median lifetime is very small. We consider a 100×100 grid of parameters values (with r ranging from 0.02 to 2.00 in increments of 0.02, and α ranging from 0.1 to 10.0 in increments of 0.1). For each point on the grid, we generate ten periods of data and fit the BG model to the resulting dataset. We compute the absolute percentage deviation of $\hat{\gamma}$ from r and the absolute percentage deviation of $\hat{\delta}$ from α, and present contour plots of these quantities as a function of r and α in Figure 3.

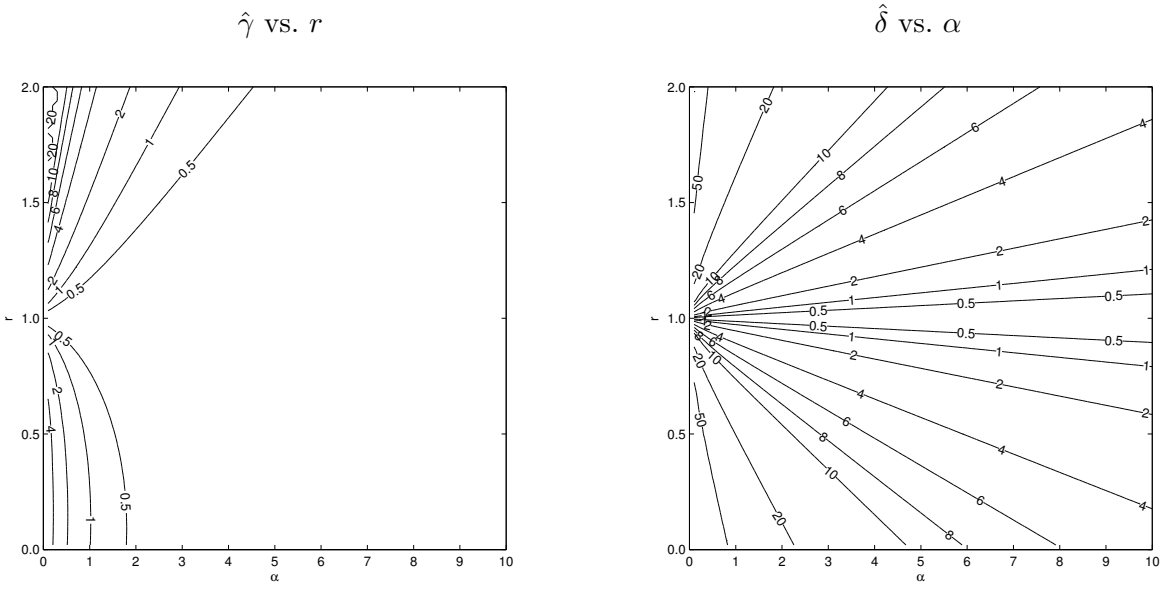


Figure 3: Contour plots of the absolute percentage deviation of the BG parameter estimates from the true P(II) values.

We notice (especially in the right-hand plot in Figure 3) a fanning out of the error around $r = 1$. Referring back to (10) and (11), we note that when $r = 1$, we have a perfect equivalence: $r = \gamma = 1$ and $\alpha = \delta$. (This perfect equivalence holds for $t = 3, 4, 5, \ldots$.) We note when r is large and α small, the errors are high. With reference to Figure 2, this corresponds to settings where the median lifetimes are very short. As the median lifetimes lengthen, the "closeness" of the model parameters increases. The message of these plots is consistent with the analytical examination of error presented above. The absolute percentage deviations for r are such that we will tend to observe $\hat{r} = \hat{\gamma}$. As observed in our empirical example, $\hat{\alpha}$ and $\hat{\delta}$ are close, but less likely to be equal.

How does this map to the similarity in fit? For each point on the grid, we compute the

percentage difference between the value of the BG log-likelihood function (evaluated at the estimated values of γ and δ) and the value of the P(II) log-likelihood function (evaluated using the corresponding values of r and α that generated the data).[7] The contour lines plotted in Figure 4 represents a 0.0001% difference; to the right of the two lines, the percentage difference is less than 0.0001%. The maximum difference observed in the lower left of the plot is 0.007%. The maximum difference observed in the upper left of the plot is 30%; with reference to Figure 2, this only occurs when the median lifetimes are extremely short. We can therefore conclude that in most realistic situations (e.g., those without very short median lifetimes), any differences between r and γ and α and δ have no impact on the fit of the two models; the BG and P(II) will provide effectively equal fits to a given dataset.[8]

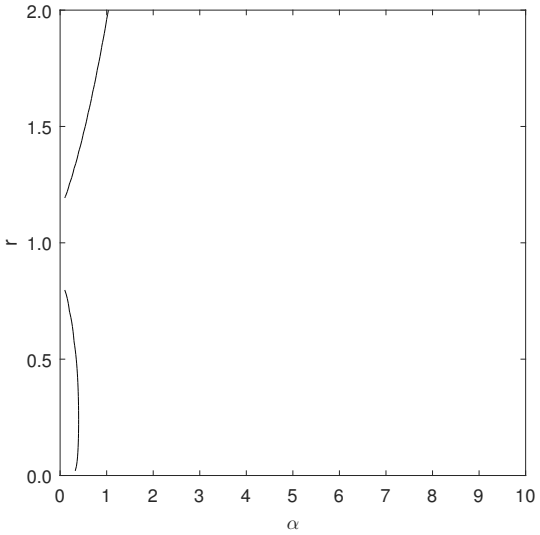


Figure 4: Contour plot of the percentage deviation of the BG and (II) log-likelihood function values as a function of r and α. (The contour lines represent a 0.0001% difference.)

5 Discussion

The beta-geometric (BG) distribution and the Pareto distribution of the second kind (P(II)) are two basic models for duration-time data that share some underlying characteristics (i.e.,

[7] We compute the percentage differences rather than the absolute differences as the value of the P(II) log-likelihood function varies by a factor of just under 40 across the grid of parameter values.

[8] This equivalence is also reflected in an examination of the maximum absolute difference between the two cdfs across the grid of parameter values.

continuous mixtures of memoryless distributions), but differ in two important respects. First, the BG is the natural model to use when the event of interest occurs in discrete time, while the P(II) is the right choice for a continuous-time setting. Second, the underlying mixing distributions (the beta and gamma for the BG and P(II), respectively) are very different. When the event can occur at any point in time, it is natural to consider a continuous-time model such as the P(II). Even in situations where events can occur at any point in time, the data-recording or reporting processes are frequently such that the data are interval-censored (as in Table 1). This raises the option of modeling such data using either a discrete-time model (such as the BG) or a continuous-time model (such as the P(II)).

Does it matter whether the statistician uses the BG or P(II) model given interval-censored data, or the BG or P(II) model given truly discrete data? The answer is no. We have shown that either model will give, for all intents and purposes, the same fit to the data. This follows from the fact that fitting the P(II) to the interval-censored data is equivalent to using a Grassia(II) mixture of geometrics. Given the closeness of the beta and Grassia (II) distributions, the closeness of the BG and P(II) distributions — as reflected in the equal maximum values of the log-likelihood functions and Figure 1 — is to be expected. Furthermore, our analysis suggests that we would expect the parameter estimates to be similar. In particular, we would expect $\hat{\gamma}$ to be very close to $\hat{r}$, with $\hat{r}$ slightly greater than $\hat{\gamma}$ if they are not equal. We would also expect $\hat{\delta}$ and $\hat{\alpha}$ to be close, with $\hat{\delta}$ slightly greater than $\hat{\alpha}$.

From a purely pragmatic perspective we would suggest that the BG model be used in any setting where the analyst needs to explain the logic of the underlying model to a non-technical audience; it can be described in terms of a coin-flipping process, where each person has a coin with a different probability of coming up heads. (It is a lot more difficult to explain the logic of an exponential distribution to an audience with limited statistical knowledge.) However, from a goodness of fit and even a parameter estimate standpoint, the BG and P(II) are two sides of the same coin.

5.1 Use in the Classroom

In the *Introduction* section, we discussed how the issues explored in this paper naturally arise in a classroom setting. How are the results presented here using the classroom? This depends

on the both technical sophistication of the student group and the time one has to explore these issues.

At the very minimum, we have a brief discussion of the equivalence of the two models in most empirical settings. If time and student interest permits, exploring the equivalence in the Section 2–4 sequence above makes sense. Otherwise, Section 3 is a good starting point as it reinforces prior learning about limits of distributions (e.g., the exponential distribution as a limit of the geometric distribution and the Poisson distribution as a limit of the binomial distribution) and it reinforces prior, more conceptual, learning about shape and scale parameters. Section 2 can then be used as a starting point for discussions about alternative mixing distributions. The material in Section 4 is reserved for the most technical student audiences.

5.2 The Value of Traditional Mixture Models

Traditional mixture models, such as the NBD (gamma mixture of Poissons), beta-binomial, and the two models considered here, typically use the beta or gamma distribution as the mixing distribution. Historically, this choice was driven by mathematical convenience; both are flexible distributions that result in closed-form marginal distributions. Use of, say, the lognormal distribution in place of the gamma distribution was avoided because researchers had to resort to numerical integration when evaluating the marginal distribution.

This was clearly an issue in the heyday of such models (say, the 1920s–1970s), when computing resources were such that evaluating integrals numerically was a costly exercise. However, it could be argued that this is not a concern these days. These traditional mixture models, with their emphasis on closed-form marginal distributions, are quaint, of historical interest, but are of no real value today.

It must not be forgotten that these "old" models have proven to be very robust over the years, and it should not be assumed that other mixing distributions are necessarily "better." To illustrate this, let us consider alternative models for solving the new product trial forecasting problem. An alternative mixing distribution for θ is the logit-normal distribution (also known as Johnsons S_B distribution) in which $\text{logit}(\theta)$ is assumed to be distributed normal. And an alternative mixing distribution for λ is the lognormal distribution. As there are no closed-form expressions for the associated marginal distributions, the associated integrals need to be solved

by numerical integration.

We fit a logit-normal mixture of geometrics and a lognormal mixture of exponentials to the dataset considered here (where the integrals are solved using Monte Carlo integration). The values of the associated log-likelihood functions are -682.4 and -682.6, respectively. The corresponding estimates of $F(52)$, the proportion of the population that have made a trial purchase by the end of week 52, are 0.106 and 0.107 respectively. This is in contrast to 0.096 for the BG and P(II) models. The empirical proportion is 0.093. Thus, in this empirical setting, these two alternative models are dominated both in-sample and out-of-sample by the traditional mixture models.

More fundamentally, it is easy for us to under-estimate the value of closed-form solutions from both a pedagogical and practical perspective. As illustrated in the Appendix, it is a very simple exercise to estimate the model parameters of such models "from scratch" in an Excel worksheet. This means we can introduce these models and, more importantly, a probability-model based approach to problem solving to student groups with limited statistical training. In doing so, we are providing them with a toolkit that they can easily apply in their work situations. In many applied settings, an 80% "solution" for 20% of the effort is very acceptable, if not desirable.

Appendix: Parameter Estimation Using Excel

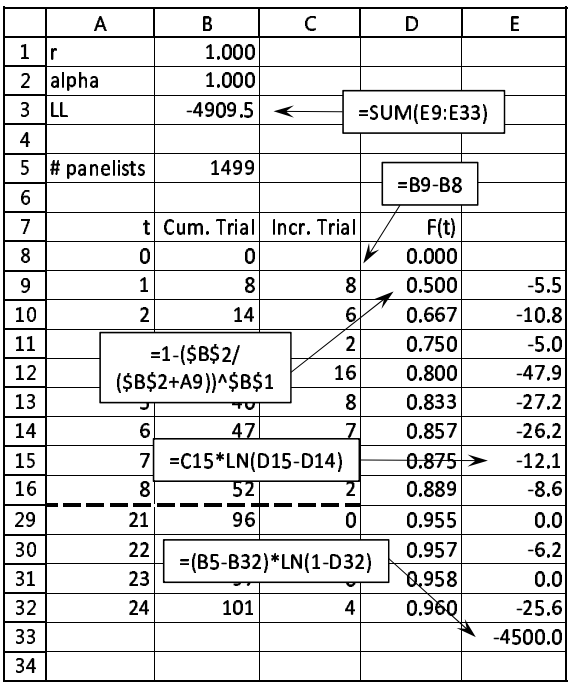
It is simple exercise to obtain the parameter estimates for the two models by coding the associated log-likelihood functions in Excel and then finding the maximum function values using the Solver add-in.

Let first consider the case of the P(II) distribution. The first thing we need to do is code-up (6) in Excel.

i) With reference to Figure A1, we first enter the cumulative trial data (Table 1) in cells `B8:B32` and then compute the incremental/weekly trial numbers (n_t in (6)) by taking the difference between adjacent rows (cells `C9:C32`).

ii) We then compute the P(II) cdf, (4), for $t = 0, 1, 2, \ldots, 24$ in cells `D8:D32`. In order to avoid the `#NUM!` error when computing $F(t)$, we need to specify the values of r and α, which we do so in cells `B1:B2`.

iii) The individual elements of (6) are computed in cells `E9:E33` and the function value reported in cell `B3`.



	A	B	C	D	E
1	r	1.000			
2	alpha	1.000			
3	LL	-4909.5			
4					
5	# panelists	1499			
6					
7	t	Cum. Trial	Incr. Trial	F(t)	
8	0	0		0.000	
9	1	8	8	0.500	-5.5
10	2	14	6	0.667	-10.8
11			2	0.750	-5.0
12			16	0.800	-47.9
13			8	0.833	-27.2
14	6	47	7	0.857	-26.2
15	7			0.875	-12.1
16	8	52	2	0.889	-8.6
29	21	96	0	0.955	0.0
30	22			0.957	-6.2
31	23			0.958	0.0
32	24	101	4	0.960	-25.6
33					-4500.0
34					

Figure A1: Coding-up the P(II) log-likelihood function in Excel.

Having coded-up the log-likelihood function, the next step is to find the values of r and α that maximize this function. This is done using the Solver add-in. With reference to Figure A2, our objective is to find the values of the model parameters (cells `B1:B2`) that maximize the value of the log-likelihood function, whose value is given in cell `B3`, subject to the constraint that the values of r and α are positive.

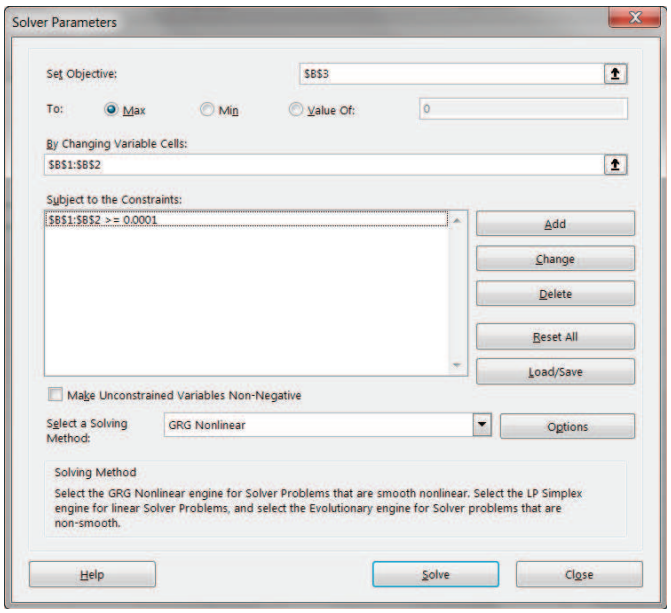


Figure A2: Solver settings.

Clicking on the "Solve" button gives us the results reported in the body of the paper.

In order to generate the model-based projections of cumulative trial out to week 52, we simply compute $F(t)$ for $t = 25, 26, \ldots, 52$ and compute the expected number of households that will have made a trial purchase by (the end of) week t as $1499 \times F(t)$. The actual cumulative trial numbers over the longitudinal validation period are given in Table A1.

Turning to the BG model, we need to code-up (3). The one key difference is that we need to compute $P(T = t)$, (1), which is the ratio of two beta functions. While Excel does not have a beta function, we can evaluate it by expressing it in terms of gamma functions,

$$B(\gamma, \delta) = \frac{\Gamma(\gamma)\Gamma(\delta)}{\Gamma(\gamma + \delta)},$$

and making use of Excel's log gamma (`gammaln()`) function. Alternatively, we can compute the

Week	# Households	Week	# Households
25	101	39	127
26	101	40	127
27	105	41	127
28	106	42	128
29	106	43	129
30	118	44	129
31	119	45	129
32	119	46	130
33	120	47	132
34	123	48	133
35	125	49	137
36	125	50	137
37	126	51	137
38	127	52	139

Table A1: Cumulative number of households that have made a trial purchase by the end of weeks 25–52.

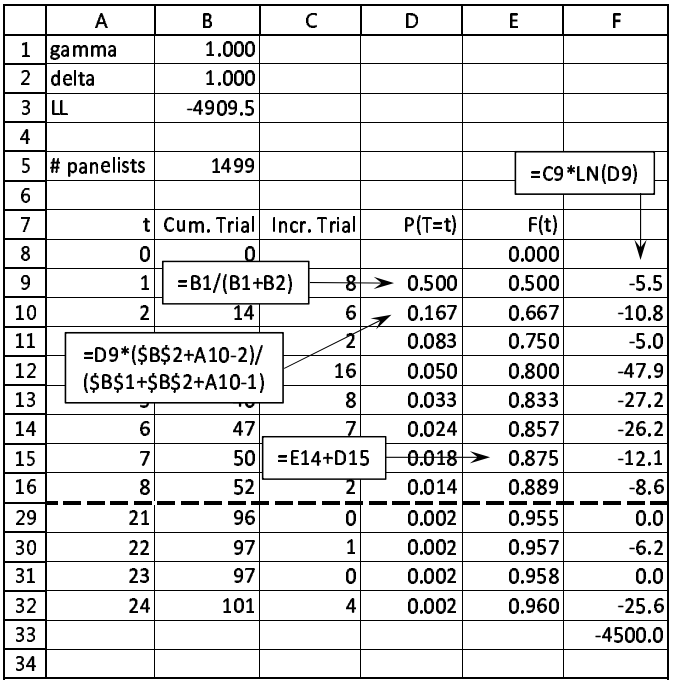
BG pmf using the following forward recursion from $P(T = 1)$:

$$P(T = t \,|\, \gamma, \delta) = \begin{cases} \dfrac{\gamma}{\gamma + \delta} & t = 1 \\ \\ \dfrac{\delta + t - 2}{\gamma + \delta + t - 1} \times P(T = t - 1) & t = 2, 3, \ldots \end{cases}$$

(The derivation of this forward-recursion is a good in-class learning exercise.) With reference to Figure A3, we compute this for $t = 1, 2, \ldots, 24$ in cells `D9:D32` and compute the cdf as $F(t) = F(t-1) + P(T = t)$ (cells `E8:E32`). As for the P(II) model, the individual elements of (3) are computed in cells `F9:F33` and the function value reported in cell `B3`. We use Solver to find the values of γ and δ that maximize the value of this function.

Reflections on Using Excel

It is very common to make use of Excel in introductory statistics courses for business students — see, for example, the reviews of associated textbooks in past issues of this journal. Nevertheless, many statisticians are quite negative about its use for teaching statistics, be it for concerns about the accuracy of Excel's statistical functions or the lack of suitable audit trails for data



	A	B	C	D	E	F
1	gamma	1.000				
2	delta	1.000				
3	LL	-4909.5				
4						
5	# panelists	1499				
6						
7	t	Cum. Trial	Incr. Trial	P(T=t)	F(t)	
8	0	0			0.000	
9	1		8	0.500	0.500	-5.5
10	2	14	6	0.167	0.667	-10.8
11			2	0.083	0.750	-5.0
12			16	0.050	0.800	-47.9
13			8	0.033	0.833	-27.2
14	6	47	7	0.024	0.857	-26.2
15	7	50		0.018	0.875	-12.1
16	8	52	2	0.014	0.889	-8.6
29	21	96	0	0.002	0.955	0.0
30	22	97	1	0.002	0.957	-6.2
31	23	97	0	0.002	0.958	0.0
32	24	101	4	0.002	0.960	-25.6
33						-4500.0
34						

Figure A3: Coding-up the BG log-likelihood function in Excel.

management (e.g., Keeling and Pavur 2011, Nash 2006, the special section on Microsoft Excel 2007 in Volume 52, Issue 10 of *Computational Statistics & Data Analysis*).

While these concerns are valid, we should first note that we are not making use of any of Excel's statistical functions. More importantly, our experience in teaching probability models (at both undergraduate and graduate level, to business and statistics students) leads us to believe strongly in the pedagogical value of "building" the log-likelihood function from scratch in a blank worksheet. We have found such an approach to be far more transparent than coding the function in, say, R — even when teaching groups of students who have some familiarity with a numerical computing environment. (With such student groups, replicating the Excel-based analyses in R is a good homework problem.)

References

Aalen, Odd O. (1987), "Two Examples of Modelling Heterogeneity in Survival Analysis," *Scandinavian Journal of Statistics*, **14** (1), 19–25.

Abramowitz, Milton and Irene A. Stegun (eds.) (1972), *Handbook of Mathematical Functions*, New York: Dover Publications.

Anscombe, F. J. (1961), "Estimating a Mixed-Exponential Response Law," *Journal of the American Statistical Association*, **56** (September), 493–502.

Consul, P. C. and G. C. Jain (1971), "On the Log-Gamma Distribution and its Properties," *Statistische Hefte*, **12** (June), 100–106.

Dubey, Satya D. (1966), "Compound Pascal Distributions," *Annals of the Institute of Statistical Mathematics*, **18** (December), 357–365.

Fader, Peter S. and Bruce G. S. Hardie (2007), "How to Project Customer Retention," *Journal of Interactive Marketing*, **21** (Winter), 76–90.

Fader, Peter S., Bruce G. S. Hardie, and Chun-Yao Huang (2004), "A Dynamic Changepoint Model for New Product Sales Forecasting," *Marketing Science*, **23** (Winter), 50–65.

Fader, Peter S., Bruce G. S. Hardie, and Subrata Sen (2014), "Stochastic Models of Buyer Behavior," in *The History of Marketing Science*, Russell S. Winer and Scott A. Neslin (eds.), Singapore: World Scientific Publishing, 165–205.

Grassia, A. (1977), "On a Family of Distributions with Argument Between 0 and 1 Obtained by Transformation of the Gamma and Derived Compound Distributions," *Australian Journal of Statistics*, **19** (2), 108–114.

Griffiths, David and Christine Schafer (1981), "Closeness of Grassia's Transformed Gammas and the Beta Distribution," *Australian Journal of Statistics*, **23** (2), 240–246.

Hardie, Bruce G. S., Peter S. Fader, and Michael Wisniewski (1998), "An Empirical Comparison of New Product Trial Forecasting Model," *Journal of Forecasting*, **17** (June–July), 209–229.

Kaplan, Edward H. (1982), "Statistical Models and Mental Health: An Analysis of Records From a Mental Health Center," M.S. Thesis, Department of Mathematics, Massachusetts Institute of Technology.

Keeling, Kellie B. and Robert J. Pavur (2011), "Statistical Accuracy of Spreadsheet Software," *The American Statistician*, **65** (November), 265–273.

Lomax, K. S. (1954), "Business Failures: Another Example of the Analysis of Failure Data," *Journal of the American Statistical Association*, **49** (December), 847–852.

Morrison, Donald G. and David C. Schmittlein (1980), "Jobs, Strikes, and Wars: Probability Models for Duration," *Organizational Behavior and Human Performance*, **25** (April), 224–251.

Nash, J. C. (2006), "Spreadsheets in Statistical Practice — Another Look," *The American Statistician*, **60** (August), 287–289.

Potter, R. G. and M. P. Parker (1964), "Predicting the Time Required to Conceive," *Population Studies*, **18** (July), 99–116.

Ratnaparkhi, Makarand V. and James E. Mosimann (1990), "On the Normality of Transformed Beta and Unit-gamma Random Variables," *Communications in Statistics — Theory and Methods*, **19** (10), 3833–3854.

Urban, Glen L. and John R. Hauser (1993), *Design and Marketing of New Products*, 2nd edn., Englewood Cliffs, NJ: Prentice Hall.