Applied Probability Models in Marketing Research: Introduction

(Supplementary Materials for the A/R/T Forum Tutorial)

Bruce G. S. Hardie
London Business School
bhardie@london.edu
www.brucehardie.com

Peter S. Fader
University of Pennsylvania
faderp@wharton.upenn.edu
www.petefader.com

1 Introduction

This note provides further details on the models presented in the Advanced Research Techniques Forum tutorial "Applied Probability Models in Marketing Research: Introduction" conducted by Bruce Hardie and Pete Fader. In particular, the models are formally derived in their general form, with the associated mathematical steps made explicit. Furthermore, methods for parameter estimation are examined and, where deemed appropriate, the mean and variance derived. Finally, the application of empirical Bayes methods is discussed and the relevant formulae derived in a step-by-step manner.

2 Preliminaries

This note assumes basic familiarity with a set of probability distributions and the associated notation. As a refresher, we briefly review the probability distributions that are the building-blocks of the probability models considered in this tutorial. For each distribution, we list its density function, mean and variance, key properties, and relevant additional information. (The parameterization of each distribution is consistent with common usage in the current marketing research literature.)

2.1 Gamma and Beta Functions

The (complete) gamma function $\Gamma(x)$ is defined by the integral

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt,\ x > 0$$

Clearly $\Gamma(1) = 1$. Integrating by parts, we get $\Gamma(x) = (x-1)\Gamma(x-1)$. It follows that if x is a positive integer, $\Gamma(x) = (x-1)!$.

The (complete) beta function $B(\alpha, \beta)$ is defined by the integral

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt,\ \alpha > 0, \beta > 0$$

The relationship between the gamma and beta functions is

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

2.2 Continuous Distributions

Exponential

The continuous random variable X is said to have an exponential distribution if it has a density function of the form

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

where $x > 0$ and $\lambda > 0$. (The parameter λ is sometimes called the rate parameter or, alternatively, the scale parameter.) The corresponding cdf is

$$F(x|\lambda) = 1 - e^{-\lambda x}$$

The mean and variance of the exponential distribution are $E(X) = 1/\lambda$ and $\text{var}(X) = 1/\lambda^2$, respectively.

Gamma

The continuous random variable X is said to have a gamma distribution if it has a density function of the form

$$f(x|r, \alpha) = \frac{\alpha^r x^{r-1} e^{-\alpha x}}{\Gamma(r)}$$

where $x > 0$ and $r, \alpha > 0$. (The parameters r and α are sometimes called the shape and scale parameters, respectively.) For non-integer r, there is no closed-form cdf for the gamma distribution. The mean and variance of the gamma distribution are $E(X) = r/\alpha$ and $\text{var}(X) = r/\alpha^2$, respectively. We note that the gamma density reduces to the exponential density when $r = 1$; furthermore, for integer r, we have the Erlang density.

The gamma distribution is a flexible, right-skewed distribution for continuous random variables defined on the positive real line (i.e., $x > 0$). For $r < 1$, the density is strictly decreasing from an infinite peak at 0. For $r = 1$, the density is strictly decreasing from the point α at $x = 0$. For $r > 1$, the density increases from the origin to a mode at $(r - 1)/\alpha$, then decreases.

Beta

The continuous random variable X is said to have a beta distribution if it has a density function of the form

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $0 \leq x \leq 1$ and $\alpha, \beta > 0$. The mean and variance of the beta distribution are $E(X) = \alpha/(\alpha+\beta)$ and $\text{var}(X) = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$, respectively.

The beta distribution is a flexible distribution for continuous random variables defined on the unit interval (i.e., $[0,1]$). Its density can take on a number of shapes, depending on the specific values of α and β. If $\alpha < 1$, the density has an infinite peak at 0 (i.e., it has a 'reverse J-shape'). If $\beta < 1$, the density has an infinite peak at 1 (i.e, it is 'J-shaped'). When $\alpha, \beta < 1$, the density is 'U-shaped.' For $\alpha = \beta = 1$, we have a uniform distribution on the unit interval. In the case of $\alpha, \beta > 1$, the density has a mode at $(\alpha-1)/(\alpha+\beta-2)$. It follows that the beta density is symmetric when $\alpha = \beta$. As α, β increase, the density tends to a spike at its mean.

Derivation: if Y_1 and Y_2 are independent gamma random variables with shape parameters α and β, respectively, and common scale parameter λ, the random variable $X = Y_1/(Y_1+Y_2)$ has a beta distribution with parameters (α, β).

Dirichlet

The continuous k-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_k)'$ is said to have a Dirichlet distribution if it has a density function of the form

$$f(\mathbf{x}|\mathbf{a}) = \frac{\Gamma(S)}{\prod_{j=1}^{k}\Gamma(a_j)} \prod_{j=1}^{k} x_j^{a_j-1}$$

where $0 \leq x_j \leq 1$ with $\sum_{j=1}^{k} x_j = 1$, $\mathbf{a} = (a_1, \ldots, a_k)'$ with $a_j > 0$, and $S \equiv \sum_{j=1}^{k} a_j$. Note that because $\sum_{j=1}^{k} x_j = 1$, this is actually a $(k-1)$-dimensional distribution since x_k is redundant and can be replaced by $1 - \sum_{j=1}^{k-1} x_j$. Consequently, the density is sometimes written as

$$f(\mathbf{x}|\mathbf{a}) = \frac{\Gamma(S)}{\prod_{j=1}^{k} \Gamma(a_j)} \left(\prod_{j=1}^{k-1} x_j^{a_j-1} \right) \left(1 - \sum_{j=1}^{k-1} x_j \right)^{a_k-1}$$

or

$$f(x_1, \ldots, x_{k-1}|\mathbf{a}) = \frac{\Gamma(S)}{\prod_{j=1}^{k} \Gamma(a_j)} \left(\prod_{j=1}^{k-1} x_j^{a_j-1} \right) \left(1 - \sum_{j=1}^{k-1} x_j \right)^{a_k-1}$$

Furthermore, because $\sum_{j=1}^{k} x_j = 1$, any integration of the complete Dirichlet pdf is performed with respect to $x_1, x_2, \ldots, x_{k-1}$, where the integration limits are $[0, 1], [0, 1 - x_1], \ldots, [0, 1 - \sum_{j=1}^{k-2} x_j]$, respectively.

The mean of the Dirichlet distribution is $E(\mathbf{X}) = \mathbf{a}/S$, with $E(X_j) = a_j/S$. The variance-covariance matrix of the Dirichlet distribution is $\text{var}(\mathbf{X}) = [\text{Diag}(S\mathbf{a}) - \mathbf{a}\mathbf{a}']/[S^2(S + 1)]$, with $\text{var}(X_j) = a_j(S - a_j)/[S^2(S + 1)]$, and $\text{cov}(X_j, X_{j'}) = -a_j a_{j'}/[S^2(S + 1)]$.

The Dirichlet distribution is the multivariate generalization of the beta distribution; for $k = 2$, we have the beta distribution with $\alpha = a_1$, $\beta = a_2$, and $x_2 = 1 - x_1$. The marginal distribution of X_j, an element of $\mathbf{X}$, is beta with parameters $(a_j, S - a_j)$.

Derivation: if $Y_1, \ldots, Y_k$ are independent gamma random variables with shape parameters a_j $(j = 1, \ldots, k)$, and common scale parameter λ, the random vector $\mathbf{X} = (X_1, \ldots, X_k)'$, where $X_j = Y_j/(\sum_{j'=1}^{k} Y_{j'})$ $(j = 1, \ldots, k)$ has a Dirichlet distribution with parameter vector $\mathbf{a} = (a_1, \ldots, a_k)'$.

2.3 Discrete Distributions

Poisson

The discrete random variable X is said to have a Poisson distribution if it has a density function of the form

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where $x = 0, 1, 2, \ldots$ and $\lambda > 0$. (The parameter λ is sometimes called the rate parameter.) The mean and variance of the Poisson distribution are $E(X) = \lambda$ and $\text{var}(X) = \lambda$, respectively.

The Poisson random variable X represents the number of occurrences of a rare event in a unit time interval or two/three dimensional space. In many applications, we are interested in the number of occurrences in a time interval of length t (or its spatial equivalent). In this case, the random variable of interest has a Poisson distribution with rate parameter λt.

If $X_1, \ldots, X_k$ are independent Poisson random variables with rate parameters λ_i $(i = 1, \ldots, k)$, then the random variable $Y = \sum_{i=1}^{k} X_i$ has a Poisson distribution with rate parameter $\lambda = \sum_{i=1}^{k} \lambda_i$. This is called the reproductive property of the Poisson distribution.

Binomial

The discrete random variable X is said to have a binomial distribution if it has a density function of the form

$$P(X = x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $x = 0, 1, 2, \ldots, n$ for positive integer n and $0 \leq p \leq 1$. The mean and variance of the binomial distribution are $E(X) = np$ and $\text{var}(X) = np(1-p)$, respectively.

The binomial random variable X is interpreted as the total number of successes occurring in n independent success/failure (i.e., Bernoulli) trials, where p is the probability of success on each individual trial.

Multinomial

The discrete k-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_k)'$ is said to have a multinomial distribution if it has a density function of the form

$$P(\mathbf{X} = \mathbf{x}|n, \mathbf{p}) = \binom{n}{x_1, \ldots, x_k} \prod_{j=1}^{k} p_j^{x_j}$$

where $x_j \in \{0, 1, 2, \ldots, n\}$ with $\sum_{j=1}^{k} x_j = n$, and $\mathbf{p} = (p_1, \ldots, p_k)'$ with $0 \leq p_j \leq 1$ and $\sum_{j=1}^{k} p_j = 1$. Note that because of the restrictions $\sum_{j=1}^{k} x_j = n$ and $\sum_{j=1}^{k} p_j = 1$, this is actually a $(k-1)$-dimensional distribution since

$x_k = n - \sum_{j=1}^{k-1} x_j$ and $p_k = 1 - \sum_{j=1}^{k-1} p_j$. Consequently, the density is sometimes written as

$$P(\mathbf{X} = \mathbf{x}|n, \mathbf{p}) = \binom{n}{x_1, \dots, x_k} \left(\prod_{j=1}^{k-1} p_j^{x_j} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{n - \sum_{j=1}^{k-1} x_j}$$

or

$$p(x_1, x_2, \dots, x_{k-1}|n, \mathbf{p}) = \\ \binom{n}{x_1, \dots, x_{k-1}, n - \sum_{j=1}^{k-1} x_j} \left(\prod_{j=1}^{k-1} p_j^{x_i} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{n - \sum_{j=1}^{k-1} x_j}$$

The random variable X_j ($j = 1, \dots, k$), an element of $\mathbf{X}$, is interpreted as the number of times outcome j occurs in n independent trials, where each trial results in one of k mutually exclusive (and collective exhaustive) outcomes, and the probability of outcome j occurring on any trial equal to p_j.

The mean of the multinomial distribution is $E(\mathbf{X}) = n\mathbf{p}$, with $E(X_j) = np_j$. The variance-covariance matrix of the multinomial distribution is $\text{var}(\mathbf{X}) = n[\text{Diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}']$, with $\text{var}(X_j) = np_j(1 - p_j)$ and $\text{cov}(X_j, X_{j'}) = -np_j p_{j'}$.

The multinomial distribution is the multivariate generalization of the binomial distribution; for $k = 2$, we have the binomial distribution with $p = p_1 = 1 - p_2$ and $x_2 = n - x_1$. The marginal distribution of X_j, an element of $\mathbf{X}$, is binomial with parameters (n, p_j).

3 The Exponential-Gamma Model

The exponential-gamma model — also known as the Lomax or Pareto distribution — results when we assume that

- the individual-level behavior of interest (e.g., time of trial purchase for a new product) is characterized by the exponential distribution with rate parameter λ, which we denote by $F(t|\lambda)$, and
- the values of λ are distributed across the population according to a gamma distribution, denoted by $g(\lambda)$.

The aggregate distribution of the behavior of interest, which we denote by $F(t)$, is obtained by weighting each $F(t|\lambda)$ by the likelihood of that value of λ occurring (i.e., $g(\lambda)$). This is formally denoted by:

$$F(t) = \int_0^\infty F(t|\lambda)g(\lambda)d\lambda$$

3.1 Model Derivation

In order to derive the aggregate distribution associated with exponentially-distributed event times at the individual-level and gamma heterogeneity, we must solve the following integral

$$P(T \leq t) = \int_0^\infty \underbrace{(1 - e^{-\lambda t})}_{\text{exponential}} \overbrace{\frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)}}^{\text{gamma}} d\lambda$$

This is done in the following manner:

1. Expand the above expression:

$$P(T \leq t) = \int_0^\infty \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)} - \int_0^\infty e^{-\lambda t} \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)}$$

2. By definition, the value of the first integral is 1; therefore we have

$$P(T \leq t) = 1 - \int_0^\infty e^{-\lambda t} \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)} d\lambda$$

3. Combine terms and move all non-λ terms to the left of the integral sign. This gives us

$$P(T \leq t) = 1 - \frac{\alpha^r}{\Gamma(r)} \int_0^\infty \lambda^{r-1} e^{-\lambda(\alpha+t)} d\lambda$$

4. We therefore have to solve the definite integral

$$\int_0^\infty \lambda^{r-1} e^{-\lambda(\alpha+t)} d\lambda$$

The "trick" is to transform the terms to the right of the integral sign into a known pdf, which integrates to 1. Looking closely at these terms, we see the heart of a gamma density with shape parameter r and scale parameter $\alpha + t$. Multiplying the integral by $[\Gamma(r)/(\alpha+t)^r]/[(\alpha+t)^r/\Gamma(r)]$, we can write our expression for $P(T \leq t)$ as

$$P(T \leq t) = 1 - \frac{\alpha^r}{\Gamma(r)} \frac{\Gamma(r)}{(\alpha+t)^r} \int_0^\infty \underbrace{\frac{(\alpha+t)^r \lambda^{r-1} e^{-\lambda(\alpha+t)}}{\Gamma(r)}}_{\text{gamma pdf}} d\lambda$$

5. As the integrand is a gamma pdf, the definite integral, by definition, equals 1, and we therefore write the equation as

$$P(T \leq t) = 1 - \left(\frac{\alpha}{\alpha+t}\right)^r$$

We call this the exponential-gamma model.

3.2 Estimating Model Parameters

In order to apply the exponential-gamma model, we must first develop estimates of the two model parameters, r and α, from the given sample data. The primary method at the modeler's disposal is the method of maximum likelihood.

In most cases, the sample data do not report the exact time at which each individual's behavior occured. Rather, we know that the behavior occurred in the time interval $(t_{i-1}, t_i]$ for $i = 1, 2, \ldots, C$. The probability of the behavior occuring in the ith time interval is given by $F(t_i) - F(t_{i-1})$. Furthermore, we typically have "right-censored" data; that is, the observation period finishes at t_C and we know that the behavior of interest has not yet occurred for a number of individuals. This implies that it will occur in the interval (t_C, ∞), and the probability that this occurs is $P(T > t_C) = 1 - F(t_C)$.

Let f_i be the number of individuals whose behavior occurred in the ith time interval $(i = 1, \ldots, C)$ and f_{C+1} be the number of right-censored individuals (e.g., those individuals who have not made a trial purchase by t_C). The log-likelihood function associated with the sample data is given by

$$LL(r,\alpha\,|\,\text{data}) = \sum_{i=1}^{C} f_i \ln\bigl[F(t_i|r,\alpha) - F(t_{i-1}|r,\alpha)\bigr] \\ + f_{C+1}\ln\bigl[1 - F(t_C|r,\alpha)\bigr],\ \text{where } t_0 = 0.$$

Using standard numerical optimization software, we find the values of r and α that maximize this log-likelihood function; these are the maximum likelihood estimates of r and α.

4 The NBD Model

The NBD model results when we assume that

- the individual-level behavior of interest is a "count" variable (e.g., number of units of a product purchased in a unit time period) and can be characterized by the Poisson distribution with rate parameter λ, which we denote by $P(X = x|\lambda)$, and
- the values of λ are distributed across the population according to a gamma distribution, denoted by $g(\lambda)$.

The aggregate distribution of the behavior of interest, which we denote by $P(X = x)$, is obtained by weighting each $P(X = x|\lambda)$ by the likelihood of that value of λ occurring (i.e., $g(\lambda)$). This is formally denoted by

$$P(X = x) = \int_0^\infty P(X = x|\lambda)g(\lambda)d\lambda$$

4.1 Model Derivation

In order to derive the aggregate distribution associated with Poisson events at the individual-level and gamma heterogeneity, we must solve the following integral:

$$P(X = x) = \int_0^\infty \underbrace{\frac{\lambda^x e^{-\lambda}}{x!}}_{\text{Poisson}} \frac{\overbrace{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}^{\text{gamma}}}{\Gamma(r)}\, d\lambda$$

This is done in the following manner:

1. Combine terms and move all non-λ terms to the left of the integral sign. This gives us

$$P(X=x)=\frac{\alpha^r}{x!\,\Gamma(r)}\int_0^\infty \lambda^{x+r-1}e^{-\lambda(\alpha+1)}d\lambda$$

2. We therefore have to solve the definite integral

$$\int_0^\infty \lambda^{x+r-1}e^{-\lambda(\alpha+1)}d\lambda$$

   The "trick" is to transform the terms to the right of the integral sign into a known pdf, which integrates to 1. Looking closely at these terms, we see the heart of a gamma density with shape parameter $x+r$ and scale parameter $\alpha+1$. Multiplying the integral by $[\Gamma(r+x)/(\alpha+1)^{r+x}]/[(\alpha+1)^{r+x}/\Gamma(r)]$, we can write our expression for $P(X=x)$ as

$$P(X=x)=\frac{\alpha^r}{x!\,\Gamma(r)}\frac{\Gamma(r+x)}{(\alpha+1)^{r+x}}\int_0^\infty \underbrace{\frac{(\alpha+1)^{r+x}\lambda^{x+r-1}e^{-\lambda(\alpha+1)}}{\Gamma(r+x)}}_{\text{gamma pdf}}d\lambda$$

3. As the integrand is a gamma pdf, the definite integral, by definition, equals 1, and we therefore write the equation as

$$\begin{aligned}P(X=x)&=\frac{\alpha^r\Gamma(r+x)}{x!\,\Gamma(r)(\alpha+1)^{r+x}}\\&=\frac{\Gamma(r+x)}{\Gamma(r)x!}\left(\frac{\alpha}{\alpha+1}\right)^r\left(\frac{1}{\alpha+1}\right)^x\end{aligned}$$

   This is called the Negative Binomial Distribution, or NBD model.

   Since $x!=\Gamma(x+1)$, we sometimes see $\Gamma(r+x)/\Gamma(r)\,x!$ expressed as $\Gamma(r+x)/\Gamma(r)\Gamma(x+1)$. Alternatively, we sometimes see $\Gamma(r+x)/\Gamma(r)\,x!$ expressed as the binomial coefficient

$$\binom{r+x-1}{x}$$

4.2 Mean and Variance of the NBD

While the mean and variance of the NBD can be derived using standard expressions (e.g., $E(X) = \sum_{x=0}^{\infty} xP(X = x)$), a more elegant approach is to compute them *by conditioning*.

Mean of the NBD

To compute the mean by conditioning, we evaluate

$$E(X) = E_Y\big[E(X|Y)\big]$$

where $E_Y[\cdot]$ denotes expectation with respect to the distribution of Y (i.e., $\int E(X|Y = y)f(y)\,dy$. For the NBD, we have

$$E(X) = E_\lambda\big[E(X|\lambda)\big]$$

Conditional on λ, X is distributed Poisson, and the mean of the Poisson distribution is λ; therefore $E(X) = E(\lambda)$. The latent variable λ has a gamma distribution, and we know that the mean of the gamma distribution is $E(\lambda) = r/\alpha$. Therefore the mean of the NBD is

$$E(X) = \frac{r}{\alpha}$$

Variance of the NBD

We can derive the formula for the variance of X is a similar manner. To compute the variance by conditioning, we evaluate

$$\text{var}(X) = E_Y\big[\text{var}(X|Y)\big] + \text{var}_Y\big[E(X|Y)\big]$$

where $\text{var}_Y[\cdot]$ denotes variance with respect to the distribution of Y. For the NBD, we have

$$\text{var}(X) = E_\lambda\big[\text{var}(X|\lambda)\big] + \text{var}_\lambda\big[E(X|\lambda)\big]$$

Conditional on λ, X is distributed Poisson, and the variance of the Poisson distribution is λ. Therefore we have

$$\text{var}(X) = E(\lambda) + \text{var}(\lambda)$$

We know that the variance of the gamma distribution is $\mathrm{var}(\lambda) = r/\alpha^2$. Therefore the variance of the NBD is

$$\mathrm{var}(X) = \frac{r}{\alpha} + \frac{r}{\alpha^2}$$

4.3 Estimating Model Parameters

In order to apply the NBD model, we must first develop estimates of the two model parameters, r and α, from the given sample data. Three methods are at the modeler's disposal: maximum likelihood, method of moments, and means and zeroes.

Approach 1: Maximum Likelihood

Let x_i be the number of events for individual i ($i = 1, \ldots, N$) in the observation period. By definition, the likelihood function is the joint density of the observed data. Assuming the x_i are independent, this is the product of NBD probabilities for each x_i. Equivalently, the log-likelihood function is given by

$$LL(r, \alpha \,|\, \mathrm{data}) = \sum_{i=1}^{N} \ln\bigl[P(X = x_i|r, \alpha)\bigr]$$

Using standard numerical optimization software, we find the values of r and α that maximize this log-likelihood function; these are the maximum likelihood estimates of r and α.

Let $x^* = \max(x_1, x_2, \ldots, x_N)$ and f_j the number of $x_i = j$. We can write the log-likelihood function as

$$LL(r, \alpha \,|\, \mathrm{data}) = \sum_{x=0}^{x^*} f_x \ln\bigl[P(X = x|r, \alpha)\bigr]$$

Censored Data: In many cases, the data available for model estimation are of the form

x	0	1	2	3+
f_x	814	128	22	7

These data are *censored* — we know 7 panelists made at least 3 purchases, but do not know the exact number of purchases they made. It is possible to estimate the model parameters using maximum likelihood methods by modifying the log-likelihood function in the following manner. Let x^+ denote the censoring point in the data — 3 in the above example. The log-likelihood function can be written as

$$\begin{aligned} LL(r,\alpha\,|\,\text{data}) &= \sum_{x=0}^{x^+-1} f_x \ln\bigl[P(X=x|r,\alpha)\bigr] + f_{x^+}\ln\bigl[P(X\ge x^+|r,\alpha)\bigr] \\ &= \sum_{x=0}^{x^+-1} f_x \ln\bigl[P(X=x|r,\alpha)\bigr] + f_{x^+}\ln\left[1-\sum_{x=0}^{x^+-1} P(X=x|r,\alpha)\right] \end{aligned}$$

Approach 2: Method of Moments

Another approach to estimating the parameters of a model from a particular dataset is to use the *method of moments*, which sees us equating the sample moments with their population counterparts. (As the NBD has two parameters, we focus on the first two moments — the mean and variance.) Denoting the sample mean by $\bar{x}$ and the sample variance by s^2, we have

$$\bar{x} = r/\alpha \tag{1}$$

$$s^2 = r/\alpha + r/\alpha^2 \tag{2}$$

Substituting (1) into (2), we get $s^2 = \bar{x} + \bar{x}/\alpha$, which implies

$$\hat{\alpha} = \frac{\bar{x}}{s^2 - \bar{x}}$$

From (1), it follows that

$$\hat{r} = \hat{\alpha}\bar{x} \tag{3}$$

Approach 3: Means and Zeros

Just as the method of moments sees us equating two sample-based moments with their population counterparts, the method of "means and zeros" sees us equating the sample mean and sample proportion of zeros with their population counterparts.

Now the proportion of zeros, as predicted by the NBD, is

$$P(X=0)=\left(\frac{\alpha}{\alpha+1}\right)^r \tag{4}$$

Let P_0 be the sample proportion of zeros, and $\bar{x}$ the sample mean. From (1) we have $r=\alpha\bar{x}$. Substituting this into (4) and equating with the sample proportion of zeros, we have

$$P_0=\left(\frac{\alpha}{\alpha+1}\right)^{\alpha\bar{x}}$$

We solve this for $\hat{\alpha}$ — using a computer — and estimate $\hat{r}$ using (3).

4.4 Computing NBD Probabilities

Given r and α, NBD probabilities can be calculated directly by evaluating the standard NBD formula, i.e.,

$$P(X=x)=\frac{\Gamma(r+x)}{\Gamma(r)x!}\left(\frac{\alpha}{\alpha+1}\right)^r\left(\frac{1}{\alpha+1}\right)^x$$

This assumes it is easy to numerically evaluate $\Gamma(\cdot)$.

Alternatively, the recursive computation of NBD probabilities is straightforward, using the following *forward recursion* formula from $P(X=0)$:

$$P(X=x)=\begin{cases}\left(\dfrac{\alpha}{\alpha+1}\right)^r & x=0\\[2ex] \dfrac{r+x-1}{x(\alpha+1)}\times P(X=x-1) & x\geq 1\end{cases}$$

4.5 The NBD for a Non-Unit Time Period

The preceding discussion and development of the NBD assumes that the length of our observation period is one unit of time. What is the form of the NBD applied to an observation period of length t time units?

Let $X(t)$ be the number of events occuring in an observation period of length t time units. If, for a unit time period, the distribution of events *at the individual-level* is Poisson with rate parameter λ, $X(t)$ has a Poisson distribution with rate parameter λt. Therefore, the expression for NBD probabilities for a time period of length t is

$$\begin{aligned}
P(X(t)=x) &= \int_0^\infty \underbrace{\frac{(\lambda t)^x e^{-\lambda t}}{x!}}_{\text{Poisson}} \frac{\overbrace{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}^{\text{gamma}}}{\Gamma(r)}\, d\lambda \\
&= \frac{\alpha^r t^x}{x!\,\Gamma(r)} \int_0^\infty \lambda^{x+r-1} e^{-\lambda(\alpha+t)} d\lambda \\
&= \frac{\alpha^r t^x}{x!\,\Gamma(r)} \frac{\Gamma(r+x)}{(\alpha+t)^{r+x}} \int_0^\infty \frac{(\alpha+t)^{r+x}\lambda^{x+r-1}e^{-\lambda(\alpha+t)}}{\Gamma(r+x)} d\lambda \\
&= \frac{\Gamma(r+x)}{\Gamma(r)x!} \left(\frac{\alpha}{\alpha+t}\right)^r \left(\frac{t}{\alpha+t}\right)^x
\end{aligned}$$

The mean and variance of $X(t)$ can easily be determined by conditioning. $E\big[X(t)\big] = E\big\{E\big[X(t)|\lambda\big]\big\}$. Since $X(t)$ is distributed Poisson with parameter λt, it follows that $E\big[X(t)\big] = E(\lambda t) = tE(\lambda) = rt/\alpha$. Similarly, the variance of $X(t)$ is given by:

$$\begin{aligned}
\text{var}\big[X(t)\big] &= E_\lambda\big[\text{var}(X(t)|\lambda)\big] + \text{var}_\lambda\big[E(X(t)|\lambda)\big] \\
&= E(\lambda t) + \text{var}(\lambda t) \\
&= tE(\lambda) + t^2\text{var}(\lambda) \\
&= \frac{rt}{\alpha} + \frac{rt^2}{\alpha^2}
\end{aligned}$$

The associated formula for computing NBD probability using *forward recursion* from $P(X=0)$ is

$$P(X=x) = \begin{cases} \left(\dfrac{\alpha}{\alpha+t}\right)^r & x=0 \\ \\ \dfrac{(r+x-1)t}{x(\alpha+t)} \times P(X=x-1) & x \geq 1 \end{cases}$$

5 The Beta-Binomial Model

The beta-binomial model results when we assume that

- the individual-level behavior of interest reflects the outcome of a series of independent choices (e.g., the number of times a target product is purchased given n category purchases) and can be characterized by the binomial distribution with parameter p, which we denote by $P(X = x|n,p)$, and
- the values of p are distributed across the population according to a beta distribution, denoted by $g(p)$.

The aggregate distribution of the behavior of interest, denoted by $P(X = x|n)$, is obtained by weighting each $P(X = x|n,p)$ by the likelihood of that value of p occurring (i.e., $g(p)$). This is formally denoted by

$$P(X = x|n) = \int_0^1 P(X = x|n,p)g(p)dp$$

5.1 Model Derivation

In order to derive the aggregate distribution associated with a binomial choice process at the individual-level and beta heterogeneity, we must solve the following integral:

$$P(X = x) = \int_0^1 \underbrace{\binom{n}{x} p^x (1-p)^{n-x}}_{\text{binomial}} \, \underbrace{\frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}}_{\text{beta}} \, dp$$

This is done in the following manner:

1. Combine terms and move all non-p terms to the left of the integral sign. This gives us

$$P(X = x) = \binom{n}{x} \frac{1}{B(\alpha,\beta)} \int_0^1 p^{\alpha+x-1}(1-p)^{\beta+n-x-1}dp$$

2. We therefore have to solve the definite integral

$$\int_0^1 p^{\alpha+x-1}(1-p)^{\beta+n-x-1}dp$$

The "trick" is to transform the terms to the right of the integral sign into a known pdf, which integrates to 1. Looking closely at this, we see that its structure mirrors the density of the beta distribution with parameters $\alpha + x$ and $\beta + n - x$. Multiplying the integral by $B(\alpha + x, \beta + n - x)/B(\alpha + x, \beta + n - x)$, we can write our expression for $P(X = x)$ as

$$P(X = x) = \binom{n}{x} \frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)} \times \int_0^1 \underbrace{\frac{1}{B(\alpha + x, \beta + n - x)} p^{\alpha+x-1}(1-p)^{\beta+n-x-1}}_{\text{beta pdf}} \, dp$$

3. As the integrand is a beta pdf, the definite integral, by definition, equals 1, and we therefore write the equation as

$$P(X = x) = \binom{n}{x} \frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)}$$

This is called the beta-binomial (or BB) model.

5.2 Mean and Variance of the Beta-Binomial

While the mean and variance of the BB can be derived using standard expressions (e.g., $E(X) = \sum_{x=0}^{n} xP(X = x)$), a more elegant approach is to compute them *by conditioning*.

Mean of the BB

To compute the mean by conditioning — see section 4.2 — we evaluate

$$E(X) = E_p\big[E(X|p)\big]$$

where $E_p[\cdot]$ denotes expectation with respect to the distribution of p. Conditional on p, X is distributed binomial, and the mean of the binomial distribution is np; therefore $E(X) = E(np)$. Since n is a constant, this is equivalent to $E(X) = nE(p)$. As the latent variable p has a beta distribution, and we know that the mean of the beta distribution is $E(p) = \alpha/(\alpha+\beta)$, it follows that the mean of the beta-binomial distribution is

$$E(X) = \frac{n\alpha}{\alpha+\beta}$$

Variance of the BB

We can derive the formula for the variance of X is a similar manner — see section 4.2 — we evaluate

$$\text{var}(X) = E_p\big[\text{var}(X|p)\big] + \text{var}_p\big[E(X|p)\big]$$

where $\text{var}_p[\cdot]$ denotes variance with respect to the distribution of p. Conditional on p, X is distributed binomial, and the variance of the binomial distribution is $np(1-p)$. Therefore we have

$$\begin{aligned}\text{var}(X) &= E\big[np(1-p)\big] + \text{var}(np)\\ &= nE(p) - nE(p^2) + n^2\text{var}(p)\end{aligned}$$

We know that the variance of the beta distribution is $\text{var}(p) = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. Recalling that $\text{var}(X) = E(X^2) - E(X)^2$, it follows that $E(p^2) = \text{var}(p) + E(p)^2$. Substituting the expressions for $E(p), E(p^2)$, and $\text{var}(p)$ into the above equation and simplifying, we arrive at the following expression for the variance of the beta-binomial distribution:

$$\text{var}(X) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

5.3 Estimating Model Parameters

In order to apply the BB model, we must first develop estimates of the two model parameters, α and β, from the given sample data. Two methods are at the modeler's disposal: method of moments and maximum likelihood. Let x_i be the number of successes out of n_i trials for individual i ($i = 1, \ldots, N$) in the observation period.

Approach 1: Maximum Likelihood

By definition, the likelihood function is the joint density of the observed data. Assuming the x_i are independent, this is the product of BB probabilities for each x_i, given n_i. The log-likelihood function is therefore

$$LL(\alpha, \beta \,|\, \text{data}) = \sum_{i=1}^{N} \ln\bigl[P(X = x_i | n_i, \alpha, \beta)\bigr]$$

Using standard numerical optimization software, we find the values of α and β that maximize this log-likelihood function; these are the maximum likelihood estimates of α and β.

In many applications of the BB, $n_i = n \; \forall \, i$. Let f_x be the number of $x_i = x$; note that $\sum_{x=0}^{n} f_x = N$. We can write the log-likelihood function as

$$LL(\alpha, \beta \,|\, \text{data}) = \sum_{x=0}^{n} f_x \ln\bigl[P(X = x | n, \alpha, \beta)\bigr]$$

Approach 2: Method of Moments

For the case of $n_i = n \; \forall \, i$, another approach to estimating the parameters of the BB model is the *method of moments*, which sees us equating the sample moments with their population counterparts. (As the BB has two parameters, we focus on the first two moments — the mean and variance.) Denoting the sample mean by $\bar{x}$ and the sample variance by s^2, we have

$$\bar{x} = \frac{n\alpha}{\alpha + \beta} \tag{5}$$

$$s^2 = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{6}$$

Solving (5) for β, we get

$$\hat{\beta} = \frac{\hat{\alpha}(n - \bar{x})}{\bar{x}} \tag{7}$$

To arrive at the method of moments estimator for α, we first note that, from (5), $\alpha + \beta = n\alpha/\bar{x}$. Substituting this expression for $\alpha + \beta$, along with that for β from (7), into (6), we solve for α. Performing the requisite algebra, we get

$$\hat{\alpha} = \frac{\bar{x}[\bar{x}(n-\bar{x}) - s^2]}{s^2 n - \bar{x}(n-\bar{x})} \tag{8}$$

6 The Dirichlet-Multinomial Model

The Dirichlet-multinomial model results when we assume that

- the individual-level behavior of interest reflects the vector of outcomes of a series of independent choices (e.g., the number of times brands A, B, C, and D are each chosen given n category purchases) and can be characterized by the multinomial distribution with parameter vector $\mathbf{p}$, which we denote by $P(\mathbf{X} = \mathbf{x}|n, \mathbf{p})$, and
- the values of $\mathbf{p}$ are distributed across the population according to a Dirichlet distribution, denoted by $g(\mathbf{p})$.

The aggregate distribution of the behavior of interest, denoted by $P(\mathbf{X} = \mathbf{x}|n)$, is obtained by weighting each $P(\mathbf{X} = \mathbf{x}|n, \mathbf{p})$, by the likelihood of that value of the vector $\mathbf{p}$ occurring (i.e., $g(\mathbf{p})$). This is denoted by

$$P(\mathbf{X} = \mathbf{x}|n) = \int P(\mathbf{X} = \mathbf{x}|n, \mathbf{p}) g(\mathbf{p}) d\mathbf{p}$$

More formally, we should note that since the elements of any $\mathbf{p}$, of length k, sum to 1, the integration is actually performed with respect to the $k-1$ variables $p_1, p_2, \ldots, p_{k-1}$, where the integration limits are $[0,1], [0, 1-p_1], \ldots, [0, 1-\sum_{j=1}^{k-2} p_j]$, respectively.

6.1 Model Derivation

In order to derive the aggregate distribution associated with a multinomial choice process at the individual level and Dirichlet heterogeneity, we must solve the following integral:

$$P(\mathbf{X}=\mathbf{x}) = \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-\sum_{j=1}^{k-2} p_j} \binom{n}{x_1, \ldots, x_k} \left(\prod_{j=1}^{k-1} p_j^{x_j} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{n - \sum_{j=1}^{k-1} x_j}$$
$$\times \frac{\Gamma(S)}{\prod_{j=1}^{k} \Gamma(a_j)} \left(\prod_{j=1}^{k-1} p_j^{a_j - 1} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{a_k - 1} \, dp_{k-1} \cdots dp_2 \, dp_1$$

This is done in the following manner:

1. Combine terms and move all non-p_j terms to the left of the integral signs. This gives us

$$P(\mathbf{X}=\mathbf{x}) = \binom{n}{x_1, \ldots, x_k} \frac{\Gamma(S)}{\prod_{j=1}^{k} \Gamma(a_j)} \times$$
$$\int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-\sum_{j=1}^{k-2} p_j} \left(\prod_{j=1}^{k-1} p_j^{a_j + x_j - 1} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{a_k + n - (\sum_{j=1}^{k-1} x_j) - 1} \, dp_{k-1} \cdots dp_2 \, dp_1$$

2. We therefore have to solve the definite integral

$$\int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-\sum_{j-1}^{k-2} p_j} \left(\prod_{j=1}^{k-1} p_j^{a_j + x_j - 1} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{a_k + n - (\sum_{j=1}^{k-1} x_j) - 1} \, dp_{k-1} \cdots dp_2 \, dp_1$$

The "trick" is to transform the terms to the right of the integral sign into a known pdf.

3. Looking closely at this, we see that its structure mirrors the density of the Dirichlet distribution with parameters $a_j + x_j$ ($j = 1, \ldots, k$); all that is missing is a $\Gamma(S+n) / \prod_{j=1}^{k} \Gamma(a_j + x_j)$ term. We can therefore write our expression for $P(\mathbf{X}=\mathbf{x})$ as

$$P(\mathbf{X}=\mathbf{x}) = \binom{n}{x_1, \ldots, x_k} \frac{\Gamma(S)}{\prod_{j=1}^{k} \Gamma(a_j)} \frac{\prod_{j=1}^{k} \Gamma(a_j + x_j)}{\Gamma(S+n)} \times$$
$$\int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-\sum_{j=1}^{k-2} p_j} \frac{\Gamma(S+n)}{\prod_{j=1}^{k} \Gamma(a_j + x_j)} \times$$
$$\left(\prod_{j=1}^{k-1} p_j^{a_j + x_j - 1} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{a_k + (n - \sum_{j=1}^{k-1} x_j) - 1} \, dp_{k-1} \cdots dp_2 \, dp_1$$

4. As the integrand is a Dirichlet pdf, the definite integral, by definition, equals 1, and we therefore write the equation as

$$P(\mathbf{X}=\mathbf{x}) = \binom{n}{x_1,\ldots,x_k} \frac{\Gamma(S)}{\prod_{j=1}^{k}\Gamma(a_j)} \frac{\prod_{j=1}^{k}\Gamma(a_j+x_j)}{\Gamma(S+n)}$$

This is called the Dirichlet-multinomial (or DM) model.

6.2 Mean and Variance of the Dirichlet-Multinomial

The mean of the DM can easily be derived *by conditioning*—see section 4.2. To do so, we evaluate

$$E(\mathbf{X}) = E_{\mathbf{p}}\big[E(\mathbf{X}|\mathbf{p})\big]$$

where $E_{\mathbf{p}}[\cdot]$ denotes expectation with respect to the distribution of the vector $\mathbf{p}$. Conditional on $\mathbf{p}$, $\mathbf{X}$ is distributed multinomial, and the mean of the multinomial distribution is $n\mathbf{p}$; therefore $E(\mathbf{X}) = E(n\mathbf{p})$. Since n is a scalar constant, this is equivalent to $E(\mathbf{X}) = nE(\mathbf{p})$. As the latent vector $\mathbf{p}$ has a Dirichlet distribution, and we know that the mean of the Dirichlet distribution is $E(\mathbf{X}) = \mathbf{a}/S$, with $E(X_j) = a_j/S$. It follows that the mean of the Dirichlet-multinomial is

$$E(\mathbf{X}) = \frac{n}{S}\mathbf{a}\ ,\ \text{with } E(X_j) = \frac{na_j}{S}$$

The derivation of the variance-covariance of the Dirichlet-multinomial is more complex and we therefore present the result without derivation:

$$\begin{aligned}\text{var}(X_j) &= \frac{na_j(S-a_j)(S+n)}{S^2(S+1)}\\ \text{cov}(X_j,X_{j'}) &= \frac{-na_ja_{j'}(S+n)}{S^2(S+1)}\end{aligned}$$

This can be re-written as:

$$\text{cov}(X_j,X_{j'}) = n\frac{a_j}{S}\left(\delta_{j=j'} - \frac{a_{j'}}{S}\right)\left(\frac{S+n}{S+1}\right)$$

where $\delta_{j=j'}$ is the Kronecker delta , defined as

$$\delta_{j=j'} = \begin{cases} 1 & \text{if } j = j' \\ 0 & \text{otherwise} \end{cases}$$

Let $\bar{\mathbf{p}}$ be the mean vector of the Dirichlet distribution with jth element $\bar{p}_j = a_j/S$. We therefore have

$$\text{cov}(X_j, X_{j'}) = n\bar{p}_j(\delta_{j=j'} - \bar{p}_{j'}) \left(\frac{S+n}{S+1}\right)$$

and can therefore express the variance-covariance of the Dirichlet-multinomial in matrix form as

$$\text{var}(\mathbf{X}) = \left(\frac{S+n}{S+1}\right) n[\text{Diag}(\bar{\mathbf{p}}) - \bar{\mathbf{p}}\bar{\mathbf{p}}']$$

6.3 Estimating Model Parameters

In order to apply the DM model, we must first develop estimates of its parameter vector $\mathbf{a}$, from the given sample data. Two methods are at the modeler's disposal: maximum likelihood and method of moments. Let $\mathbf{x}_i$ be the vector of purchases made by household i $(i = 1, \ldots , N)$ across the k brands, and n_i the number of category purchases $(n_i = \sum_{j=1}^{k} x_{ij})$; x_{ij} denotes the number of times outcome j occurs in n_i independent trials.

Approach 1: Maximum Likelihood

By definition likelihood function is the joint density of the observed data. Assuming the observations are independent, this is the product of the DM probabilities for each $\mathbf{x}_i$. The log-likelihood function is therefore

$$LL(\mathbf{a} \,|\, \text{data}) = \sum_{i=1}^{N} \ln\big[P(\mathbf{X} = \mathbf{x}_i | n_i, \mathbf{a})\big]$$

Using standard numerical optimization software, we find the value of the parameter vector $\mathbf{a}$ that maximizes this log-likelihood function; this is the maximum likelihood estimate of $\mathbf{a}$.

Approach 2: Method of Moments

For the case of $n_i = n \; \forall \, i$, another approach to estimating the parameters of the DM model is the *method of moments*.

Let us denote the sample mean vector by $\bar{\mathbf{x}}$, the jth element of which is

$$\bar{x}_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij} \;, \qquad j = 1, \ldots, k$$

Equating the sample mean vector with its population counterpart, we have

$$\bar{x}_j = \frac{n a_j}{S} \tag{9}$$

Given an estimate of S, it follows that

$$\hat{a}_j = \frac{\hat{S}\bar{x}_j}{n}, \qquad j = 1, \ldots, k$$

We therefore need an estimate of S; there are several means of doing this, two of which are:

- Let us denote the sample variance of X_j by s_j^2. Equating this with its population counterpart, we have

  $$s_j^2 = \frac{n a_j (S - a_j)(S + n)}{S^2 (S + 1)} \tag{10}$$

  From (9) we have $\bar{x}_j / n = a_j / S$. Substituting this into (10), we have

  $$s_j^2 = \frac{n \bar{x}_j (n - \bar{x}_j)(S + n)}{n(S + 1)}$$

  Solving this for S, we get

  $$\hat{S} = \frac{n\left[\bar{x}_j (n - \bar{x}_j) - s_j^2\right]}{n s_j^2 - \bar{x}_j (n - \bar{x}_j)}$$

- Recall that the variance-covariance of the Dirichlet-multinomial is

  $$\text{var}(\mathbf{X}) = \left(\frac{S + n}{S + 1}\right) n[\text{Diag}(\bar{\mathbf{p}}) - \bar{\mathbf{p}}\bar{\mathbf{p}}']$$

where $\bar{\mathbf{p}}$ is the mean vector of the Dirichlet distribution with jth element $\bar{p}_j = a_j/S$. Looking closely at this expression, we see that it is $(S+n)/(S+1) \times$ the variance-covariance matrix of the multinomial distribution computed with $\bar{\mathbf{p}}$. This leads to the following procedure for developing an estimate of S:

1. Let $\widehat{\mathbf{\Sigma}}_{DM}$ be the *sample* variance-covariance matrix generated using the observed data, and $\widehat{\mathbf{\Sigma}}_M$ the multinomial variance-covariance matrix generated using $\bar{\mathbf{x}}/n$ as our estimate of $\bar{\mathbf{p}}$.

2. Dropping the kth row and column of each matrix, we get $\widehat{\mathbf{\Sigma}}'_{DM}$ and $\widehat{\mathbf{\Sigma}}'_M$; both matrices are of order $(k-1) \times (k-1)$. (We do this as the rank of both $\widehat{\mathbf{\Sigma}}_{DM}$ and $\widehat{\mathbf{\Sigma}}_M$ is $k-1$.) Recall from basic matrix algebra that, for scalar b and $n \times n$ matrix $\mathbf{A}$, $|b\mathbf{A}| = b^n|\mathbf{A}|$, where $|\cdot|$ denotes the determinant. It follows that

$$|\widehat{\mathbf{\Sigma}}'_{DM}| = \left(\frac{S+n}{S+1}\right)^{k-1} |\widehat{\mathbf{\Sigma}}'_M|$$

   Let

$$\gamma = \frac{|\widehat{\mathbf{\Sigma}}'_{DM}|}{|\widehat{\mathbf{\Sigma}}'_M|}$$

3. Solving

$$\gamma = \left(\frac{S+n}{S+1}\right)^{k-1}$$

   for S, we get

$$\hat{S} = \frac{n - \sqrt[k-1]{\gamma}}{\sqrt[k-1]{\gamma} - 1}$$

The second approach is probably more desirable as it develops an estimate of S from all the variances and covariances, as opposed to the variance of *only* one variable — for any $j = 1, \ldots, k$.

7 Empirical Bayes Methods

At the heart of any probability modeling effort is the assumption that the observed individual-level behavior x is the realization of a random process with density $f(x|\theta)$, which has unknown parameter(s) θ. By assuming a particular distribution for θ, we are able to derive an aggregate-level model without specific knowledge of any given individual's latent parameter(s), and therefore solve the management problem motivating the modeling exercise.

In many cases, however, we are interested in estimating a given individual's latent "trait" (i.e., θ). This may be because we wish to rank the individuals on the basis of their true underlying behavioral tendency or because we wish to forecast their behavior in a future period. In either case, the challenge is to make interferences regarding θ, given the individual's observed behavior x. In order to address this problem, we make use of Bayes theorem.

Definitions

- The **prior distribution** $g(\theta)$ represents our opinion about the possible values θ can take on, prior to collecting any information about the specific individual.
- The **model distribution** $f(x|\theta)$ is the density function for the observed data, given a specific value of the latent parameter θ. (Note that this is the same as the likelihood function $L(\theta|x)$ and consequently many textbooks on Bayesian methods use this alternative terminology and notation.)
- The **marginal distribution** of x is given by

$$f(x) = \int f(x|\theta)g(\theta)\,d\theta$$

- The **posterior distribution** $g(\theta|x)$ is the conditional distribution of θ, given the observed data x. It represents our updated opinion about the possible values θ can take on, now that we have some information x about the specific individual.

According to Bayes theorem, the posterior distribution is computed as

$$\begin{aligned} g(\theta|x) &= \frac{f(x|\theta)g(\theta)}{\int f(x|\theta)g(\theta)\,d\theta} \\ &= \frac{f(x|\theta)g(\theta)}{f(x)} \end{aligned}$$

This is sometimes expressed as posterior $\propto$ likelihood × prior.

In applying Bayes theorem to the types of problems noted above, we use the mixing distribution, which captures the heterogeneity in the individual-level latent variables, as the prior distribution. Using the estimated parameters of this mixing distribution, along with the observed data x, we arrive at the posterior distribution using the above formula.

Formally, this approach is known as parametric empirical Bayes:

- *parametric* because we specific a parametric distribution (e.g, gamma, beta) for the prior. (Alternatively, some modelers use a *nonparametric* prior distribution, but this is beyond the scope of this note.)
- *empirical* because we estimate the parameters of this prior distribution using the sample data, as opposed to using analyst-specified values as in the case of "traditional" Bayesian analysis.

In applied marketing settings, we very rarely focus on the posterior distribution as an end result. Rather we may:

1. Compute the **predictive distribution** $f(y|x)$, which is the distribution of a new behavior y given the observed data x. For example, what is the distribution of purchases in a future period for an individual who made x purchases in the current period?
2. Compute the **conditional expectation** of the future behavior, given the observed data, i.e., $E(y|x)$. (This is the mean of the predictive distribution.)
3. Compute the **conditional expectation** of the latent variable θ, given the observed data, i.e., $E(\theta|x)$.

7.1 The NBD Model

Consider a behavior (e.g., product purchasing) that can be characterized by the NBD model (i.e., Poisson counts at the individual-level with gamma heterogeneity). A model has been calibrated using data from Period 1 (of unit length) and we are interested in predicting individual-level behavior in a non-overlapping Period 2 (also of unit length). Let the random variables X_1 and X_2 be the counts for Periods 1 and 2, respectively.

- If we knew nothing about an individual's purchasing in Period 1, what would be our best guess as to the distribution of the individual's buying rate, λ? Our best guess would be that the individual's buying rate is distributed according to the population gamma distribution with parameters r and α. Consequently, $E(\lambda) = r/\alpha$ and therefore $E(X_2) = r/\alpha$.
- If we know that the individual made x purchases in Period 1, we may be tempted to say that this individual will make x purchases in Period 2; i.e., $E(X_2|X_1 = x) = x$. However, this does not take into account the assumed stochastic nature of buying behavior. Moreover, it provides no insight into the distribution of the individual's buying rate.

Therefore, our objective is to derive the distribution of the individual's buying rate, λ, taking into consideration the fact that he purchased x units in Period 1.

Applying Bayes theorem, we have

$$\begin{aligned} g(\lambda|x) &= \frac{\overbrace{\frac{\lambda^x e^{-\lambda}}{x!}}^{P(X_1=x|\lambda)} \; \overbrace{\frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)}}^{g(\lambda|r,\alpha)}}{\underbrace{\frac{\Gamma(r+x)}{\Gamma(r)\,x!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^x}_{P(X_1=x)}} \\ &= \frac{(\alpha+1)^{r+x} \lambda^{r+x-1} e^{-\lambda(\alpha+1}}{\Gamma(r+x)} \\ &= \text{gamma}(r+x, \alpha+1) \end{aligned}$$

That is, the updated distribution of λ, assuming $X_1 = x$ and a gamma prior distribution, is itself gamma with parameters $r + x$ and $\alpha + 1$.

It follows that the distribution of X_2, conditional on $X_1 = x_1$ is

$$P(X_2|X_1 = x_1) = \frac{\Gamma(r + x_1 + x_2)}{\Gamma(r + x_1)\, x_2!} \left(\frac{\alpha + 1}{\alpha + 2}\right)^{r+x_1} \left(\frac{1}{\alpha + 2}\right)^{x_2}$$

Also note that the expected value of X_2, conditioned on the fact that $X_1 = x$ (i.e., the conditional expectation of X_2) is

$$E(X_2|X_1 = x) = \frac{r + x}{\alpha + 1}$$

This can be written as

$$E(X_2|X_1 = x) = \left(\frac{\alpha}{\alpha + 1}\right) \frac{r}{\alpha} + \left(\frac{1}{\alpha + 1}\right) x$$

which implies that the expectation of future behavior, conditional on observed behavior, is a weighted average of the observed value (x) and the population mean (r/α). Therefore, the "regression to the mean" phenomenon applies to NBD-based conditional expectations. We note that the larger the value of α, the greater the regression to the mean effect.

7.2 The Beta-Binomial Model

Consider a phenomenon (e.g., brand choice) that can be characterized by the BB model (i.e., a binomial "choice" process at the individual-level with beta heterogeneity). A model has been calibrated using data of the form (x_i, n_i), $i = 1, \ldots, N$, where x_i is the number of times individual i chooses the focal brand from a total of n_i purchasing occasions. We are interested in estimating the individual's underlying choice probability, p.

- If we knew nothing about an individual's choice behavior, what would be our best guess as to the distribution of the individual's choice probability, p? Our best guess would be that p is distributed according to the population beta distribution with parameters α and β. Consequently, $E(p) = \alpha/(\alpha + \beta)$.

- If we know that the individual chose the focal brand x out of n times, we may be tempted to say that our best guess of this individual's choice probability is x/n. However, this does not take into account the assumed stochastic nature of the choice process. Moreover, it provides no insight into the distribution of the individual's choice probability.

Therefore, our objective is to derive the distribution of the individual's choice probability, p, taking into consideration the fact that he chose the brand of interest x out of n times.

Applying Bayes theorem, we have

$$\begin{aligned} g(p|x,n) &= \frac{\overbrace{\binom{n}{x} p^x (1-p)^{n-x}}^{P(X=x|n,p)} \overbrace{\frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}}^{g(p|\alpha,\beta)}}{\underbrace{\binom{n}{x}\frac{B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}}_{P(X=x|n)}} \\ &= \frac{1}{B(\alpha+x,\beta+n-x)} p^{\alpha+x-1}(1-p)^{\beta+n-x-1} \\ &= \text{beta}(\alpha+x,\beta+n-x) \end{aligned}$$

That is, the updated distribution of p, given x and n and a beta prior distribution, is itself beta with parameters $\alpha + x$ and $\beta + n - x$. Therefore, the expected value of p, conditional on x and n (i.e., the conditional expectation of p) is

$$E(p|x,n) = \frac{\alpha+x}{\alpha+\beta+n}$$

This can be written as

$$E(p|x,n) = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta} + \left(\frac{n}{\alpha+\beta+n}\right)\frac{x}{n}$$

This is a weighted average of the predictions based on the observed choice probability (x/n) and the population mean $(\alpha/(\alpha+\beta))$. The larger the value of $\alpha+\beta$, relative to n, the greater the regression to the mean effect.

It follows that the distribution of X^*, the number of times the brand is chosen out of n^* purchase occasions, conditional on $X = x$ is

$$P(X^* = x^*|X = x, n, n^*) = \binom{n^*}{x^*} \frac{B(\alpha + x + x^*, \beta + n - x + n^* - x^*)}{B(\alpha + x, \beta + n - x)}$$

The expected value of X^*, conditional on x and n (i.e., the conditional expectation of X^*) is

$$E(X^*|x, n, n^*) = n^* \frac{\alpha + x}{\alpha + \beta + n}$$

7.3 The Dirichlet-Multinomial Model

Consider a phenomenon (e.g., brand choice) that can be characterized by the Dirichlet-multinomial model (i.e., a multinomial "choice" process at the individual-level with Dirichlet heterogeneity). A model has been calibrated using data of the form $(\mathbf{x}_i, n_i)$, $i = 1, \ldots, N$, where $\mathbf{x}_i$ is individual i's vector of purchases across k brands and $n_i = \sum_{j=1}^{k} x_{ij}$ is the total number of purchase occasions for this individual. We are interested in estimating the individual's underlying choice probability vector, $\mathbf{p}$.

- If we knew nothing about an individual's choice behavior, what would be our best guess as to the distribution of the individual's choice probability vector, $\mathbf{p}$? Our best guess would be that $\mathbf{p}$ is distributed according to the population Dirichlet distribution with parameters a_j, $j = 1, \ldots, k$ and $S = \sum_{j=1}^{k} a_j$. Consequently, $E(p_j) = a_j/S$.
- If we know the individual's purchase vector, $\mathbf{x}_i$, we may be tempted to say that our best guess of this individual's choice probability vector is $\mathbf{x}_i/n_i$. However, this does not take into account the assumed stochastic nature of the choice process. Moreover, it provides no insight into the distribution of the individual's choice probability vector.

Therefore, our objective is to derive the distribution of the individual's choice probability vector, $\mathbf{p}$, taking into consideration his purchases given by $\mathbf{x}$.

According to Bayes theorem, we have:

$$g(\mathbf{p}|\mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x}|n, \mathbf{p})g(\mathbf{p}|\mathbf{a})}{P(\mathbf{X} = \mathbf{x}|\mathbf{a}, n)}$$

Substituting the relevant expressions, we have

$$g(\mathbf{p}|\mathbf{x}) = \binom{n}{x_1, \ldots, x_k} \left(\prod_{j=1}^{k-1} p_j^{x_j}\right) \left(1 - \sum_{j=1}^{k-1} p_j\right)^{n - \sum_{j=1}^{k-1} x_j} \times$$
$$\frac{\Gamma(S)}{\prod_{j=1}^{k} \Gamma(a_j)} \left(\prod_{j=1}^{k-1} p_j^{a_j - 1}\right) \left(1 - \sum_{j=1}^{k-1} p_j\right)^{a_k - 1} \Bigg/$$
$$\binom{n}{x_1, \ldots, x_k} \frac{\Gamma(S)}{\prod_{j=1}^{k} \Gamma(a_j)} \frac{\prod_{j=1}^{k} \Gamma(a_j + x_j)}{\Gamma(S+n)}$$

Simplifying the above expression, we get

$$\begin{aligned} g(\mathbf{p}|\mathbf{x}, n) &= \frac{\Gamma(S+n)}{\prod_{j=1}^{k} \Gamma(a_j + x_j)} \left(\prod_{j=1}^{k-1} p_j^{a_j + x_j - 1}\right) \left(1 - \sum_{j=1}^{k-1} p_j\right)^{a_k + (n - \sum_{j=1}^{k-1} x_j) - 1} \\ &= \text{Dirichlet}(\mathbf{a} + \mathbf{x}) \end{aligned}$$

That is, the updated distribution of $\mathbf{p}$, given $\mathbf{x}$ and a Dirichlet prior distribution, is itself Dirichlet with parameter vector $\mathbf{a} + \mathbf{x}$. Therefore, the expected value of p_j, conditional on $\mathbf{x}$ (i.e., the conditional expectation of p_j) is

$$E(p_j|\mathbf{x}) = \frac{a_j + x_j}{S + n}$$

This can be written as

$$E(p_j|\mathbf{x}) = \left(\frac{S}{S+n}\right) \frac{a_j}{S} + \left(\frac{n}{S+n}\right) \frac{x_j}{n}$$

This is a weighted average of the predictions based on the observed choice probability (x_j/n) and the population mean (a_j/S). The larger the value of S, relative to n, the greater the regression to the mean effect.