Applied Probability Models in Marketing Research: Introduction

Peter S. Fader
University of Pennsylvania
www.petefader.com

Bruce G. S. Hardie
London Business School
www.brucehardie.com

16th Annual Advanced Research Techniques Forum
June 12–15, 2005

Problem 1: Predicting New Product Trial

(Modeling Timing Data)

Background

Ace Snackfoods, Inc. has developed a new snack product called Krunchy Bits. Before deciding whether or not to "go national" with the new product, the marketing manager for Krunchy Bits has decided to commission a year-long test market using IRI's BehaviorScan service, with a view to getting a clearer picture of the product's potential.

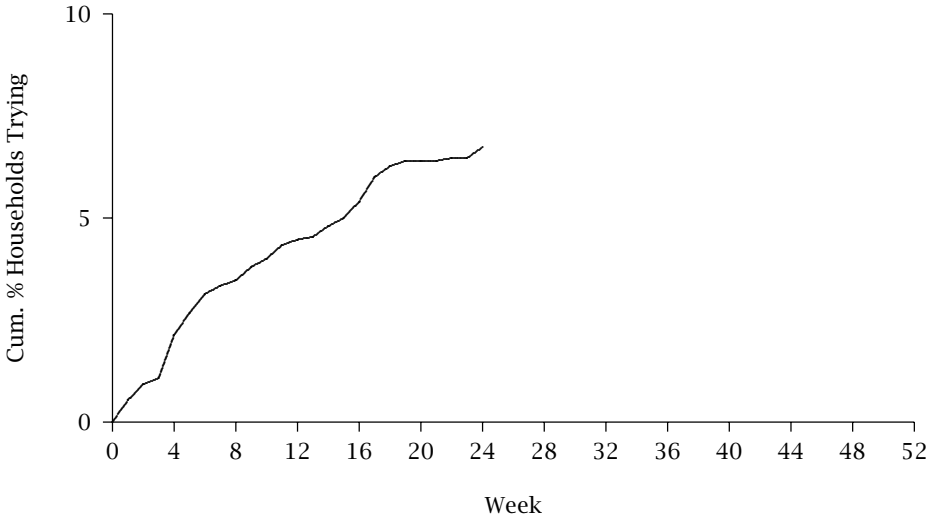
The product has now been under test for 24 weeks. On hand is a dataset documenting the number of households that have made a trial purchase by the end of each week. (The total size of the panel is 1499 households.)

The marketing manager for Krunchy Bits would like a forecast of the product's year-end performance in the test market. First, she wants a forecast of the percentage of households that will have made a trial purchase by week 52.

Krunchy Bits Cumulative Trial

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

Krunchy Bits Cumulative Trial



5

Approaches to Forecasting Trial

- French curve
- "Curve fitting" — specify a flexible functional form, fit it to the data, and project into the future.
- Probability model

Developing a Model of Trial Purchasing

- Start at the individual-level then aggregate.
  - **Q:** What is the individual-level behavior of interest?
  - **A:** Time (since new product launch) of trial purchase.
- We don't know exactly what is driving the behavior ⇒ treat it as a random variable.

The Individual-Level Model

- Let T denote the random variable of interest, and t denote a particular realization.
- Assume time-to-trial is distributed exponentially.
- The probability that an individual has tried by time t is given by:

$$F(t) = P(T \leq t) = 1 - e^{-\lambda t}$$

- λ represents the individual's trial rate.

The Market-Level Model

Assume two segments of consumers:

Segment	Description	Size	λ
1	ever triers	p	θ
2	never triers	$1-p$	0

$$\begin{aligned} P(T \le t) = \; & P(T \le t | \text{ever trier}) \times P(\text{ever trier}) + \\ & P(T \le t | \text{never trier}) \times P(\text{never trier}) \\ = \; & pF(t|\lambda = \theta) + (1-p)F(t|\lambda = 0) \\ = \; & p(1 - e^{-\theta t}) \end{aligned}$$

→ the "exponential w/ never triers" model

Estimating Model Parameters

- Let us assume that the Krunchy Bits data are the outcome of a process characterized by the "exponential w/ never triers" model.
- Which set of model parameters are more likely to have "generated" the data?

p	θ	$P(\text{data})$	$\ln\left(P(\text{data})\right)$
0.5	0.10	1.8×10^{-539}	-1240.5
0.5	0.05	3.9×10^{-443}	-1018.7

Problem 1 -- Model 1

	A	B	C	D	E	F	G	H	I
1	Product:	Krunchy Bits				p	0.5		
2	Panelists:	1499				\theta	0.1		
3						LL =	=SUM(G6:G30)		
4		Cum_Trl							
5	Week	# HHs	Incr_Trl		P(T <= t)	P(try week t)			E[T(t)]
6	1	8	=B6		=G$1*(1-EXP(-G$2*A6))	=E6	=C6*LN(F6)		=B$2*E6
7	2	14	=B7-B6		=G$1*(1-EXP(-G$2*A7))	=E7-E6	=C7*LN(F7)		=B$2*E7
8	3	16	=B8-B7		=G$1*(1-EXP(-G$2*A8))	=E8-E7	=C8*LN(F8)		=B$2*E8
9	4	32	=B9-B8		=G$1*(1-EXP(-G$2*A9))	=E9-E8	=C9*LN(F9)		=B$2*E9
10	5	40	=B10-B9		=G$1*(1-EXP(-G$2*A10))	=E10-E9	=C10*LN(F10)		=B$2*E10
11	6	47	=B11-B10		=G$1*(1-EXP(-G$2*A11))	=E11-E10	=C11*LN(F11)		=B$2*E11
12	7	50	=B12-B11		=G$1*(1-EXP(-G$2*A12))	=E12-E11	=C12*LN(F12)		=B$2*E12
13	8	52	=B13-B12		=G$1*(1-EXP(-G$2*A13))	=E13-E12	=C13*LN(F13)		=B$2*E13
14	9	57	=B14-B13		=G$1*(1-EXP(-G$2*A14))	=E14-E13	=C14*LN(F14)		=B$2*E14
15	10	60	=B15-B14		=G$1*(1-EXP(-G$2*A15))	=E15-E14	=C15*LN(F15)		=B$2*E15
16	11	65	=B16-B15		=G$1*(1-EXP(-G$2*A16))	=E16-E15	=C16*LN(F16)		=B$2*E16
17	12	67	=B17-B16		=G$1*(1-EXP(-G$2*A17))	=E17-E16	=C17*LN(F17)		=B$2*E17
18	13	68	=B18-B17		=G$1*(1-EXP(-G$2*A18))	=E18-E17	=C18*LN(F18)		=B$2*E18
19	14	72	=B19-B18		=G$1*(1-EXP(-G$2*A19))	=E19-E18	=C19*LN(F19)		=B$2*E19
20	15	75	=B20-B19		=G$1*(1-EXP(-G$2*A20))	=E20-E19	=C20*LN(F20)		=B$2*E20
21	16	81	=B21-B20		=G$1*(1-EXP(-G$2*A21))	=E21-E20	=C21*LN(F21)		=B$2*E21
22	17	90	=B22-B21		=G$1*(1-EXP(-G$2*A22))	=E22-E21	=C22*LN(F22)		=B$2*E22
23	18	94	=B23-B22		=G$1*(1-EXP(-G$2*A23))	=E23-E22	=C23*LN(F23)		=B$2*E23
24	19	96	=B24-B23		=G$1*(1-EXP(-G$2*A24))	=E24-E23	=C24*LN(F24)		=B$2*E24
25	20	96	=B25-B24		=G$1*(1-EXP(-G$2*A25))	=E25-E24	=C25*LN(F25)		=B$2*E25
26	21	96	=B26-B25		=G$1*(1-EXP(-G$2*A26))	=E26-E25	=C26*LN(F26)		=B$2*E26
27	22	97	=B27-B26		=G$1*(1-EXP(-G$2*A27))	=E27-E26	=C27*LN(F27)		=B$2*E27
28	23	97	=B28-B27		=G$1*(1-EXP(-G$2*A28))	=E28-E27	=C28*LN(F28)		=B$2*E28
29	24	101	=B29-B28		=G$1*(1-EXP(-G$2*A29))	=E29-E28	=C29*LN(F29)		=B$2*E29
30	25	101			=G$1*(1-EXP(-G$2*A30))	=E30-E29	=(B2-B29)*LN(1-E29)		=B$2*E30
31	26	101			=G$1*(1-EXP(-G$2*A31))	=E31-E30			=B$2*E31
32	27	105			=G$1*(1-EXP(-G$2*A32))	=E32-E31			=B$2*E32
33	28	106			=G$1*(1-EXP(-G$2*A33))	=E33-E32			=B$2*E33
34	29	106			=G$1*(1-EXP(-G$2*A34))	=E34-E33			=B$2*E34
35	30	118			=G$1*(1-EXP(-G$2*A35))	=E35-E34			=B$2*E35
36	31	119			=G$1*(1-EXP(-G$2*A36))	=E36-E35			=B$2*E36
37	32	119			=G$1*(1-EXP(-G$2*A37))	=E37-E36			=B$2*E37
38	33	120			=G$1*(1-EXP(-G$2*A38))	=E38-E37			=B$2*E38
39	34	123			=G$1*(1-EXP(-G$2*A39))	=E39-E38			=B$2*E39
40	35	125			=G$1*(1-EXP(-G$2*A40))	=E40-E39			=B$2*E40
41	36	125			=G$1*(1-EXP(-G$2*A41))	=E41-E40			=B$2*E41
42	37	126			=G$1*(1-EXP(-G$2*A42))	=E42-E41			=B$2*E42
43	38	127			=G$1*(1-EXP(-G$2*A43))	=E43-E42			=B$2*E43
44	39	127			=G$1*(1-EXP(-G$2*A44))	=E44-E43			=B$2*E44
45	40	127			=G$1*(1-EXP(-G$2*A45))	=E45-E44			=B$2*E45
46	41	127			=G$1*(1-EXP(-G$2*A46))	=E46-E45			=B$2*E46
47	42	128			=G$1*(1-EXP(-G$2*A47))	=E47-E46			=B$2*E47
48	43	129			=G$1*(1-EXP(-G$2*A48))	=E48-E47			=B$2*E48
49	44	129			=G$1*(1-EXP(-G$2*A49))	=E49-E48			=B$2*E49
50	45	129			=G$1*(1-EXP(-G$2*A50))	=E50-E49			=B$2*E50
51	46	130			=G$1*(1-EXP(-G$2*A51))	=E51-E50			=B$2*E51
52	47	132			=G$1*(1-EXP(-G$2*A52))	=E52-E51			=B$2*E52
53	48	133			=G$1*(1-EXP(-G$2*A53))	=E53-E52			=B$2*E53
54	49	137			=G$1*(1-EXP(-G$2*A54))	=E54-E53			=B$2*E54
55	50	137			=G$1*(1-EXP(-G$2*A55))	=E55-E54			=B$2*E55
56	51	137			=G$1*(1-EXP(-G$2*A56))	=E56-E55			=B$2*E56
57	52	139			=G$1*(1-EXP(-G$2*A57))	=E57-E56			=B$2*E57

Estimating Model Parameters

We estimate the model parameters using the method of *maximum likelihood*.

- The likelihood function is defined as the probability of observing all of the data points
- This probability is computed using the model and is viewed as a function of the model parameters:

$$L(\text{parameters}) = p(\text{data}|\text{parameters})$$

- For any given set of parameters, $L(\cdot)$ tells us the probability of obtaining the actual data
- For a given dataset, the maximum likelihood estimates of the model parameters are those values that maximize $L(\cdot)$

Estimating Model Parameters

The log-likelihood function is defined as:

$$\begin{aligned} LL(p, \theta|\text{data}) = {} & 8 \times \ln[P(0 < T \leq 1)] + \\ & 6 \times \ln[P(1 < T \leq 2)] + \\ & \ldots + \\ & 4 \times \ln[P(23 < T \leq 24)] + \\ & (1499 - 101) \times \ln[P(T > 24)] \end{aligned}$$

The maximum value of the log-likelihood function is $LL = -680.9$, which occurs at $\hat{p} = 0.085$ and $\hat{\theta} = 0.066$.

Problem 1 -- Model 1

	A	B	C	D	E	F	G	H	I
1	Product:	Krunchy Bits				p	0.085		
2	Panelists:	1499				\theta	0.066		
3						LL =	-680.9		
4		Cum_Trl							
5	Week	# HHs	Incr_Trl		P(T <= t)	P(try week t)			E[T(t)]
6	1	8	8		0.00543	0.00543	-41.723		8.14
7	2	14	6		0.01052	0.00508	-31.691		15.76
8	3	16	2		0.01527	0.00476	-10.696		22.89
9	4	32	16		0.01972	0.00445	-86.633		29.57
10	5	40	8		0.02389	0.00417	-43.848		35.81
11	6	47	7		0.02779	0.00390	-38.832		41.65
12	7	50	3		0.03143	0.00365	-16.841		47.12
13	8	52	2		0.03485	0.00341	-11.360		52.24
14	9	57	5		0.03804	0.00319	-28.733		57.02
15	10	60	3		0.04103	0.00299	-17.439		61.50
16	11	65	5		0.04383	0.00280	-29.397		65.70
17	12	67	2		0.04644	0.00262	-11.892		69.62
18	13	68	1		0.04889	0.00245	-6.012		73.29
19	14	72	4		0.05118	0.00229	-24.314		76.72
20	15	75	3		0.05333	0.00214	-18.435		79.94
21	16	81	6		0.05533	0.00201	-37.268		82.95
22	17	90	9		0.05721	0.00188	-56.500		85.76
23	18	94	4		0.05897	0.00176	-25.377		88.39
24	19	96	2		0.06061	0.00164	-12.821		90.86
25	20	96	0		0.06215	0.00154	0.000		93.16
26	21	96	0		0.06359	0.00144	0.000		95.32
27	22	97	1		0.06494	0.00135	-6.610		97.34
28	23	97	0		0.06620	0.00126	0.000		99.23
29	24	101	4		0.06738	0.00118	-26.970		101.00
30	25	101			0.06848	0.00110	-97.518		102.65
31	26	101			0.06951	0.00103			104.20
32	27	105			0.07048	0.00097			105.65
33	28	106			0.07139	0.00090			107.01
34	29	106			0.07223	0.00085			108.28
35	30	118			0.07302	0.00079			109.46
36	31	119			0.07377	0.00074			110.57
37	32	119			0.07446	0.00069			111.61
38	33	120			0.07511	0.00065			112.59
39	34	123			0.07572	0.00061			113.50
40	35	125			0.07628	0.00057			114.35
41	36	125			0.07682	0.00053			115.15
42	37	126			0.07731	0.00050			115.89
43	38	127			0.07778	0.00047			116.59
44	39	127			0.07821	0.00044			117.24
45	40	127			0.07862	0.00041			117.85
46	41	127			0.07900	0.00038			118.43
47	42	128			0.07936	0.00036			118.96
48	43	129			0.07969	0.00033			119.46
49	44	129			0.08001	0.00031			119.93
50	45	129			0.08030	0.00029			120.37
51	46	130			0.08057	0.00027			120.78
52	47	132			0.08083	0.00026			121.16
53	48	133			0.08107	0.00024			121.52
54	49	137			0.08129	0.00022			121.86
55	50	137			0.08150	0.00021			122.17
56	51	137			0.08170	0.00020			122.47
57	52	139			0.08188	0.00018			122.74

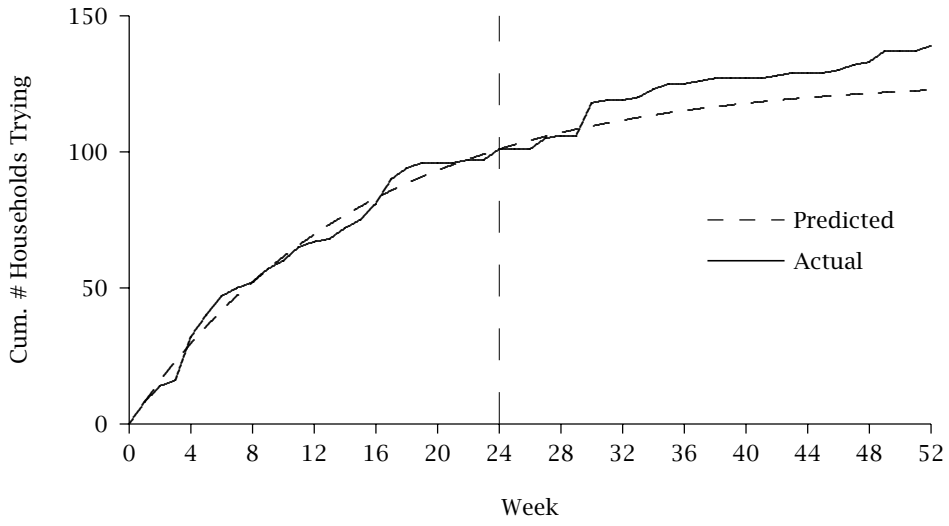
Forecasting Trial

- $F(t)$ represents the probability that a randomly chosen household has made a trial purchase by time t, where $t = 0$ corresponds to the launch of the new product.
- Let $T(t)$ = cumulative # households that have made a trial purchase by time t:

$$\begin{aligned} E[T(t)] &= N \times \hat{F}(t) \\ &= N\hat{p}(1 - e^{-\hat{\theta}t}), \quad t = 1, 2, \ldots \end{aligned}$$

  where N is the panel size.
- Use projection factors for market-level estimates.

Cumulative Trial Forecast



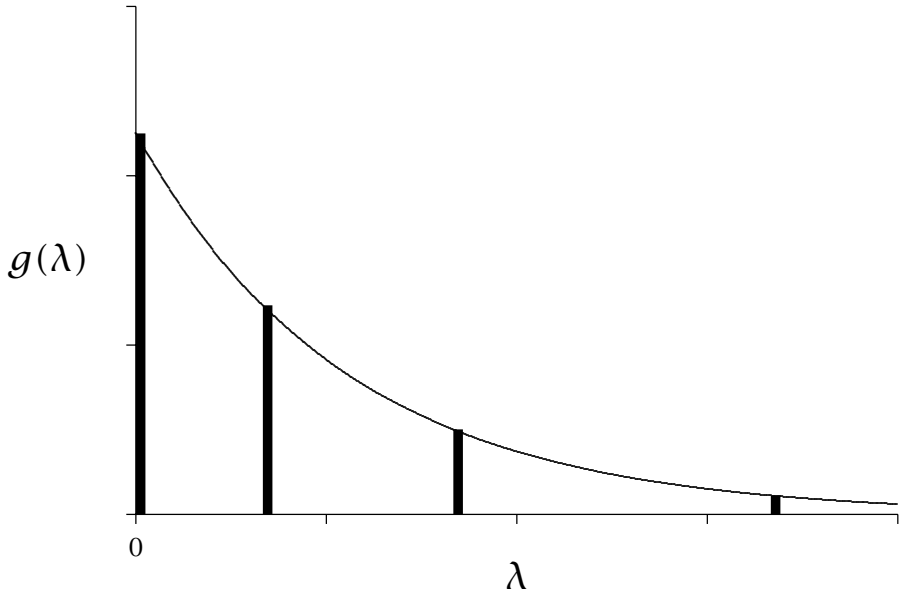
Extending the Basic Model

- The "exponential w/ never triers" model assumes all triers have the same underlying trial rate θ — a bit simplistic.
- Allow for multiple trier "segments" each with a different (latent) trial rate:

$$F(t) = \sum_{s=1}^{S} p_s F(t|\lambda_s), \quad \lambda_1 = 0, \sum_{s=1}^{S} p_s = 1$$

- Replace the discrete distribution with a continuous distribution.

Distribution of Trial Rates



Distribution of Trial Rates

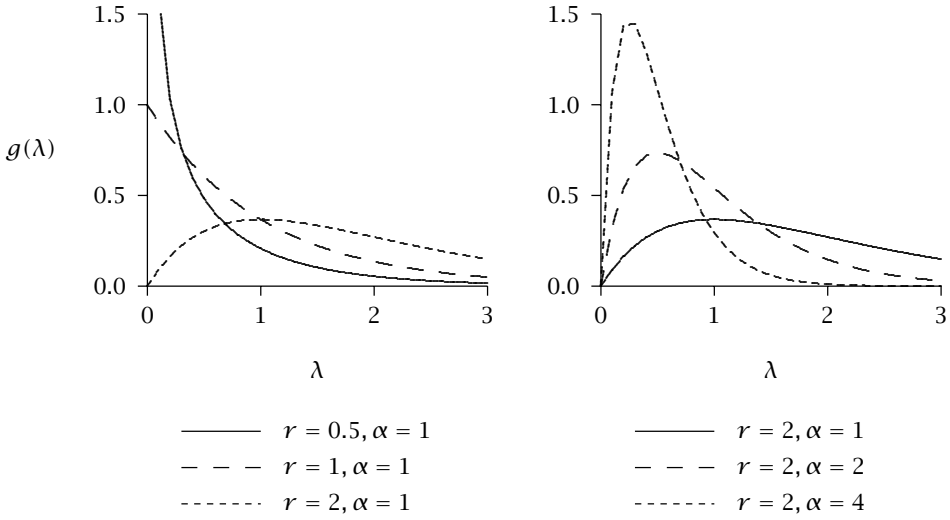
- Assume trial rates are distributed across the population according to a gamma distribution:

$$g(\lambda) = \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)}$$

  where r is the "shape" parameter and α is the "scale" parameter.

- The gamma distribution is a flexible (unimodal) distribution …and is mathematically convenient.

Illustrative Gamma Density Functions



Alternative Market-Level Model

The cumulative distribution of time-to-trial at the market-level is given by:

$$\begin{aligned} P(T \le t) &= \int_0^\infty P(T \le t|\lambda)\, g(\lambda)\, d\lambda \\ &= 1 - \left(\frac{\alpha}{\alpha + t}\right)^r \end{aligned}$$

We call this the "exponential-gamma" model.

Estimating Model Parameters

The log-likelihood function is defined as:

$$\begin{aligned} LL(r, \alpha|\text{data}) = {} & 8 \times \ln[P(0 < T \le 1)] \quad + \\ & 6 \times \ln[P(1 < T \le 2)] \quad + \\ & \ldots \quad + \\ & 4 \times \ln[P(23 < T \le 24)] + \\ & (1499 - 101) \times \ln[P(T > 24)] \end{aligned}$$

The maximum value of the log-likelihood function is $LL = -681.4$, which occurs at $\hat{r} = 0.050$ and $\hat{\alpha} = 7.973$.

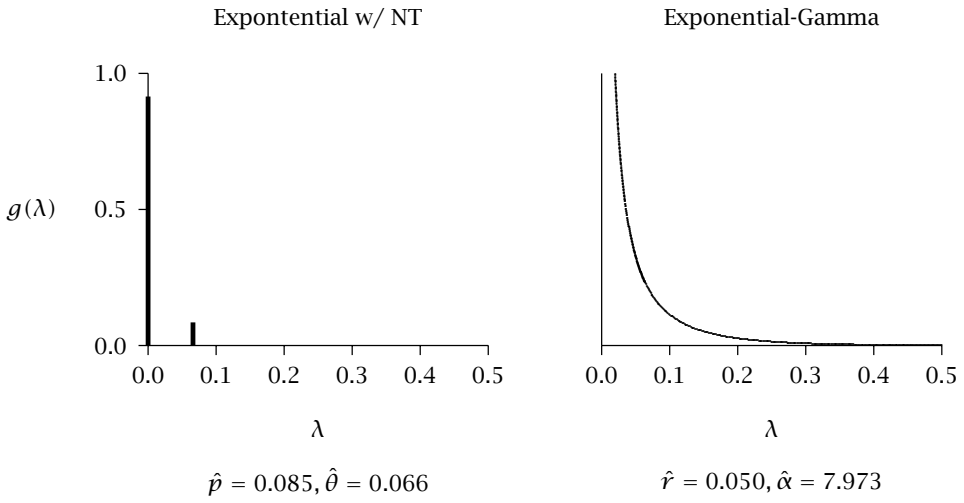
Problem 1 -- Model 2

	A	B	C	D	E	F	G	H	I
1	Product:	Krunchy Bits				r	1		
2	Panelists:	1499				\alpha	1		
3						LL =	=SUM(G6:G30)		
4		Cum_Trl							
5	Week	# HHs	Incr_Trl		P(T <= t)	P(try week t)			E[T(t)]
6	1	8	=B6		=1-(G$2/(G$2+A6))^G$1	=E6	=C6*LN(F6)		=B$2*E6
7	2	14	=B7-B6		=1-(G$2/(G$2+A7))^G$1	=E7-E6	=C7*LN(F7)		=B$2*E7
8	3	16	=B8-B7		=1-(G$2/(G$2+A8))^G$1	=E8-E7	=C8*LN(F8)		=B$2*E8
9	4	32	=B9-B8		=1-(G$2/(G$2+A9))^G$1	=E9-E8	=C9*LN(F9)		=B$2*E9
10	5	40	=B10-B9		=1-(G$2/(G$2+A10))^G$1	=E10-E9	=C10*LN(F10)		=B$2*E10
11	6	47	=B11-B10		=1-(G$2/(G$2+A11))^G$1	=E11-E10	=C11*LN(F11)		=B$2*E11
12	7	50	=B12-B11		=1-(G$2/(G$2+A12))^G$1	=E12-E11	=C12*LN(F12)		=B$2*E12
13	8	52	=B13-B12		=1-(G$2/(G$2+A13))^G$1	=E13-E12	=C13*LN(F13)		=B$2*E13
14	9	57	=B14-B13		=1-(G$2/(G$2+A14))^G$1	=E14-E13	=C14*LN(F14)		=B$2*E14
15	10	60	=B15-B14		=1-(G$2/(G$2+A15))^G$1	=E15-E14	=C15*LN(F15)		=B$2*E15
16	11	65	=B16-B15		=1-(G$2/(G$2+A16))^G$1	=E16-E15	=C16*LN(F16)		=B$2*E16
17	12	67	=B17-B16		=1-(G$2/(G$2+A17))^G$1	=E17-E16	=C17*LN(F17)		=B$2*E17
18	13	68	=B18-B17		=1-(G$2/(G$2+A18))^G$1	=E18-E17	=C18*LN(F18)		=B$2*E18
19	14	72	=B19-B18		=1-(G$2/(G$2+A19))^G$1	=E19-E18	=C19*LN(F19)		=B$2*E19
20	15	75	=B20-B19		=1-(G$2/(G$2+A20))^G$1	=E20-E19	=C20*LN(F20)		=B$2*E20
21	16	81	=B21-B20		=1-(G$2/(G$2+A21))^G$1	=E21-E20	=C21*LN(F21)		=B$2*E21
22	17	90	=B22-B21		=1-(G$2/(G$2+A22))^G$1	=E22-E21	=C22*LN(F22)		=B$2*E22
23	18	94	=B23-B22		=1-(G$2/(G$2+A23))^G$1	=E23-E22	=C23*LN(F23)		=B$2*E23
24	19	96	=B24-B23		=1-(G$2/(G$2+A24))^G$1	=E24-E23	=C24*LN(F24)		=B$2*E24
25	20	96	=B25-B24		=1-(G$2/(G$2+A25))^G$1	=E25-E24	=C25*LN(F25)		=B$2*E25
26	21	96	=B26-B25		=1-(G$2/(G$2+A26))^G$1	=E26-E25	=C26*LN(F26)		=B$2*E26
27	22	97	=B27-B26		=1-(G$2/(G$2+A27))^G$1	=E27-E26	=C27*LN(F27)		=B$2*E27
28	23	97	=B28-B27		=1-(G$2/(G$2+A28))^G$1	=E28-E27	=C28*LN(F28)		=B$2*E28
29	24	101	=B29-B28		=1-(G$2/(G$2+A29))^G$1	=E29-E28	=C29*LN(F29)		=B$2*E29
30	25	101			=1-(G$2/(G$2+A30))^G$1	=E30-E29	=(B2-B29)*LN(1-E29)		=B$2*E30
31	26	101			=1-(G$2/(G$2+A31))^G$1	=E31-E30			=B$2*E31
32	27	105			=1-(G$2/(G$2+A32))^G$1	=E32-E31			=B$2*E32
33	28	106			=1-(G$2/(G$2+A33))^G$1	=E33-E32			=B$2*E33
34	29	106			=1-(G$2/(G$2+A34))^G$1	=E34-E33			=B$2*E34
35	30	118			=1-(G$2/(G$2+A35))^G$1	=E35-E34			=B$2*E35
36	31	119			=1-(G$2/(G$2+A36))^G$1	=E36-E35			=B$2*E36
37	32	119			=1-(G$2/(G$2+A37))^G$1	=E37-E36			=B$2*E37
38	33	120			=1-(G$2/(G$2+A38))^G$1	=E38-E37			=B$2*E38
39	34	123			=1-(G$2/(G$2+A39))^G$1	=E39-E38			=B$2*E39
40	35	125			=1-(G$2/(G$2+A40))^G$1	=E40-E39			=B$2*E40
41	36	125			=1-(G$2/(G$2+A41))^G$1	=E41-E40			=B$2*E41
42	37	126			=1-(G$2/(G$2+A42))^G$1	=E42-E41			=B$2*E42
43	38	127			=1-(G$2/(G$2+A43))^G$1	=E43-E42			=B$2*E43
44	39	127			=1-(G$2/(G$2+A44))^G$1	=E44-E43			=B$2*E44
45	40	127			=1-(G$2/(G$2+A45))^G$1	=E45-E44			=B$2*E45
46	41	127			=1-(G$2/(G$2+A46))^G$1	=E46-E45			=B$2*E46
47	42	128			=1-(G$2/(G$2+A47))^G$1	=E47-E46			=B$2*E47
48	43	129			=1-(G$2/(G$2+A48))^G$1	=E48-E47			=B$2*E48
49	44	129			=1-(G$2/(G$2+A49))^G$1	=E49-E48			=B$2*E49
50	45	129			=1-(G$2/(G$2+A50))^G$1	=E50-E49			=B$2*E50
51	46	130			=1-(G$2/(G$2+A51))^G$1	=E51-E50			=B$2*E51
52	47	132			=1-(G$2/(G$2+A52))^G$1	=E52-E51			=B$2*E52
53	48	133			=1-(G$2/(G$2+A53))^G$1	=E53-E52			=B$2*E53
54	49	137			=1-(G$2/(G$2+A54))^G$1	=E54-E53			=B$2*E54
55	50	137			=1-(G$2/(G$2+A55))^G$1	=E55-E54			=B$2*E55
56	51	137			=1-(G$2/(G$2+A56))^G$1	=E56-E55			=B$2*E56
57	52	139			=1-(G$2/(G$2+A57))^G$1	=E57-E56			=B$2*E57

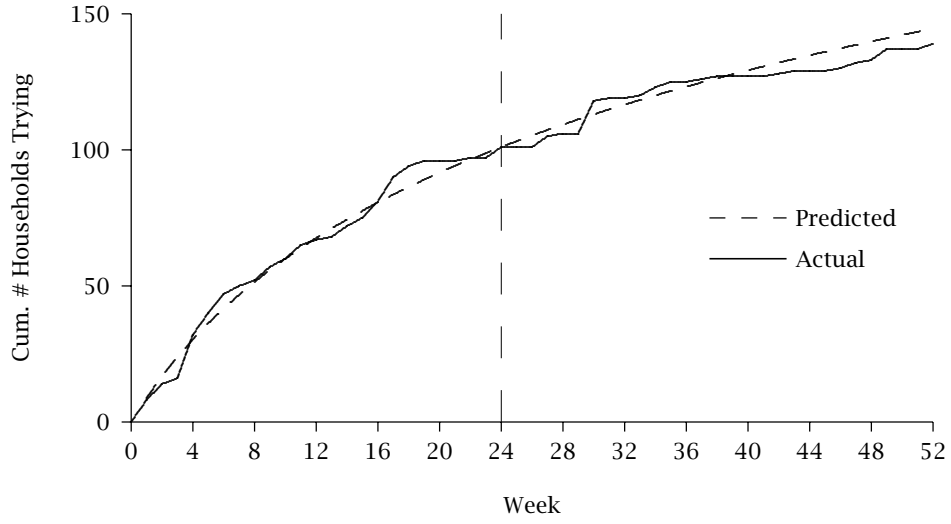
Problem 1 -- Model 2

	A	B	C	D	E	F	G	H	I
1	Product:	Krunchy Bits				r	0.050		
2	Panelists:	1499				\alpha	7.973		
3						LL =	-681.4		
4		Cum_Trl							
5	Week	# HHs	Incr_Trl		P(T <= t)	P(try week t)			E[T(t)]
6	1	8	8		0.00592	0.00592	-41.036		8.87
7	2	14	6		0.01118	0.00526	-31.482		16.76
8	3	16	2		0.01592	0.00474	-10.705		23.86
9	4	32	16		0.02022	0.00430	-87.175		30.31
10	5	40	8		0.02416	0.00394	-44.291		36.22
11	6	47	7		0.02780	0.00363	-39.322		41.67
12	7	50	3		0.03117	0.00337	-17.078		46.72
13	8	52	2		0.03431	0.00314	-11.526		51.43
14	9	57	5		0.03725	0.00294	-29.144		55.84
15	10	60	3		0.04002	0.00277	-17.672		59.98
16	11	65	5		0.04262	0.00261	-29.746		63.89
17	12	67	2		0.04509	0.00247	-12.009		67.59
18	13	68	1		0.04743	0.00234	-6.057		71.10
19	14	72	4		0.04966	0.00223	-24.429		74.44
20	15	75	3		0.05178	0.00212	-18.465		77.62
21	16	81	6		0.05381	0.00203	-37.205		80.66
22	17	90	9		0.05575	0.00194	-56.202		83.57
23	18	94	4		0.05761	0.00186	-25.147		86.36
24	19	96	2		0.05940	0.00179	-12.654		89.04
25	20	96	0		0.06112	0.00172	0.000		91.62
26	21	96	0		0.06277	0.00166	0.000		94.10
27	22	97	1		0.06437	0.00160	-6.440		96.49
28	23	97	0		0.06591	0.00154	0.000		98.80
29	24	101	4		0.06740	0.00149	-26.036		101.04
30	25	101			0.06884	0.00144	-97.554		103.20
31	26	101			0.07024	0.00140			105.29
32	27	105			0.07159	0.00135			107.32
33	28	106			0.07291	0.00131			109.29
34	29	106			0.07419	0.00128			111.20
35	30	118			0.07543	0.00124			113.06
36	31	119			0.07663	0.00121			114.87
37	32	119			0.07781	0.00117			116.63
38	33	120			0.07895	0.00114			118.35
39	34	123			0.08007	0.00112			120.02
40	35	125			0.08115	0.00109			121.65
41	36	125			0.08222	0.00106			123.24
42	37	126			0.08325	0.00104			124.80
43	38	127			0.08426	0.00101			126.31
44	39	127			0.08525	0.00099			127.80
45	40	127			0.08622	0.00097			129.25
46	41	127			0.08717	0.00095			130.67
47	42	128			0.08810	0.00093			132.05
48	43	129			0.08900	0.00091			133.42
49	44	129			0.08989	0.00089			134.75
50	45	129			0.09076	0.00087			136.05
51	46	130			0.09162	0.00085			137.33
52	47	132			0.09245	0.00084			138.59
53	48	133			0.09328	0.00082			139.82
54	49	137			0.09408	0.00081			141.03
55	50	137			0.09487	0.00079			142.22
56	51	137			0.09565	0.00078			143.38
57	52	139			0.09641	0.00076			144.53

Estimated Distribution of λ



Cumulative Trial Forecast



Further Model Extensions

- Combine a “never triers” term with the “exponential-gamma” model.
- Incorporate the effects of marketing covariates.
- Model repeat sales using a “depth of repeat” formulation, where transitions from one repeat class to the next are modeled using an “exponential-gamma”-type model.

Concepts and Tools Introduced

- Probability models
- (Single-event) timing processes
- Models of new product trial/adoption

Further Reading

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

Fader, Peter S., Bruce G. S. Hardie, and Robert Zeithammer (2003), "Forecasting New Product Trial in a Controlled Test Market Environment," *Journal of Forecasting*, **22** (August), 391–410.

Kalbfleisch, John D. and Ross L. Prentice (2002), *The Statistical Analysis of Failure Time Data*, 2nd edn., New York: Wiley.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, New York: Wiley.

Introduction to Probability Models

The Logic of Probability Models

- Many researchers attempt to describe/predict behavior using observed variables.
- However, they still use random components in recognition that not all factors are included in the model.
- We treat behavior as if it were “random” (probabilistic, stochastic).
- We propose a model of individual-level behavior which is “summed” across individuals (taking individual differences into account) to obtain a model of aggregate behavior.

Uses of Probability Models

- Understanding market-level behavior patterns
- Prediction
  - To settings (e.g., time periods) beyond the observation period
  - Conditional on past behavior
- Profiling behavioral propensities of individuals
- Benchmarks/norms

Building a Probability Model

(i) Determine the marketing decision problem/ information needed.

(ii) Identify the *observable* individual-level behavior of interest.

- We denote this by x.

(iii) Select a probability distribution that characterizes this individual-level behavior.

- This is denoted by $f(x|\theta)$.
- We view the parameters of this distribution as individual-level *latent traits*.

Building a Probability Model

(iv) Specify a distribution to characterize the distribution of the latent trait variable(s) across the population.

- We denote this by $g(\theta)$.
- This is often called the *mixing distribution*.

(v) Derive the corresponding *aggregate* or *observed* distribution for the behavior of interest:

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

Building a Probability Model

(vi) Estimate the parameters (of the mixing distribution) by fitting the aggregate distribution to the observed data.

(vii) Use the model to solve the marketing decision problem/provide the required information.

Outline

- Problem 1: Predicting New Product Trial
  (Modeling Timing Data)
- Problem 2: Estimating Billboard Exposures
  (Modeling Count Data)
- Problem 3: Test/Roll Decisions in Segmentation-based Direct Marketing
  (Modeling "Choice" Data)
- Further applications and tools/modeling issues

Problem 2:
Estimating Billboard Exposures

(Modeling Count Data)

Background

One advertising medium at the marketer's disposal is the outdoor billboard. The unit of purchase for this medium is usually a "monthly showing," which comprises a specific set of billboards carrying the advertiser's message in a given market.

The effectiveness of a monthly showing is evaluated in terms of three measures: reach, (average) frequency, and gross rating points (GRPs). These measures are determined using data collected from a sample of people in the market.

Respondents record their daily travel on maps. From each respondent's travel map, the total frequency of exposure to the showing over the survey period is counted. An "exposure" is deemed to occur each time the respondent travels by a billboard in the showing, on the street or road closest to that billboard, going towards the billboard's face.

Background

The standard approach to data collection requires each respondent to fill out daily travel maps for *an entire month*. The problem with this is that it is difficult and expensive to get a high proportion of respondents to do this accurately.

B&P Research is interested in developing a means by which it can generate effectiveness measures for a monthly showing from a survey in which respondents fill out travel maps for *only one week*.

Data have been collected from a sample of 250 residents who completed daily travel maps for one week. The sampling process is such that approximately one quarter of the respondents fill out travel maps during each of the four weeks in the target month.

Effectiveness Measures

The effectiveness of a monthly showing is evaluated in terms of three measures:

- Reach: the proportion of the population exposed to the billboard message at least once in the month.
- Average Frequency: the average number of exposures (per month) among those people reached.
- Gross Rating Points (GRPs): the mean number of exposures per 100 people.

Distribution of Billboard Exposures (1 week)

# Exposures	# People	# Exposures	# People
0	48	12	5
1	37	13	3
2	30	14	3
3	24	15	2
4	20	16	2
5	16	17	2
6	13	18	1
7	11	19	1
8	9	20	2
9	7	21	1
10	6	22	1
11	5	23	1

Average # Exposures = 4.456

Modeling Objective

Develop a model that enables us to estimate a billboard showing's reach, average frequency, and GRPs for the month using the one-week data.

Modeling Issues

- Modeling the exposures to showing in a week.
- Estimating summary statistics of the exposure distribution for a longer period of time (i.e., one month).

Modeling One Week Exposures

- Let the random variable X denote the number of exposures to the showing in a week.
- At the individual-level, X is assumed to be Poisson distributed with (exposure) rate parameter λ:

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

- Exposure rates (λ) are distributed across the population according to a gamma distribution:

$$g(\lambda) = \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)}$$

Modeling One Week Exposures

- The distribution of exposures at the population-level is given by:

$$\begin{aligned} P(X = x) &= \int_0^\infty P(X = x|\lambda)\, g(\lambda)\, d\lambda \\ &= \frac{\Gamma(r + x)}{\Gamma(r)x!} \left(\frac{\alpha}{\alpha + 1}\right)^r \left(\frac{1}{\alpha + 1}\right)^x \end{aligned}$$

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

- The mean of the NBD is given by $E(X) = r/\alpha$.

Computing NBD Probabilities

- Note that

$$\frac{P(X = x)}{P(X = x - 1)} = \frac{r + x - 1}{x(\alpha + 1)}$$

- We can therefore compute NBD probabilities using the following *forward recursion* formula:

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

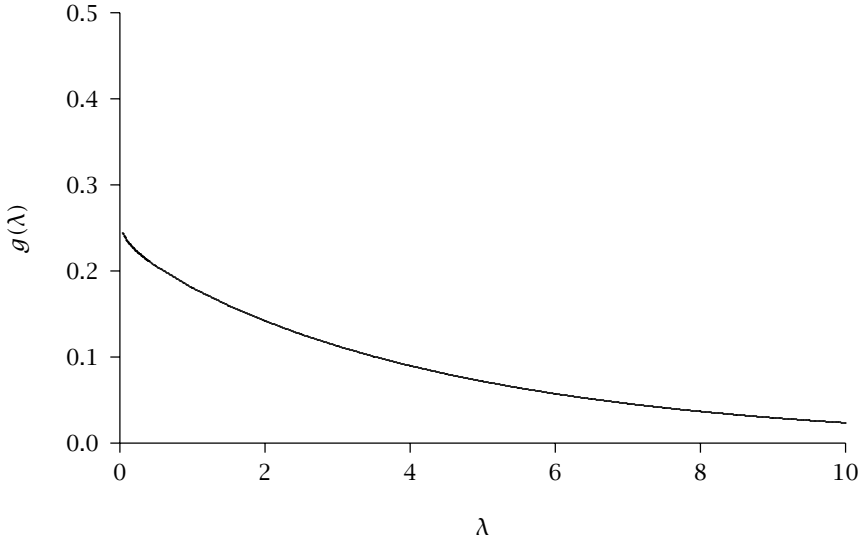
Estimating Model Parameters

The log-likelihood function is defined as:

$$\begin{aligned} LL(r, \alpha|\text{data}) = 48 \times \ln[P(X=0)] &+ \\ 37 \times \ln[P(X=1)] &+ \\ 30 \times \ln[P(X=2)] &+ \\ \ldots &+ \\ 1 \times \ln[P(X=23)] & \end{aligned}$$

The maximum value of the log-likelihood function is $LL = -649.7$, which occurs at $\hat{r} = 0.969$ and $\hat{\alpha} = 0.218$.

Estimated Distribution of λ



$\hat{r} = 0.969, \hat{\alpha} = 0.218$

Problem 2 -- Parameter Estimation

	A	B	C	D
1	r	1		
2	\alpha	1	LL=	=SUM(D5:D28)
3				
4	x	f_x	P(X=x)	
5	0	48	=(B2/(B2+1))^B1	=B5*LN(C5)
6	1	37	=C5*(B$1+A6-1)/(A6*(B$2+1))	=B6*LN(C6)
7	2	30	=C6*(B$1+A7-1)/(A7*(B$2+1))	=B7*LN(C7)
8	3	24	=C7*(B$1+A8-1)/(A8*(B$2+1))	=B8*LN(C8)
9	4	20	=C8*(B$1+A9-1)/(A9*(B$2+1))	=B9*LN(C9)
10	5	16	=C9*(B$1+A10-1)/(A10*(B$2+1))	=B10*LN(C10)
11	6	13	=C10*(B$1+A11-1)/(A11*(B$2+1))	=B11*LN(C11)
12	7	11	=C11*(B$1+A12-1)/(A12*(B$2+1))	=B12*LN(C12)
13	8	9	=C12*(B$1+A13-1)/(A13*(B$2+1))	=B13*LN(C13)
14	9	7	=C13*(B$1+A14-1)/(A14*(B$2+1))	=B14*LN(C14)
15	10	6	=C14*(B$1+A15-1)/(A15*(B$2+1))	=B15*LN(C15)
16	11	5	=C15*(B$1+A16-1)/(A16*(B$2+1))	=B16*LN(C16)
17	12	5	=C16*(B$1+A17-1)/(A17*(B$2+1))	=B17*LN(C17)
18	13	3	=C17*(B$1+A18-1)/(A18*(B$2+1))	=B18*LN(C18)
19	14	3	=C18*(B$1+A19-1)/(A19*(B$2+1))	=B19*LN(C19)
20	15	2	=C19*(B$1+A20-1)/(A20*(B$2+1))	=B20*LN(C20)
21	16	2	=C20*(B$1+A21-1)/(A21*(B$2+1))	=B21*LN(C21)
22	17	2	=C21*(B$1+A22-1)/(A22*(B$2+1))	=B22*LN(C22)
23	18	1	=C22*(B$1+A23-1)/(A23*(B$2+1))	=B23*LN(C23)
24	19	1	=C23*(B$1+A24-1)/(A24*(B$2+1))	=B24*LN(C24)
25	20	2	=C24*(B$1+A25-1)/(A25*(B$2+1))	=B25*LN(C25)
26	21	1	=C25*(B$1+A26-1)/(A26*(B$2+1))	=B26*LN(C26)
27	22	1	=C26*(B$1+A27-1)/(A27*(B$2+1))	=B27*LN(C27)
28	23	1	=C27*(B$1+A28-1)/(A28*(B$2+1))	=B28*LN(C28)

Problem 2 -- Parameter Estimation

	A	B	C	D
1	r	0.96926		
2	\alpha	0.21752	LL=	-649.6888
3				
4	x	f_x	P(X=x)	
5	0	48	0.18837	-80.128
6	1	37	0.14996	-70.203
7	2	30	0.12128	-63.291
8	3	24	0.09859	-55.603
9	4	20	0.08035	-50.427
10	5	16	0.06559	-43.589
11	6	13	0.05360	-38.041
12	7	11	0.04383	-34.402
13	8	9	0.03586	-29.953
14	9	7	0.02935	-24.699
15	10	6	0.02403	-22.370
16	11	5	0.01969	-19.639
17	12	5	0.01613	-20.636
18	13	3	0.01321	-12.979
19	14	3	0.01083	-13.576
20	15	2	0.00888	-9.449
21	16	2	0.00728	-9.846
22	17	2	0.00597	-10.243
23	18	1	0.00489	-5.320
24	19	1	0.00401	-5.519
25	20	2	0.00329	-11.434
26	21	1	0.00270	-5.915
27	22	1	0.00221	-6.113
28	23	1	0.00182	-6.312

NBD for a Non-Unit Time Period

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

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

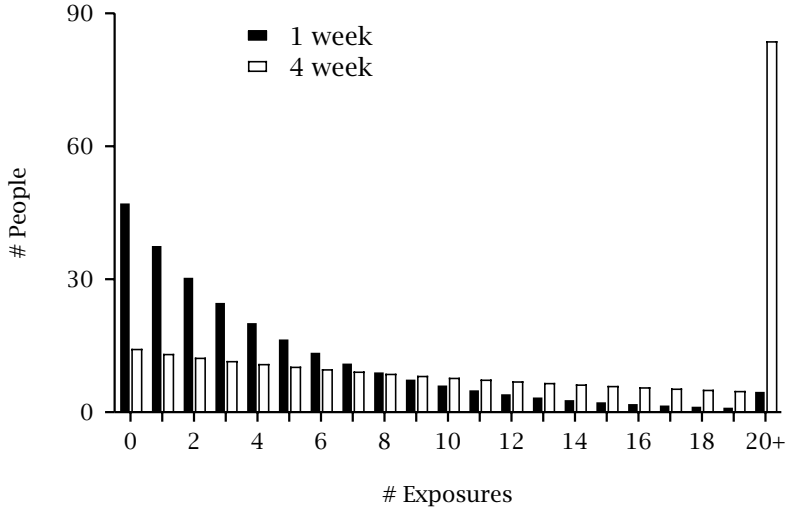
NBD for a Non-Unit Time Period

- The distribution of exposures at the population-level is given by:

$$\begin{aligned} P(X(t) = x) &= \int_0^\infty P(X(t) = x|\lambda)\, g(\lambda)\, d\lambda \\ &= \frac{\Gamma(r+x)}{\Gamma(r)x!} \left(\frac{\alpha}{\alpha+t}\right)^r \left(\frac{t}{\alpha+t}\right)^x \end{aligned}$$

- The mean of this distribution is given by $E[X(t)] = rt/\alpha$.

Exposure Distributions: 1 week vs. 4 week



Effectiveness of Monthly Showing

- For $t = 4$, we have:
  - $P(X(t) = 0) = 0.056$, and
  - $E[X(t)] = 17.82$
- It follows that:
  - Reach $= 1 - P(X(t) = 0)$
    $= 94.4\%$
  - Frequency $= E[X(t)]/(1 - P(X(t) = 0))$
    $= 18.9$
  - GRPs $= 100 \times E[X(t)]$
    $= 1782$

Problem 2 -- Solution

	A	B
1	r	='Parameter Estimation'!B1
2	\alpha	='Parameter Estimation'!B2
3	t	4
4		
5	P(X(t)=0)	=(B2/(B2+B3))^B1
6	E[X(t)]	=B1*B3/B2
7		
8	Reach	=1-B5
9	Frequency	=B6/B8
10	GRPs	=100*B6

Problem 2 -- Solution

	A	B
1	r	0.96926
2	\alpha	0.21752
3	t	4
4		
5	P(X(t)=0)	0.056
6	E[X(t)]	17.82
7		
8	Reach	94.4%
9	Frequency	18.9
10	GRPs	1782

Concepts and Tools Introduced

- Counting processes
- The NBD model
- Extrapolating an observed histogram over time
- Using models to estimate "exposure distributions" for media vehicles

Further Reading

Greene, Jerome D. (1982), *Consumer Behavior Models for Non-Statisticians*, New York: Praeger.

Morrison, Donald G. and David C. Schmittlein (1988), "Generalizing the NBD Model for Customer Purchases: What Are the Implications and Is It Worth the Effort?" *Journal of Business and Economic Statistics*, **6** (April), 145–159.

Ehrenberg, A. S. C. (1988), *Repeat-Buying*, 2nd edn., London: Charles Griffin & Company, Ltd. (Available online at <http://www.empgens.com/repeat-buying.htm>.)

Problem 3:
Test/Roll Decisions in Segmentation-based Direct Marketing

(Modeling "Choice" Data)

The "Segmentation" Approach

1. Divide the customer list into a set of (homogeneous) segments.
2. Test customer response by mailing to a random sample of each segment.
3. Rollout to segments with a response rate (RR) above some cut-off point,

$$\text{e.g., RR} > \frac{\text{cost of each mailing}}{\text{unit margin}}$$

Ben's Knick Knacks, Inc.

- A consumer durable product (unit margin = $161.50, mailing cost per 10,000 = $3343)
- 126 segments formed from customer database on the basis of past purchase history information
- Test mailing to 3.24% of database

Ben's Knick Knacks, Inc.

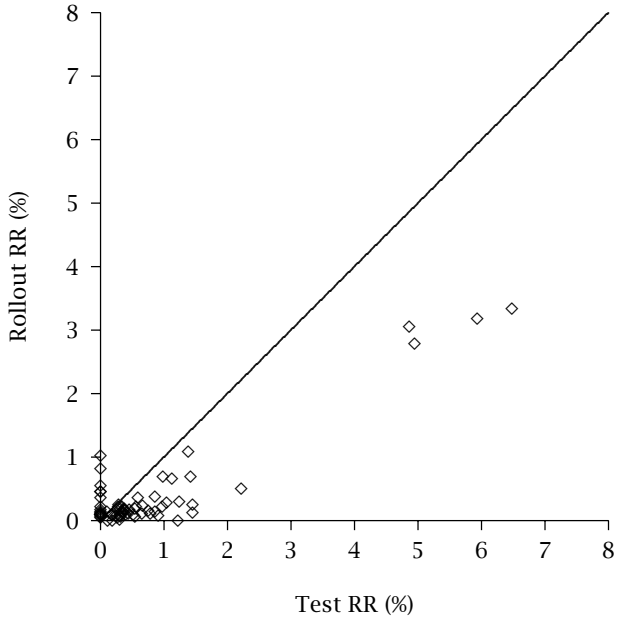
Standard approach:

- Rollout to all segments with

$$\text{Test RR} > \frac{3343/10,000}{161.50} = 0.00207$$

- 51 segments pass this hurdle

Test vs. Actual Response Rate



55

Modeling Objective

Develop a model that leverages the whole data set to make better informed decisions.

Model Development

Notation:

$$N_s = \text{size of segment } s \ (s = 1, \ldots, S)$$
$$m_s = \#\text{ members of segment } s \text{ tested}$$
$$X_s = \#\text{ responses to test in segment } s$$

Assume: All members of segment s have the same (unknown) response probability $p_s \Rightarrow X_s$ is a binomial random variable

$$P(X_s = x_s | m_s, p_s) = \binom{m_s}{x_s} p_s^{x_s} (1 - p_s)^{m_s - x_s}$$

Distribution of Response Probabilities

- Heterogeneity in p_s is captured using a beta distribution:

$$g(p_s) = \frac{1}{B(\alpha, \beta)} p_s^{\alpha - 1} (1 - p_s)^{\beta - 1}$$

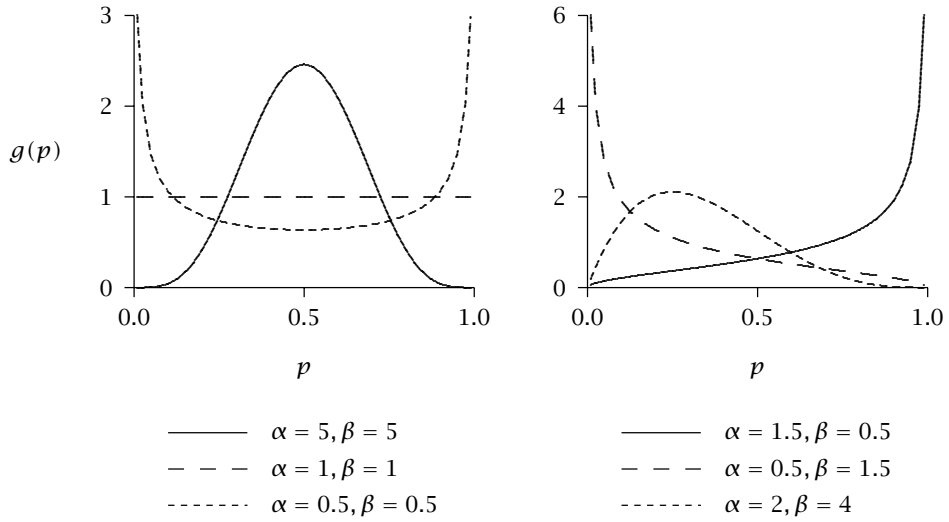
- The beta function, $B(\alpha, \beta)$, can be expressed as

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

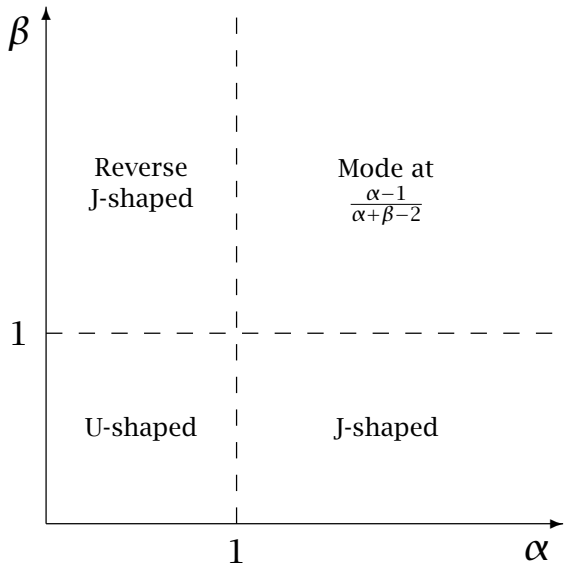
- The mean of the beta distribution is given by

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

Illustrative Beta Density Functions



Shape of the Beta Density



The Beta Binomial Model

The aggregate distribution of responses to a mailing of size m_s is given by

$$\begin{aligned} P(X_s = x_s | m_s) &= \int_0^1 P(X_s = x_s | m_s, p_s)\, g(p_s)\, dp_s \\ &= \binom{m_s}{x_s} \frac{B(\alpha + x_s, \beta + m_s - x_s)}{B(\alpha, \beta)} \end{aligned}$$

Estimating Model Parameters

The log-likelihood function is defined as:

$$\begin{aligned} LL(\alpha, \beta | \text{data}) &= \sum_{s=1}^{126} \ln[P(X_s = x_s | m_s)] \\ &= \sum_{s=1}^{126} \ln\Bigg[\frac{m_s!}{(m_s - x_s)!\, x_s!} \underbrace{\frac{\Gamma(\alpha + x_s)\Gamma(\beta + m_s - x_s)}{\Gamma(\alpha + \beta + m_s)}}_{B(\alpha + x_s, \beta + m_s - x_s)} \underbrace{\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{1/B(\alpha, \beta)} \Bigg] \end{aligned}$$

The maximum value of the log-likelihood function is $LL = -200.5$, which occurs at $\hat{\alpha} = 0.439$ and $\hat{\beta} = 95.411$.

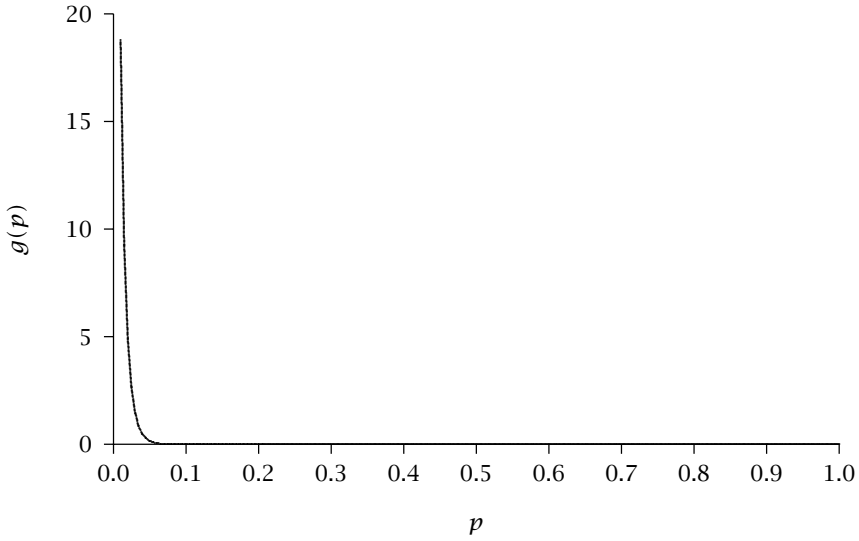
Problem 3 -- Model (a)

	A	B	C	D	E	F
1	\alpha	1		B(\alpha,\beta)	=EXP(GAMMALN(B1)+GAMMALN(B2)-GAMMALN(B1+B2))	
2	\beta	1				
3					LL =	=SUM(F6:F131)
4						
5	Segment	m_s	x_s		P(X=x|m)	
6	1	34	0		=COMBIN(B6,C6)*EXP(GAMMALN(B$1+C6)+GAMMALN(B$2+B6-C6)-GAMMALN(B$1+B$2+B6))/E$1	=LN(E6)
7	2	102	1		=COMBIN(B7,C7)*EXP(GAMMALN(B$1+C7)+GAMMALN(B$2+B7-C7)-GAMMALN(B$1+B$2+B7))/E$1	=LN(E7)
8	3	53	0		=COMBIN(B8,C8)*EXP(GAMMALN(B$1+C8)+GAMMALN(B$2+B8-C8)-GAMMALN(B$1+B$2+B8))/E$1	=LN(E8)
9	4	145	2		=COMBIN(B9,C9)*EXP(GAMMALN(B$1+C9)+GAMMALN(B$2+B9-C9)-GAMMALN(B$1+B$2+B9))/E$1	=LN(E9)
10	5	1254	62		=COMBIN(B10,C10)*EXP(GAMMALN(B$1+C10)+GAMMALN(B$2+B10-C10)-GAMMALN(B$1+B$2+B10))/E$1	=LN(E10)
11	6	144	7		=COMBIN(B11,C11)*EXP(GAMMALN(B$1+C11)+GAMMALN(B$2+B11-C11)-GAMMALN(B$1+B$2+B11))/E$1	=LN(E11)
12	7	1235	80		=COMBIN(B12,C12)*EXP(GAMMALN(B$1+C12)+GAMMALN(B$2+B12-C12)-GAMMALN(B$1+B$2+B12))/E$1	=LN(E12)
13	8	573	34		=COMBIN(B13,C13)*EXP(GAMMALN(B$1+C13)+GAMMALN(B$2+B13-C13)-GAMMALN(B$1+B$2+B13))/E$1	=LN(E13)
14	9	1083	24		=COMBIN(B14,C14)*EXP(GAMMALN(B$1+C14)+GAMMALN(B$2+B14-C14)-GAMMALN(B$1+B$2+B14))/E$1	=LN(E14)
15	10	352	5		=COMBIN(B15,C15)*EXP(GAMMALN(B$1+C15)+GAMMALN(B$2+B15-C15)-GAMMALN(B$1+B$2+B15))/E$1	=LN(E15)
16	11	817	7		=COMBIN(B16,C16)*EXP(GAMMALN(B$1+C16)+GAMMALN(B$2+B16-C16)-GAMMALN(B$1+B$2+B16))/E$1	=LN(E16)
17	12	118	0		=COMBIN(B17,C17)*EXP(GAMMALN(B$1+C17)+GAMMALN(B$2+B17-C17)-GAMMALN(B$1+B$2+B17))/E$1	=LN(E17)
18	13	1049	3		=COMBIN(B18,C18)*EXP(GAMMALN(B$1+C18)+GAMMALN(B$2+B18-C18)-GAMMALN(B$1+B$2+B18))/E$1	=LN(E18)
19	14	452	3		=COMBIN(B19,C19)*EXP(GAMMALN(B$1+C19)+GAMMALN(B$2+B19-C19)-GAMMALN(B$1+B$2+B19))/E$1	=LN(E19)
20	15	338	2		=COMBIN(B20,C20)*EXP(GAMMALN(B$1+C20)+GAMMALN(B$2+B20-C20)-GAMMALN(B$1+B$2+B20))/E$1	=LN(E20)
21	16	168	0		=COMBIN(B21,C21)*EXP(GAMMALN(B$1+C21)+GAMMALN(B$2+B21-C21)-GAMMALN(B$1+B$2+B21))/E$1	=LN(E21)
22	17	242	3		=COMBIN(B22,C22)*EXP(GAMMALN(B$1+C22)+GAMMALN(B$2+B22-C22)-GAMMALN(B$1+B$2+B22))/E$1	=LN(E22)
23	18	185	1		=COMBIN(B23,C23)*EXP(GAMMALN(B$1+C23)+GAMMALN(B$2+B23-C23)-GAMMALN(B$1+B$2+B23))/E$1	=LN(E23)
24	19	116	0		=COMBIN(B24,C24)*EXP(GAMMALN(B$1+C24)+GAMMALN(B$2+B24-C24)-GAMMALN(B$1+B$2+B24))/E$1	=LN(E24)
25	20	69	1		=COMBIN(B25,C25)*EXP(GAMMALN(B$1+C25)+GAMMALN(B$2+B25-C25)-GAMMALN(B$1+B$2+B25))/E$1	=LN(E25)
26	21	193	1		=COMBIN(B26,C26)*EXP(GAMMALN(B$1+C26)+GAMMALN(B$2+B26-C26)-GAMMALN(B$1+B$2+B26))/E$1	=LN(E26)
27	22	82	1		=COMBIN(B27,C27)*EXP(GAMMALN(B$1+C27)+GAMMALN(B$2+B27-C27)-GAMMALN(B$1+B$2+B27))/E$1	=LN(E27)
28	23	265	1		=COMBIN(B28,C28)*EXP(GAMMALN(B$1+C28)+GAMMALN(B$2+B28-C28)-GAMMALN(B$1+B$2+B28))/E$1	=LN(E28)
29	24	171	0		=COMBIN(B29,C29)*EXP(GAMMALN(B$1+C29)+GAMMALN(B$2+B29-C29)-GAMMALN(B$1+B$2+B29))/E$1	=LN(E29)
30	25	1554	7		=COMBIN(B30,C30)*EXP(GAMMALN(B$1+C30)+GAMMALN(B$2+B30-C30)-GAMMALN(B$1+B$2+B30))/E$1	=LN(E30)
31	26	1339	4		=COMBIN(B31,C31)*EXP(GAMMALN(B$1+C31)+GAMMALN(B$2+B31-C31)-GAMMALN(B$1+B$2+B31))/E$1	=LN(E31)
32	27	1167	4		=COMBIN(B32,C32)*EXP(GAMMALN(B$1+C32)+GAMMALN(B$2+B32-C32)-GAMMALN(B$1+B$2+B32))/E$1	=LN(E32)
33	28	621	2		=COMBIN(B33,C33)*EXP(GAMMALN(B$1+C33)+GAMMALN(B$2+B33-C33)-GAMMALN(B$1+B$2+B33))/E$1	=LN(E33)
34	29	1013	1		=COMBIN(B34,C34)*EXP(GAMMALN(B$1+C34)+GAMMALN(B$2+B34-C34)-GAMMALN(B$1+B$2+B34))/E$1	=LN(E34)
35	30	544	1		=COMBIN(B35,C35)*EXP(GAMMALN(B$1+C35)+GAMMALN(B$2+B35-C35)-GAMMALN(B$1+B$2+B35))/E$1	=LN(E35)
36	31	731	1		=COMBIN(B36,C36)*EXP(GAMMALN(B$1+C36)+GAMMALN(B$2+B36-C36)-GAMMALN(B$1+B$2+B36))/E$1	=LN(E36)
37	32	326	0		=COMBIN(B37,C37)*EXP(GAMMALN(B$1+C37)+GAMMALN(B$2+B37-C37)-GAMMALN(B$1+B$2+B37))/E$1	=LN(E37)
38	33	772	1		=COMBIN(B38,C38)*EXP(GAMMALN(B$1+C38)+GAMMALN(B$2+B38-C38)-GAMMALN(B$1+B$2+B38))/E$1	=LN(E38)
39	34	335	1		=COMBIN(B39,C39)*EXP(GAMMALN(B$1+C39)+GAMMALN(B$2+B39-C39)-GAMMALN(B$1+B$2+B39))/E$1	=LN(E39)
40	35	235	0		=COMBIN(B40,C40)*EXP(GAMMALN(B$1+C40)+GAMMALN(B$2+B40-C40)-GAMMALN(B$1+B$2+B40))/E$1	=LN(E40)
41	36	218	0		=COMBIN(B41,C41)*EXP(GAMMALN(B$1+C41)+GAMMALN(B$2+B41-C41)-GAMMALN(B$1+B$2+B41))/E$1	=LN(E41)
42	37	221	0		=COMBIN(B42,C42)*EXP(GAMMALN(B$1+C42)+GAMMALN(B$2+B42-C42)-GAMMALN(B$1+B$2+B42))/E$1	=LN(E42)
43	38	103	1		=COMBIN(B43,C43)*EXP(GAMMALN(B$1+C43)+GAMMALN(B$2+B43-C43)-GAMMALN(B$1+B$2+B43))/E$1	=LN(E43)
44	39	170	0		=COMBIN(B44,C44)*EXP(GAMMALN(B$1+C44)+GAMMALN(B$2+B44-C44)-GAMMALN(B$1+B$2+B44))/E$1	=LN(E44)
45	40	45	0		=COMBIN(B45,C45)*EXP(GAMMALN(B$1+C45)+GAMMALN(B$2+B45-C45)-GAMMALN(B$1+B$2+B45))/E$1	=LN(E45)

Problem 3 -- Model

	A	B	C	D	E	F	G	H	I
1	\alpha	0.439	B(\alpha,\beta)		0.273				
2	\beta	95.411							
3					LL =	-200.548		cutoff	0.00207
4									
5	Segment	m_s	x_s		P(X=x|m)			E[p_s|x_s]	Roll?
6	1	34	0		0.87448	-0.134		0.00338	Y
7	2	102	1		0.16556	-1.798		0.00727	Y
8	3	53	0		0.82334	-0.194		0.00295	Y
9	4	145	2		0.07694	-2.565		0.01013	Y
10	5	1254	62		0.00015	-8.793		0.04626	Y
11	6	144	7		0.00301	-5.805		0.03101	Y
12	7	1235	80		0.00003	-10.403		0.06044	Y
13	8	573	34		0.00014	-8.869		0.05149	Y
14	9	1083	24		0.00362	-5.622		0.02073	Y
15	10	352	5		0.03010	-3.503		0.01214	Y
16	11	817	7		0.02810	-3.572		0.00815	Y
17	12	118	0		0.70182	-0.354		0.00205	N
18	13	1049	3		0.06653	-2.710		0.00300	Y
19	14	452	3		0.06735	-2.698		0.00628	Y
20	15	338	2		0.09913	-2.311		0.00562	Y
21	16	168	0		0.63981	-0.447		0.00166	N
22	17	242	3		0.05465	-2.907		0.01018	Y
23	18	185	1		0.18091	-1.710		0.00512	Y
24	19	116	0		0.70473	-0.350		0.00207	Y
25	20	69	1		0.14588	-1.925		0.00873	Y
26	21	193	1		0.18122	-1.708		0.00498	Y
27	22	82	1		0.15531	-1.862		0.00809	Y
28	23	265	1		0.18042	-1.712		0.00399	Y
29	24	171	0		0.63664	-0.452		0.00164	N
30	25	1554	7		0.03089	-3.477		0.00451	Y
31	26	1339	4		0.05107	-2.975		0.00309	Y
32	27	1167	4		0.05197	-2.957		0.00352	Y
33	28	621	2		0.09808	-2.322		0.00340	Y
34	29	1013	1		0.13667	-1.990		0.00130	N
35	30	544	1		0.16210	-1.820		0.00225	Y
36	31	731	1		0.15052	-1.894		0.00174	N
37	32	326	0		0.52048	-0.653		0.00104	N
38	33	772	1		0.14826	-1.909		0.00166	N
39	34	335	1		0.17658	-1.734		0.00334	Y
40	35	235	0		0.57918	-0.546		0.00133	N
41	36	218	0		0.59277	-0.523		0.00140	N
42	37	221	0		0.59030	-0.527		0.00139	N
43	38	103	1		0.16596	-1.796		0.00724	Y
44	39	170	0		0.63769	-0.450		0.00165	N
45	40	45	0		0.84365	-0.170		0.00312	Y
46	41	237	0		0.57764	-0.549		0.00132	N
47	42	86	0		0.75377	-0.283		0.00241	Y
48	43	297	1		0.17887	-1.721		0.00366	Y
49	44	415	0		0.47847	-0.737		0.00086	N
50	45	187	0		0.62053	-0.477		0.00155	N
51	46	248	0		0.56944	-0.563		0.00128	N

Estimated Distribution of p



$$\hat{\alpha} = 0.439, \hat{\beta} = 95.411, \bar{p} = 0.0046$$

Applying the Model

What is our best guess of p_s given a response of x_s to a test mailing of size m_s?

Intuitively, we would expect

$$E(p_s|x_s, m_s) \approx \omega \frac{\alpha}{\alpha + \beta} + (1 - \omega) \frac{x_s}{m_s}$$

Bayes Theorem

- The *prior distribution* $g(p)$ captures the possible values p can take on, prior to collecting any information about the specific individual.
- The *posterior distribution* $g(p|x)$ is the conditional distribution of p, given the observed data x. It represents our updated opinion about the possible values p can take on, now that we have some information x about the specific individual.
- According to Bayes theorem:

$$g(p|x) = \frac{f(x|p)g(p)}{\int f(x|p)g(p)\,dp}$$

Bayes Theorem

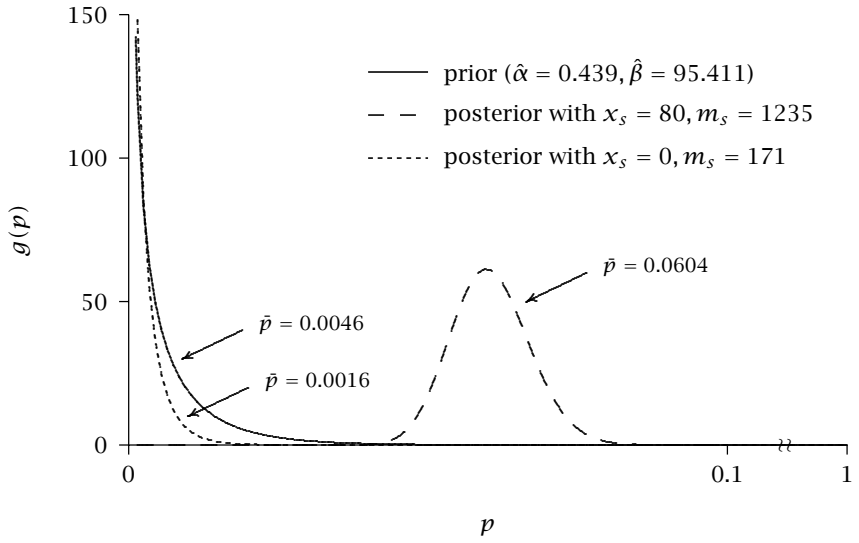
For the beta-binomial model, we have:

$$g(p_s|X_s = x_s, m_s) = \frac{\overbrace{P(X_s = x_s|m_s, p_s)}^{\text{binomial}}\overbrace{g(p_s)}^{\text{beta}}}{\underbrace{\int_0^1 P(X_s = x_s|m_s, p_s)\, g(p_s)\, dp_s}_{\text{beta-binomial}}}$$

$$= \frac{1}{B(\alpha + x_s, \beta + m_s - x_s)} p_s^{\alpha + x_s - 1}(1 - p_s)^{\beta + m_s - x_s - 1}$$

which is a beta distribution with parameters $\alpha + x_s$ and $\beta + m_s - x_s$.

Distribution of p



Applying the Model

Recall that the mean of the beta distribution is $\alpha/(\alpha+\beta)$. Therefore

$$E(p_s|X_s = x_s, m_s) = \frac{\alpha + x_s}{\alpha + \beta + m_s}$$

which can be written as

$$\left(\frac{\alpha+\beta}{\alpha+\beta+m_s}\right)\frac{\alpha}{\alpha+\beta} + \left(\frac{m_s}{\alpha+\beta+m_s}\right)\frac{x_s}{m_s}$$

- a weighted average of the test RR (x_s/m_s) and the population mean ($\alpha/(\alpha+\beta)$).
- "Regressing the test RR to the mean"

Model-Based Decision Rule

- Rollout to segments with:

$$E(p_s|X_s = x_s, m_s) > \frac{3343/10,000}{161.5} = 0.00207$$

- 66 segments pass this hurdle
- To test this model, we compare model predictions with managers' actions. (We also examine the performance of the "standard" approach.)

Results

	Standard	Manager	Model
# Segments (Rule)	51		66
# Segments (Act.)	46	71	53
Contacts	682,392	858,728	732,675
Responses	4,463	4,804	4,582
Profit	$492,651	$488,773	$495,060

Use of model results in a profit increase of $6287; 126,053 fewer contacts, saved for another offering.

Problem 3 -- Model (b)

	A	B	C	D	E	F	G	H	I
1	\alpha	0.439		B(\alpha,\beta)	0.2733				
2	\beta	95.411							
3					LL =	-200.548		cutoff	=(3343/10000)/161.5
4									
5	Segment	m_s	x_s		P(X=x|m)			E[p_s|x_s]	Roll?
6	1	34	0		0.8745	-0.1341		=(B$1+C6)/(B$1+B$2+B6)	=IF(H6>=I$3,"Y","N")
7	2	102	1		0.1656	-1.7984		=(B$1+C7)/(B$1+B$2+B7)	=IF(H7>=I$3,"Y","N")
8	3	53	0		0.8233	-0.1944		=(B$1+C8)/(B$1+B$2+B8)	=IF(H8>=I$3,"Y","N")
9	4	145	2		0.0769	-2.5647		=(B$1+C9)/(B$1+B$2+B9)	=IF(H9>=I$3,"Y","N")
10	5	1254	62		0.0002	-8.7933		=(B$1+C10)/(B$1+B$2+B10)	=IF(H10>=I$3,"Y","N")
11	6	144	7		0.003	-5.8046		=(B$1+C11)/(B$1+B$2+B11)	=IF(H11>=I$3,"Y","N")
12	7	1235	80		0	-10.4029		=(B$1+C12)/(B$1+B$2+B12)	=IF(H12>=I$3,"Y","N")
13	8	573	34		0.0001	-8.8693		=(B$1+C13)/(B$1+B$2+B13)	=IF(H13>=I$3,"Y","N")
14	9	1083	24		0.0036	-5.6216		=(B$1+C14)/(B$1+B$2+B14)	=IF(H14>=I$3,"Y","N")
15	10	352	5		0.0301	-3.5032		=(B$1+C15)/(B$1+B$2+B15)	=IF(H15>=I$3,"Y","N")
16	11	817	7		0.0281	-3.5719		=(B$1+C16)/(B$1+B$2+B16)	=IF(H16>=I$3,"Y","N")
17	12	118	0		0.7018	-0.3541		=(B$1+C17)/(B$1+B$2+B17)	=IF(H17>=I$3,"Y","N")
18	13	1049	3		0.0665	-2.7102		=(B$1+C18)/(B$1+B$2+B18)	=IF(H18>=I$3,"Y","N")
19	14	452	3		0.0674	-2.6978		=(B$1+C19)/(B$1+B$2+B19)	=IF(H19>=I$3,"Y","N")
20	15	338	2		0.0991	-2.3113		=(B$1+C20)/(B$1+B$2+B20)	=IF(H20>=I$3,"Y","N")
21	16	168	0		0.6398	-0.4466		=(B$1+C21)/(B$1+B$2+B21)	=IF(H21>=I$3,"Y","N")
22	17	242	3		0.0547	-2.9067		=(B$1+C22)/(B$1+B$2+B22)	=IF(H22>=I$3,"Y","N")
23	18	185	1		0.1809	-1.7098		=(B$1+C23)/(B$1+B$2+B23)	=IF(H23>=I$3,"Y","N")
24	19	116	0		0.7047	-0.3499		=(B$1+C24)/(B$1+B$2+B24)	=IF(H24>=I$3,"Y","N")
25	20	69	1		0.1459	-1.925		=(B$1+C25)/(B$1+B$2+B25)	=IF(H25>=I$3,"Y","N")
26	21	193	1		0.1812	-1.708		=(B$1+C26)/(B$1+B$2+B26)	=IF(H26>=I$3,"Y","N")
27	22	82	1		0.1553	-1.8623		=(B$1+C27)/(B$1+B$2+B27)	=IF(H27>=I$3,"Y","N")
28	23	265	1		0.1804	-1.7125		=(B$1+C28)/(B$1+B$2+B28)	=IF(H28>=I$3,"Y","N")
29	24	171	0		0.6366	-0.4516		=(B$1+C29)/(B$1+B$2+B29)	=IF(H29>=I$3,"Y","N")
30	25	1554	7		0.0309	-3.4774		=(B$1+C30)/(B$1+B$2+B30)	=IF(H30>=I$3,"Y","N")
31	26	1339	4		0.0511	-2.9745		=(B$1+C31)/(B$1+B$2+B31)	=IF(H31>=I$3,"Y","N")
32	27	1167	4		0.052	-2.9572		=(B$1+C32)/(B$1+B$2+B32)	=IF(H32>=I$3,"Y","N")
33	28	621	2		0.0981	-2.3219		=(B$1+C33)/(B$1+B$2+B33)	=IF(H33>=I$3,"Y","N")
34	29	1013	1		0.1367	-1.9902		=(B$1+C34)/(B$1+B$2+B34)	=IF(H34>=I$3,"Y","N")
35	30	544	1		0.1621	-1.8195		=(B$1+C35)/(B$1+B$2+B35)	=IF(H35>=I$3,"Y","N")
36	31	731	1		0.1505	-1.8936		=(B$1+C36)/(B$1+B$2+B36)	=IF(H36>=I$3,"Y","N")

Concepts and Tools Introduced

- "Choice" processes
- The Beta Binomial model
- "Regression-to-the-mean" and the use of models to capture such an effect
- Bayes theorem (and "empirical Bayes" methods)
- Using "empirical Bayes" methods in the development of targeted marketing campaigns

Further Reading

Colombo, Richard and Donald G. Morrison (1988), "Blacklisting Social Science Departments with Poor Ph.D. Submission Rates," *Management Science*, **34** (June), 696–706.

Morwitz, Vicki G. and David C. Schmittlein (1998), "Testing New Direct Marketing Offerings: The Interplay of Management Judgment and Statistical Models," *Management Science*, **44** (May), 610–628.

Sabavala, Darius J. and Donald G. Morrison (1977), "A Model of TV Show Loyalty," *Journal of Advertising Research*, **17** (December), 35–43.

Further Applications and Tools/ Modeling Issues

Recap

- The preceding three problems introduce simple models for three behavioral processes:
  - Timing ⟶ "when"
  - Counting ⟶ "how many"
  - "Choice" ⟶ "whether/which"
- Each of these simple models has multiple applications.
- More complex behavioral phenomena can be captured by combining models from each of these processes.

Further Applications: Timing Models

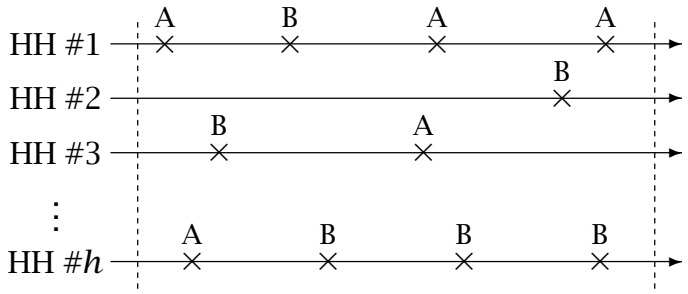
- Repeat purchasing of new products
- Response times:
  - Coupon redemptions
  - Survey response
  - Direct mail (response, returns, repeat sales)
- Customer retention/attrition
- Other durations:
  - Salesforce job tenure
  - Length of web site browsing session

Further Applications: Count Models

- Repeat purchasing
- Customer concentration (“80/20” rules)
- Salesforce productivity/allocation
- Number of page views during a web site browsing session

Further Applications: "Choice" Models

- Brand choice



- Media exposure
- Multibrand choice (BB → Dirichlet Multinomial)
- Taste tests (discrimination tests)
- "Click-through" behavior

Integrated Models

- Counting + Timing
  - catalog purchases (purchasing|"alive" & "death" process)
  - "stickiness" (# visits & duration/visit)
- Counting + Counting
  - purchase volume (# transactions & units/transaction)
  - page views/month (# visits & pages/visit)
- Counting + Choice
  - brand purchasing (category purchasing & brand choice)
  - "conversion" behavior (# visits & buy/not-buy)

A Template for Integrated Models

		Stage 2		
		Counting	Timing	Choice
Stage 1	Counting			
	Timing			
	Choice			

Further Issues

Relaxing usual assumptions:

- Non-exponential purchasing (greater regularity) ⟶ non-Poisson counts
- Non-gamma/beta heterogeneity (e.g., “hard core” nonbuyers, “hard core” loyals)
- Nonstationarity — latent traits vary over time

The basic models are quite robust to these departures.

Extensions

- Latent class/finite mixture models
- Introducing covariate effects
- Hierarchical Bayes methods

The Excel spreadsheets associated with this tutorial, along with electronic copies of the tutorial materials and a "supplementary materials handout" that works through the math of the models, can be found at:

`http://brucehardie.com/talks.html`

An annotated list of key books for those interested in applied probability modelling can be found at:

`http://brucehardie.com/notes/001/`

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

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

or

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

The random variable X_j ($j = 1, \ldots, k$), 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 $\text{cov}(X_j, X_{j'}) = -np_jp_{j'}$.

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

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

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:

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

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 $\text{var}(\lambda) = r/\alpha^2$. Therefore the variance of the NBD is

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

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

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

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

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

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

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

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 $P(X = 0)$:

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

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

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

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

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

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

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

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 $\times$ 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

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

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.