An Introduction to Probability Models for Marketing Research

Peter S. Fader
University of Pennsylvania

Bruce G. S. Hardie
London Business School

27th Annual Advanced Research Techniques Forum
June 26–29, 2016

Problem 1:
Projecting Customer Retention Rates

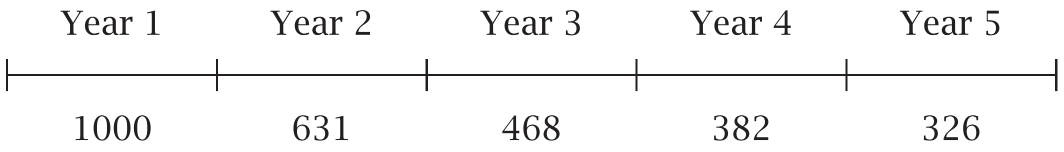
(Modelling Discrete-Time Duration Data)

Motivating Problem

Consider a company with a subscription-based business model. 1000 customers are acquired at the beginning of Year 1 with the following renewal patterns:

ID	Year 1	Year 2	Year 3	Year 4	Year 5
0001	1	1	0	0	0
0002	1	0	0	0	0
0003	1	1	1	0	0
0004	1	1	0	0	0
0005	1	1	1	1	1
⋮		⋮		⋮	
0998	1	0	0	0	0
0999	1	1	1	0	0
1000	1	0	0	0	0
	1000	631	468	382	326

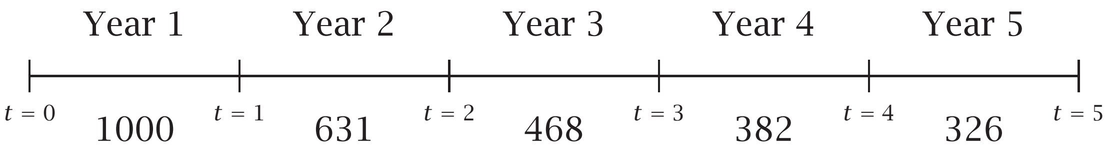
Motivating Problem



Year 1	Year 2	Year 3	Year 4	Year 5
1000	631	468	382	326

- How many customers will “survive” to Year 6, 7, . . . , 13?
- What will the retention rates for this cohort look like for the next 8 years?

Notation and Terminology



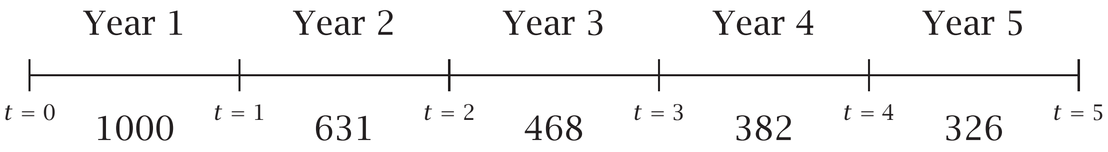
The *survivor function* $S(t)$ is the proportion of the cohort that continue as a customer beyond t.

$$S(0) = ?$$

$$S(1) = ?$$

$$S(2) = ?$$

Notation and Terminology



The *retention rate* is the ratio of customers retained to the number at risk.

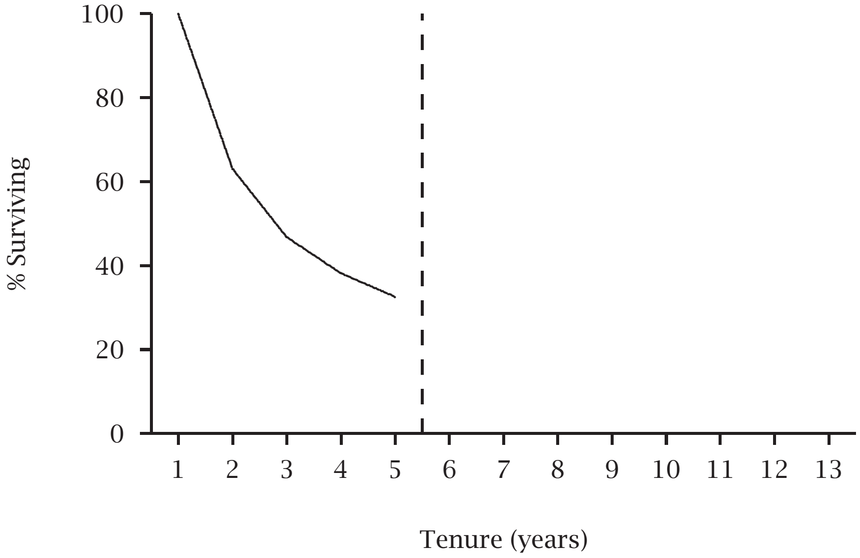
$$r(1) = ?$$

$$r(2) = ?$$

For survivor function $S(t)$, $r(t) = \dfrac{S(t)}{S(t-1)}$.

Modelling Objective

We want to derive a mathematical expression for $S(t)$, which can then be used to generate the desired forecasts.



Natural Starting Point

Project the survival curve using functions of time:

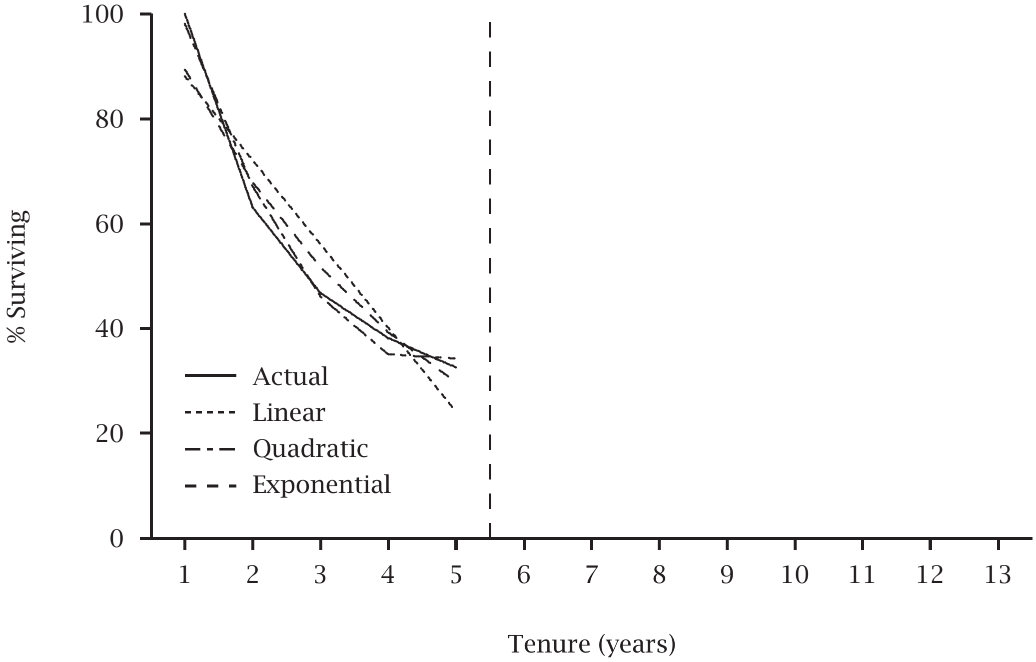
- Consider linear, quadratic, and exponential functions
- Let $y = S(t)$:

$$y = 0.881 - 0.160t \qquad R^2 = 0.868$$

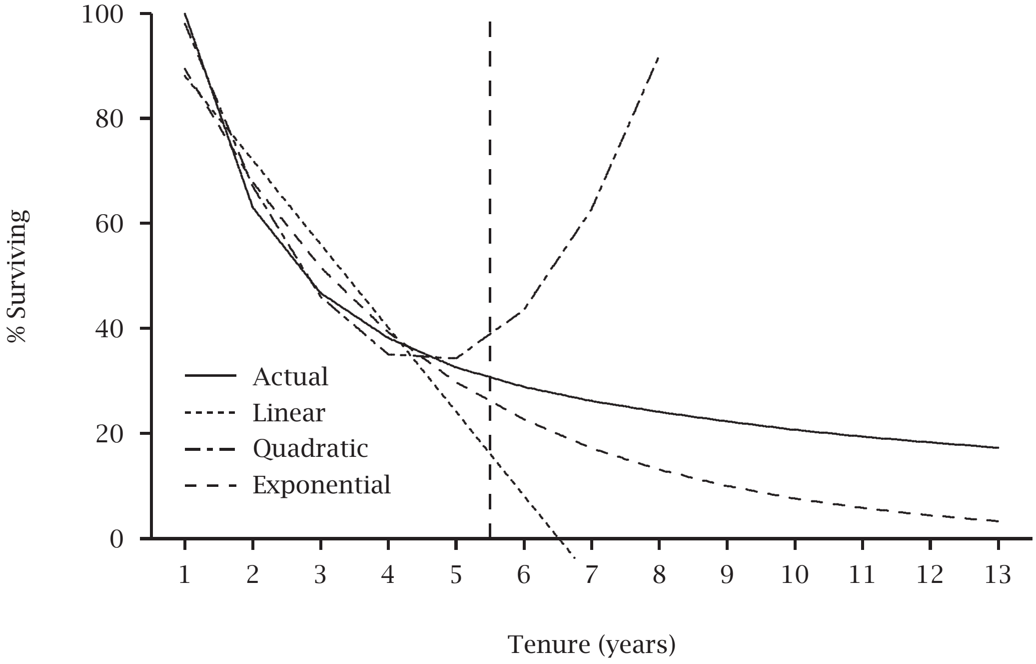
$$y = 0.981 - 0.361t + 0.050t^2 \qquad R^2 = 0.989$$

$$\ln(y) = -0.112 - 0.274t \qquad R^2 = 0.954$$

Model Fit



Survival Curve Projections



Developing a Better Model (I)

At the end of each contract period, a customer makes the renewal decision by tossing a coin: H → renew, T → don't renew

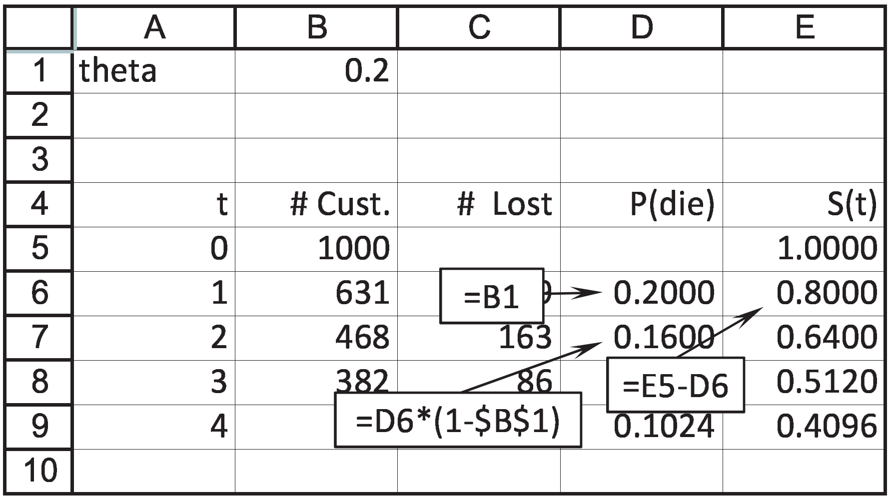
Length of relationship			
1 period	T		
2 periods	H	T	
3 periods	H	H	T
. . .			

$$P(t \text{ periods}) = \begin{cases} P(\mathsf{T}) & t = 1 \\ P(\mathsf{H}) \times P(t-1 \text{ periods}) & t = 2, 3, \ldots \end{cases}$$

Developing a Better Model (I)

i) $P(\mathsf{H}) = 1 - \theta$ is constant and unobserved.

ii) All customers have the same "churn probability" θ.



	A	B	C	D	E
1	theta	0.2			
2					
3					
4	t	# Cust.	# Lost	P(die)	S(t)
5	0	1000			1.0000
6	1	631		0.2000	0.8000
7	2	468	163	0.1600	0.6400
8	3	382	86		0.5120
9	4			0.1024	0.4096
10					

=B1

=D6*(1-$B$1)

=E5-D6

Developing a Better Model (I)

More formally:

- Let the random variable T denote the duration of the customer's relationship with the firm.
- We assume that the random variable T is distributed geometric with parameter θ:

$$
\begin{aligned}
P(T = t \mid \theta) &= \theta(1-\theta)^{t-1}, \quad t = 1, 2, 3, \ldots \\
S(t \mid \theta) &= P(T > t \mid \theta) \\
&= (1-\theta)^{t}, \quad t = 0, 1, 2, 3, \ldots
\end{aligned}
$$

Estimating Model Parameters

Assuming

i) the observed data were generated according to the "coin flipping" story of contract renewal, and

ii) we know $P(\mathsf{T}) = \theta$,

the probability of the observed pattern of renewals is:

$$
\begin{aligned}
&[P(T = 1 \mid \theta)]^{369}[P(T = 2 \mid \theta)]^{163}[P(T = 3 \mid \theta)]^{86} \\
&\qquad \times [P(T = 4 \mid \theta)]^{56}[S(t \mid \theta)]^{326} \\
&= [\theta]^{369}[\theta(1-\theta)]^{163}[\theta(1-\theta)^{2}]^{86} \\
&\qquad \times [\theta(1-\theta)^{3}]^{56}[(1-\theta)^{4}]^{326}
\end{aligned}
$$

Estimating Model Parameters

- Suppose we have two candidate coins:

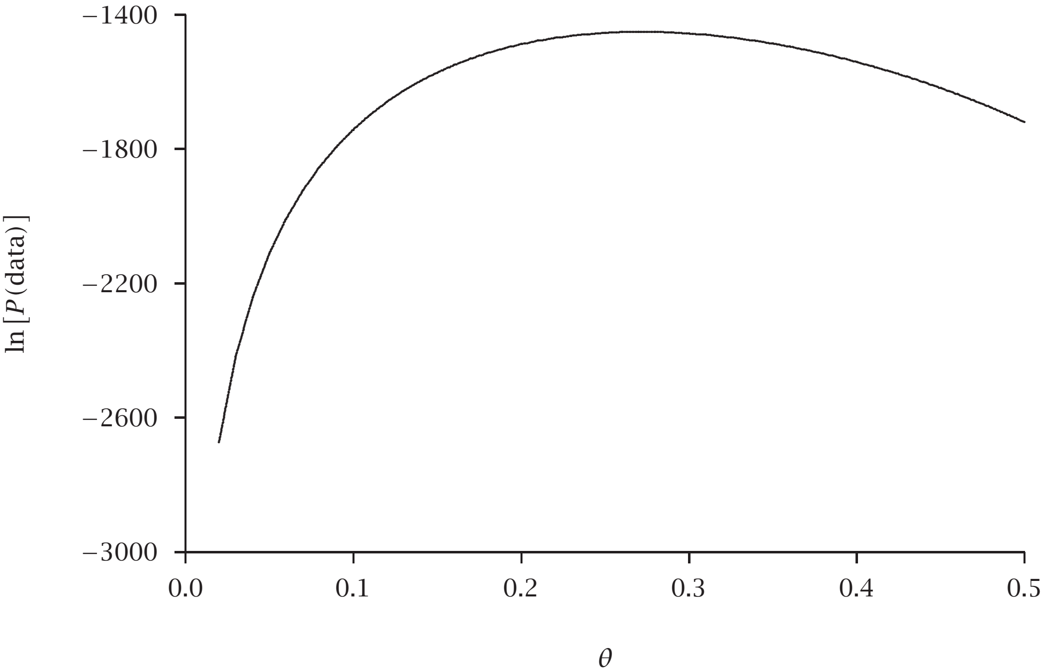
Coin A: $\theta = 0.2$

Coin B: $\theta = 0.5$

- Which coin is more likely to have generated the observed pattern of renewals across this set of 1000 customers?

θ	$P(\text{data} \mid \theta)$	$\ln[P(\text{data} \mid \theta)]$
0.2	6.00×10^{-647}	−1488.0
0.5	1.40×10^{-747}	−1719.7

Estimating Model Parameters



Estimating Model Parameters

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

- The likelihood function is defined as the probability of observing the data for a given set of the (unknown) model parameters.
- It is computed using the model and is viewed as a function of the model parameters:

$$L(\text{parameters} \mid \text{data}) = p(\text{data} \mid \text{parameters})\,.$$

- For a given dataset, the maximum likelihood estimates of the model parameters are those values that maximize $L(\cdot)$.
- It is typically more convenient to use the natural logarithm of the likelihood function — the log-likelihood function.

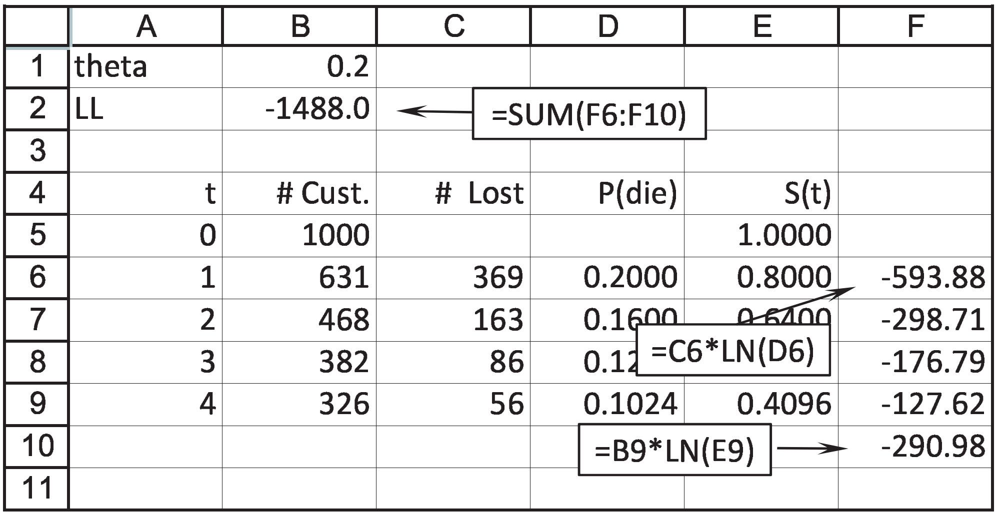
Estimating Model Parameters

The log-likelihood function is given by:

$$\begin{aligned} LL(\theta \mid \text{data}) = {} & 369 \times \ln[P(T = 1 \mid \theta)] + \\ & 163 \times \ln[P(T = 2 \mid \theta)] + \\ & 86 \times \ln[P(T = 3 \mid \theta)] + \\ & 56 \times \ln[P(T = 4 \mid \theta)] + \\ & 326 \times \ln[S(4 \mid \theta)] \end{aligned}$$

The maximum value of the log-likelihood function is $LL = -1451.2$, which occurs at $\hat{\theta} = 0.272$.

Estimating Model Parameters



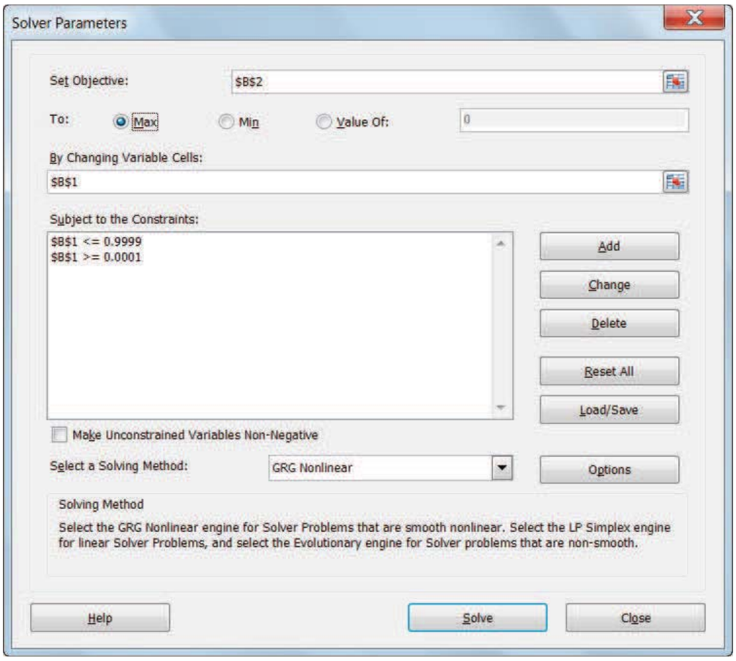
	A	B	C	D	E	F
1	theta	0.2				
2	LL	-1488.0				
3						
4	t	# Cust.	# Lost	P(die)	S(t)	
5	0	1000			1.0000	
6	1	631	369	0.2000	0.8000	-593.88
7	2	468	163	0.1600	0.6400	-298.71
8	3	382	86	0.12		-176.79
9	4	326	56	0.1024	0.4096	-127.62
10						-290.98
11						

=SUM(F6:F10)

=C6*LN(D6)

=B9*LN(E9)

Estimating Model Parameters



Solver Parameters

Set Objective: $B$2

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells: $B$1

Subject to the Constraints:

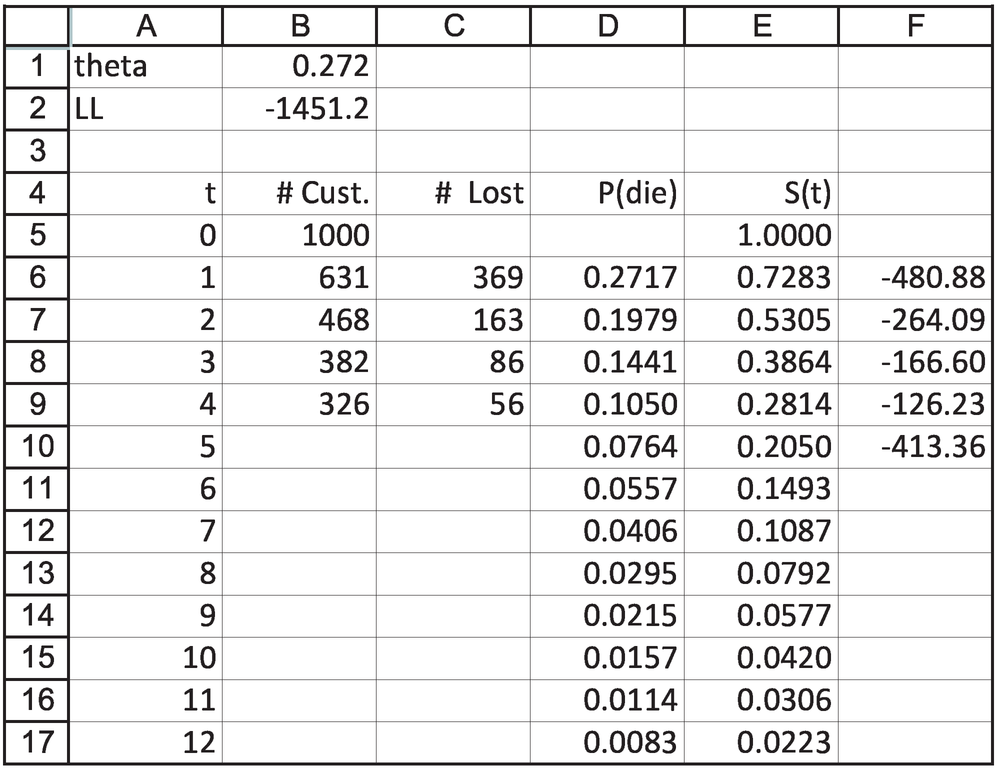
$B$1 <= 0.9999
$B$1 >= 0.0001

☐ Make Unconstrained Variables Non-Negative

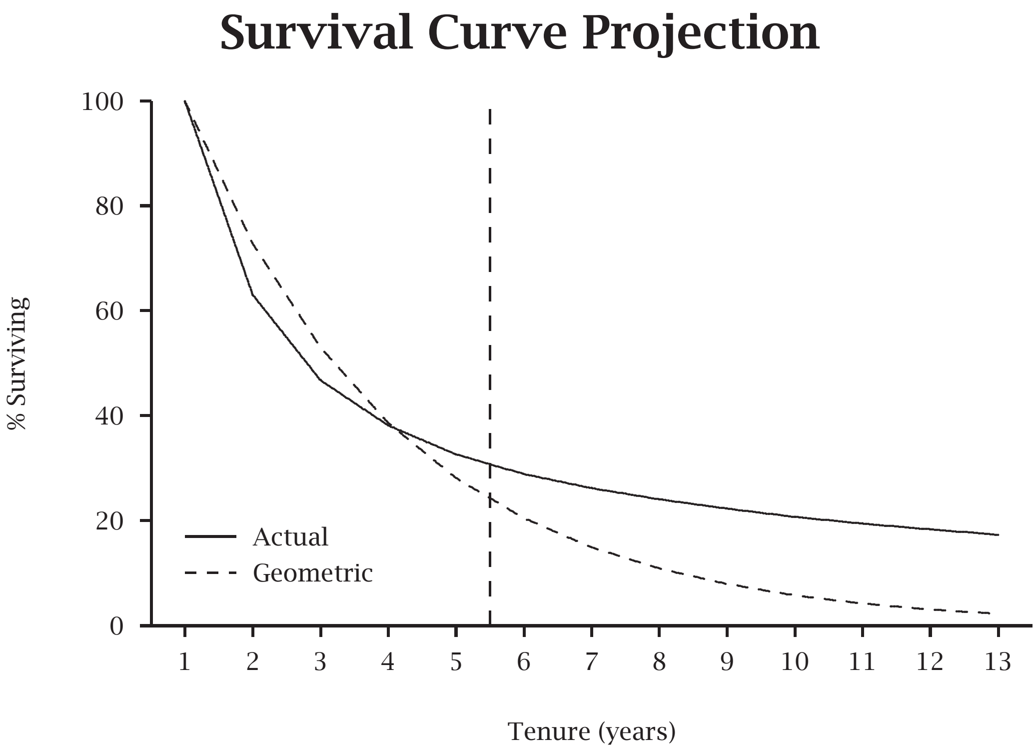
Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

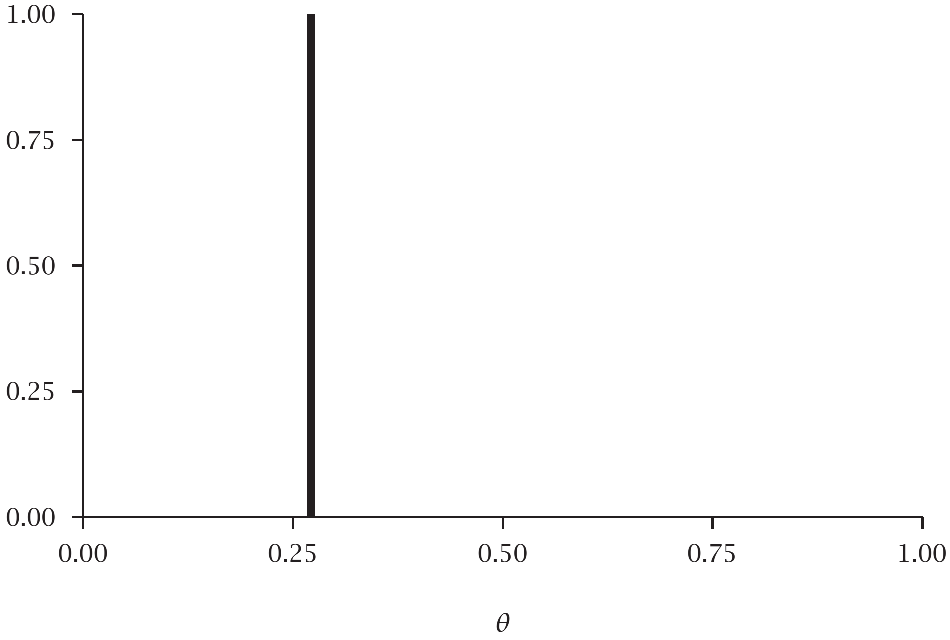


	A	B	C	D	E	F
1	theta	0.272				
2	LL	-1451.2				
3						
4	t	# Cust.	# Lost	P(die)	S(t)	
5	0	1000			1.0000	
6	1	631	369	0.2717	0.7283	-480.88
7	2	468	163	0.1979	0.5305	-264.09
8	3	382	86	0.1441	0.3864	-166.60
9	4	326	56	0.1050	0.2814	-126.23
10	5			0.0764	0.2050	-413.36
11	6			0.0557	0.1493	
12	7			0.0406	0.1087	
13	8			0.0295	0.0792	
14	9			0.0215	0.0577	
15	10			0.0157	0.0420	
16	11			0.0114	0.0306	
17	12			0.0083	0.0223	



What’s wrong with this story of customer contract-renewal behavior?

Visualizing Parameter Estimates



Developing a Better Model (II)

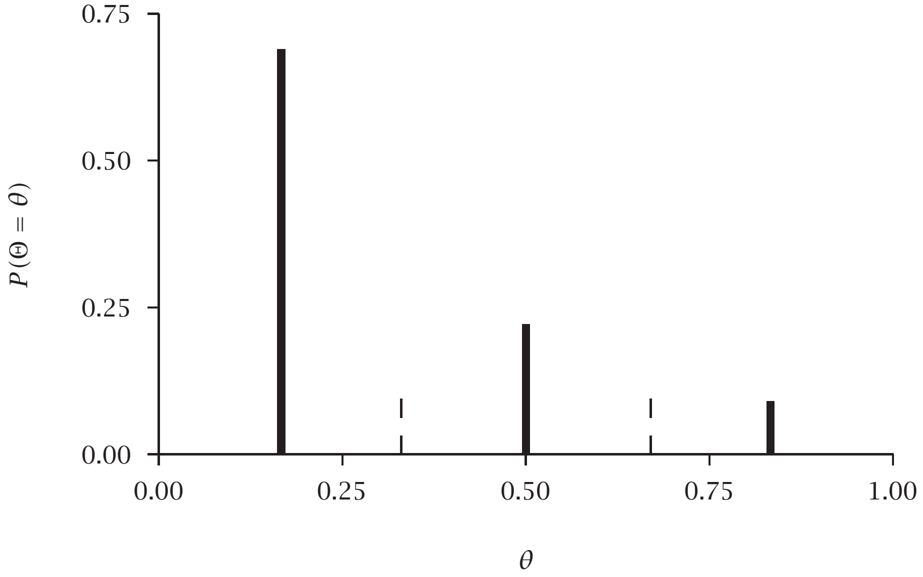
- Consider the following story of customer behavior:
  i) At the end of each period, an individual renews his contract with (constant and unobserved) probability $1 - \theta$.
  ii) "Churn probabilities" vary across customers.
- Since we don't know any given customer's true value of θ, we treat it as a realization of a random variable (Θ).
- We need to specify a probability distribution that captures how θ varies across customers (by giving us the probability of each possible value of θ).

Accommodating Heterogeneity in θ

- Suppose we divide $(0, 1)$ into three sub-intervals of equal width: 0.000–0.333, 0.333–0.667, 0.667–1.000
- We allow θ to take on the value of the mid-point of each sub-interval:

$$\Theta = \begin{cases} 0.167 & \text{with probability } P(\Theta = 0.167) \\ 0.500 & \text{with probability } P(\Theta = 0.500) \\ 0.833 & \text{with probability } P(\Theta = 0.833) \end{cases}$$

Accommodating Heterogeneity in θ



$$P(\Theta = 0.167) + P(\Theta = 0.500) + P(\Theta = 0.833) = 1$$

Accommodating Heterogeneity in θ

What is the probability that a randomly chosen new customer will cancel their contract at the end of period t?

i) If we knew their θ, it would simply be $P(T = t \mid \theta)$.

ii) Since we only know the distribution of Θ across the population, we evaluate $P(T = t \mid \theta)$ for each possible value of θ, weighting it by the probability of a randomly chosen new customer having that value of θ:

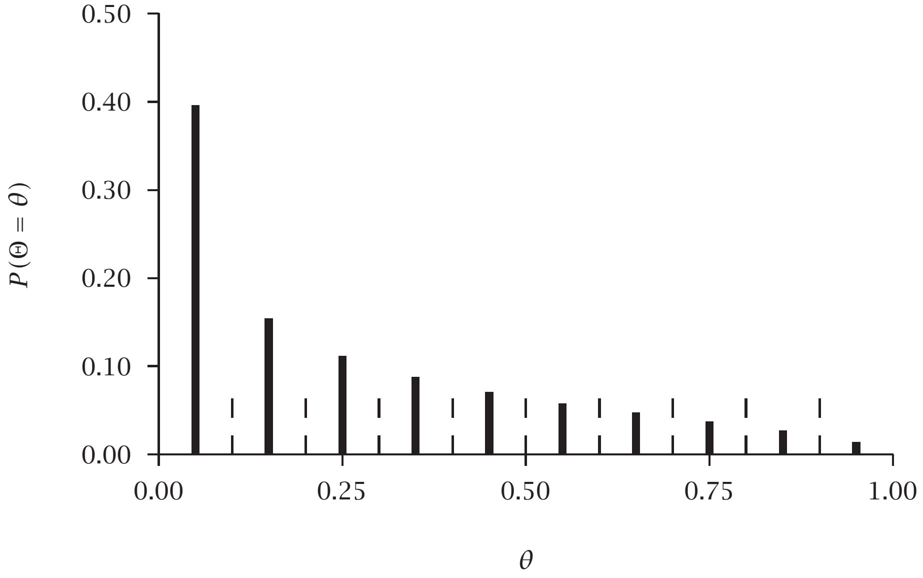
$$\begin{aligned} P(T = t) = {} & P(T = t \mid \Theta = 0.167)\, P(\Theta = 0.167) \\ & + P(T = t \mid \Theta = 0.500)\, P(\Theta = 0.500) \\ & + P(T = t \mid \Theta = 0.833)\, P(\Theta = 0.833) \end{aligned}$$

Accommodating Heterogeneity in θ

- The problem with this is that $E(\Theta)$ is bounded between 0.167 and 0.833.
- To allow for greater flexibility, suppose we divide $(0,1)$ into ten sub-intervals of equal width and allow θ to take on the value of the mid-point of each sub-interval:

$$\Theta = \begin{cases} 0.05 & \text{with probability } P(\Theta = 0.05) \\ 0.15 & \text{with probability } P(\Theta = 0.15) \\ & \quad \cdots \\ 0.95 & \text{with probability } P(\Theta = 0.95) \end{cases}$$

Accommodating Heterogeneity in θ



$$P(\Theta = 0.05) + P(\Theta = 0.15) + \cdots + P(\Theta = 0.95) = 1$$

Accommodating Heterogeneity in θ

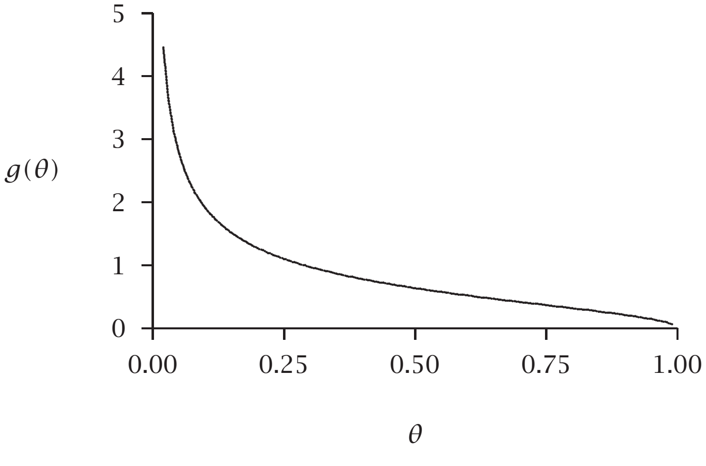
- The probability that a randomly chosen new customer will cancel their contract at the end of period t is

$$\begin{aligned} P(T=t) = {} & P(T=t \mid \Theta=0.05)\, P(\Theta=0.05) \\ & + P(T=t \mid \Theta=0.15)\, P(\Theta=0.15) \\ & + \cdots \\ & + P(T=t \mid \Theta=0.95)\, P(\Theta=0.95) \end{aligned}$$

- This ten sub-interval solution is more flexible — $E(\Theta)$ is now bounded between 0.05 and 0.95 — but is less parsimonious.

Accommodating Heterogeneity in θ

In order to increase flexibility without sacrificing parsimony, we let the number of sub-intervals go to infinity and represent the probabilities in terms of a simple continuous function $g(\theta)$ defined over $(0, 1)$:



Accommodating Heterogeneity in θ

- The probability of getting a specific value of θ is infinitesimally small.
- Discrete ⟶ continuous $\Rightarrow \sum \longrightarrow \int$
- By definition, $P(0 \le \Theta \le 1) = 1 \Leftrightarrow$ area under the curve, $\int_0^1 g(\theta)\,d\theta$, equals one.
- For a randomly chosen customer,

$$P(T = t) = \int_0^1 P(T = t \mid \theta) g(\theta)\,d\theta$$

The Beta Distribution

- The beta distribution is a flexible (and mathematically convenient) two-parameter distribution bounded between 0 and 1:

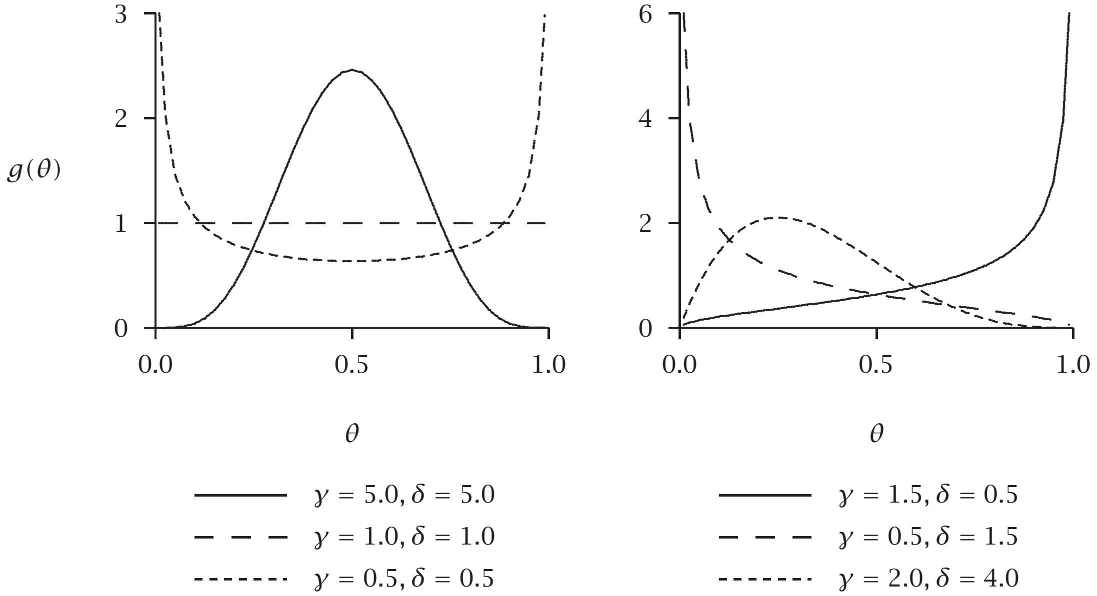
$$g(\theta \mid \gamma, \delta) = \frac{\theta^{\gamma-1}(1-\theta)^{\delta-1}}{B(\gamma, \delta)},$$

  where $\gamma, \delta > 0$ and $B(\gamma, \delta)$ is the beta function.
- The mean and variance of the beta distribution are

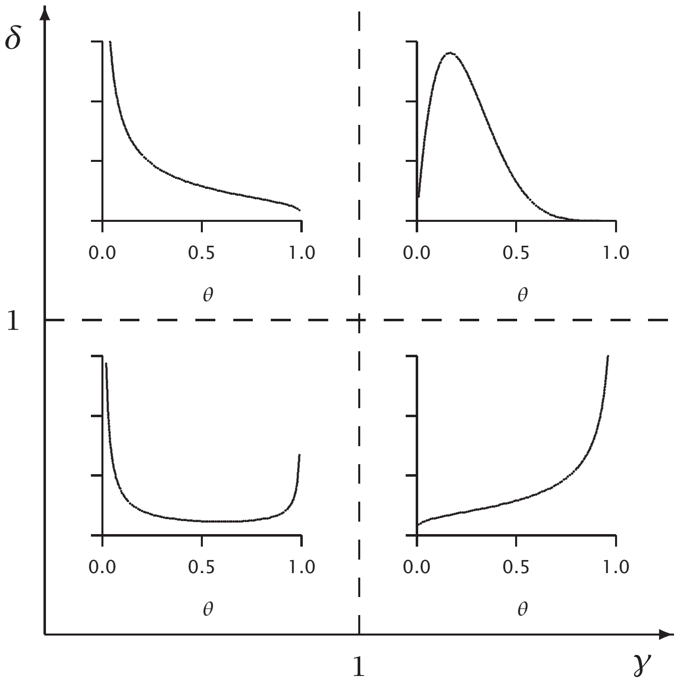
$$E(\Theta) = \frac{\gamma}{\gamma + \delta}$$

$$\text{var}(\Theta) = \frac{\gamma\delta}{(\gamma+\delta)^2(\gamma+\delta+1)}$$

Illustrative Beta Distributions



Five General Shapes of the Beta Distribution



The Beta Function

- The beta function $B(\gamma,\delta)$ is defined by the integral

$$B(\gamma,\delta) := \int_0^1 t^{\gamma-1}(1-t)^{\delta-1}dt, \quad \gamma,\delta > 0\,,$$

  and can be expressed in terms of gamma functions:

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

- The gamma function $\Gamma(\gamma)$ is a generalized factorial, which has the recursive property $\Gamma(\gamma+1) = \gamma\Gamma(\gamma)$. Since $\Gamma(0) = 1$, $\Gamma(n) = (n-1)!$ for positive integer n.

Developing a Better Model (II)

For a randomly chosen individual,

$$\begin{aligned} P(T = t \mid \gamma,\delta) &= \int_0^1 P(T = t \mid \theta) g(\theta \mid \gamma,\delta)\, d\theta \\ &= \int_0^1 \theta(1-\theta)^{t-1} \frac{\theta^{\gamma-1}(1-\theta)^{\delta-1}}{B(\gamma,\delta)}\, d\theta \\ &= \frac{1}{B(\gamma,\delta)} \int_0^1 \theta^{\gamma}(1-\theta)^{\delta+t-2}\, d\theta \\ &= \frac{B(\gamma+1,\delta+t-1)}{B(\gamma,\delta)}\,. \end{aligned}$$

Developing a Better Model (II)

Similarly,

$$\begin{aligned} S(t \mid \gamma, \delta) &= \int_0^1 S(t \mid \theta) g(\theta \mid \gamma, \delta)\, d\theta \\ &= \int_0^1 (1-\theta)^t \frac{\theta^{\gamma-1}(1-\theta)^{\delta-1}}{B(\gamma,\delta)}\, d\theta \\ &= \frac{1}{B(\gamma,\delta)} \int_0^1 \theta^{\gamma-1}(1-\theta)^{\delta+t-1}\, d\theta \\ &= \frac{B(\gamma, \delta+t)}{B(\gamma,\delta)}\,. \end{aligned}$$

We call this *continuous mixture* model the beta-geometric (BG) distribution.

Developing a Better Model (II)

We can compute BG probabilities using the following forward-recursion formula from $P(T=1)$:

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

Estimating Model Parameters

Assuming

i) the observed data were generated according to the heterogeneous "coin flipping" story of contract renewal, and

ii) we know γ and δ,

the probability of the observed pattern of renewals is:

$$[P(T=1\mid\gamma,\delta)]^{369}[P(T=2\mid\gamma,\delta)]^{163}[P(T=3\mid\gamma,\delta)]^{86}$$
$$\times[P(T=4\mid\gamma,\delta)]^{56}[S(4\mid\gamma,\delta)]^{326}$$

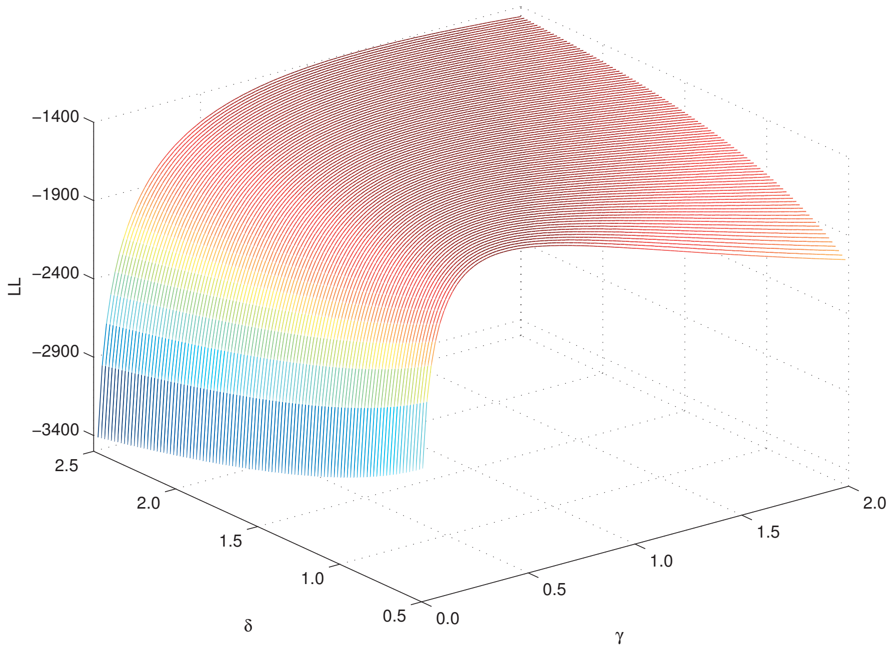
Estimating Model Parameters

The log-likelihood function is given by:

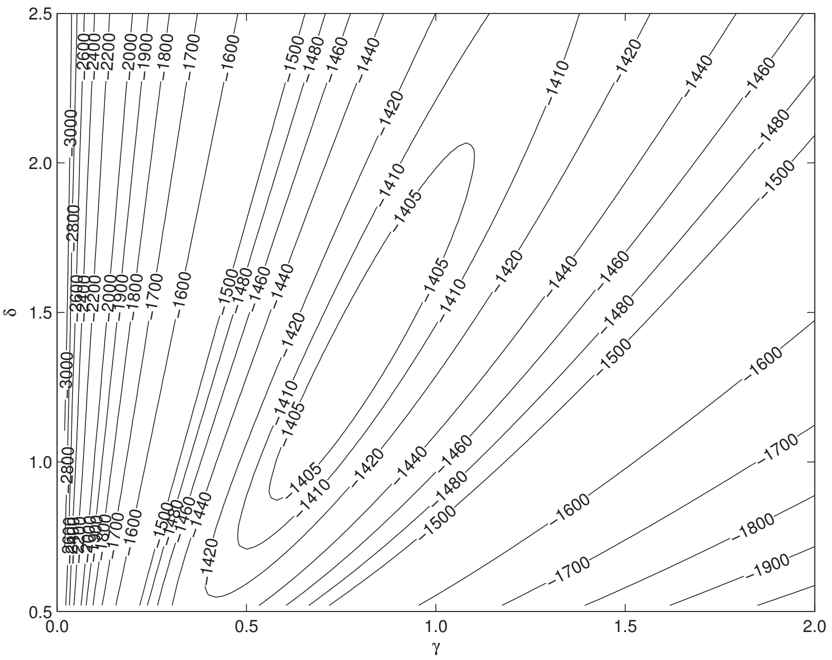
$$\begin{aligned} LL(\gamma,\delta\mid\text{data}) = {} & 369\times\ln[P(T=1\mid\gamma,\delta)]+\\ & 163\times\ln[P(T=2\mid\gamma,\delta)]+\\ & 86\times\ln[P(T=3\mid\gamma,\delta)]+\\ & 56\times\ln[P(T=4\mid\gamma,\delta)]+\\ & 326\times\ln[S(4\mid\gamma,\delta)]\end{aligned}$$

The maximum value of the log-likelihood function is $LL=-1401.6$, which occurs at $\hat{\gamma}=0.764$ and $\hat{\delta}=1.296$.

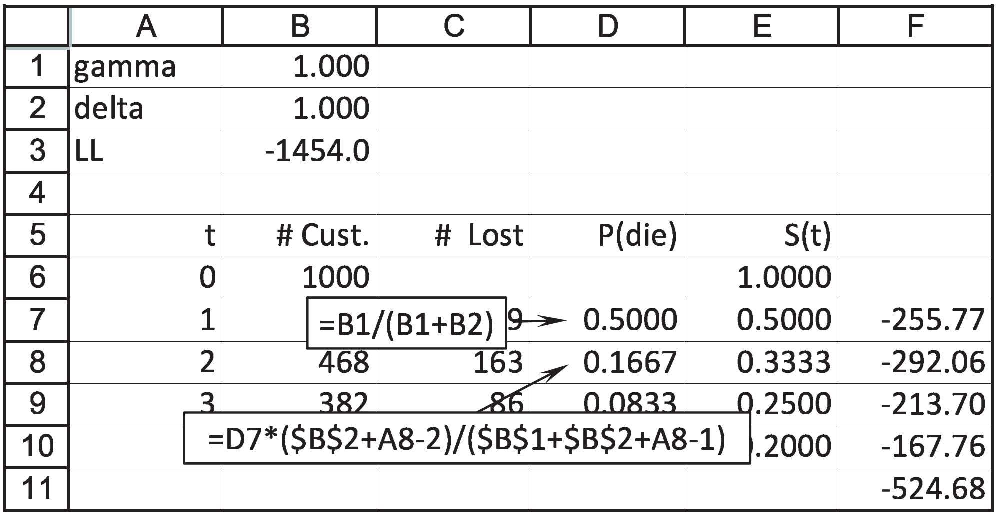
Surface Plot of BG LL Function



Contour Plot of BG LL Function



Estimating Model Parameters

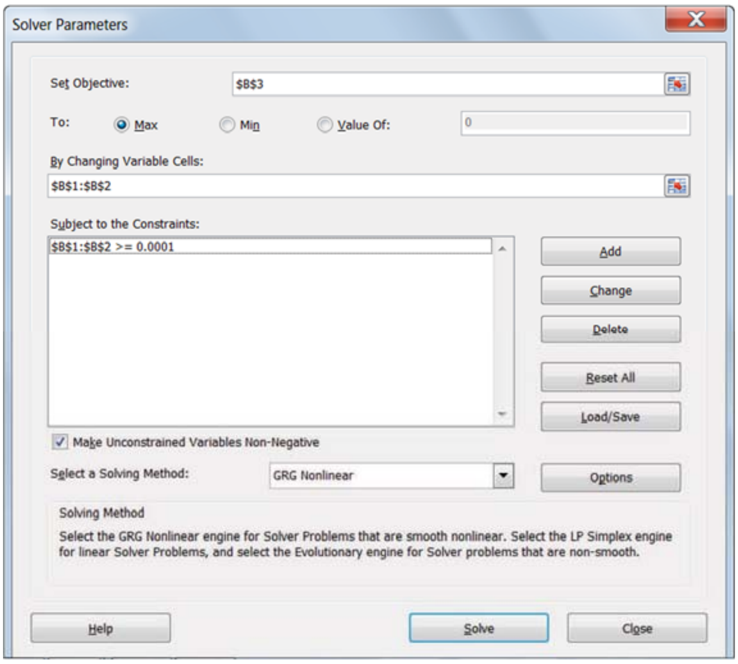


	A	B	C	D	E	F
1	gamma	1.000				
2	delta	1.000				
3	LL	-1454.0				
4						
5	t	# Cust.	# Lost	P(die)	S(t)	
6	0	1000			1.0000	
7	1		9	0.5000	0.5000	-255.77
8	2	468	163	0.1667	0.3333	-292.06
9	3	382	86	0.0833	0.2500	-213.70
10					.2000	-167.76
11						-524.68

=B1/(B1+B2)

=D7*($B$2+A8-2)/($B$1+$B$2+A8-1)

Estimating Model Parameters



Set Objective: $B$3

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells: $B$1:$B$2

Subject to the Constraints:
$B$1:$B$2 >= 0.0001

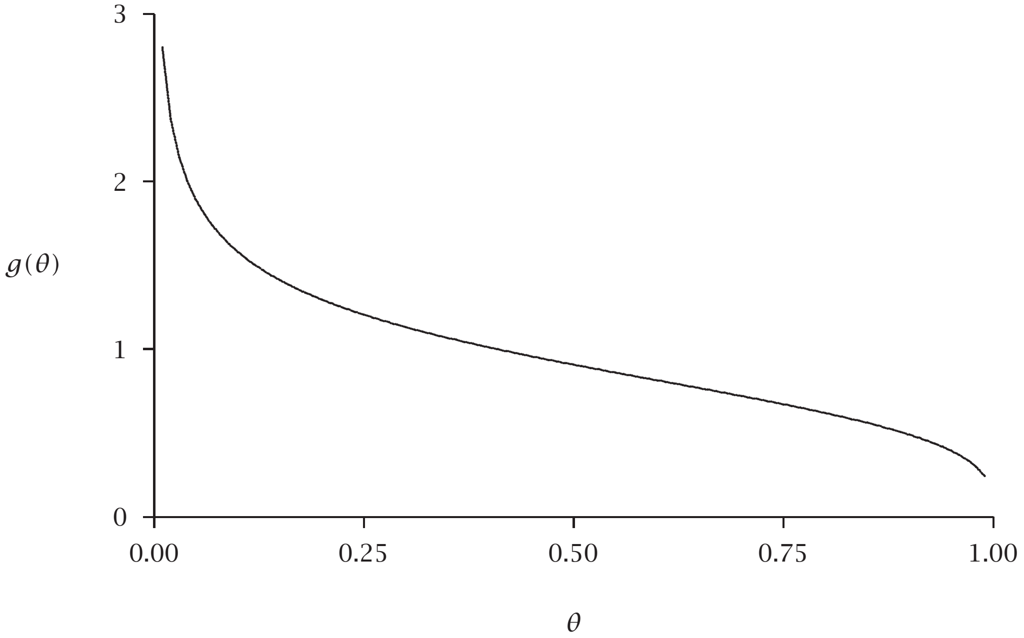
☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: GRG Nonlinear

Solving Method

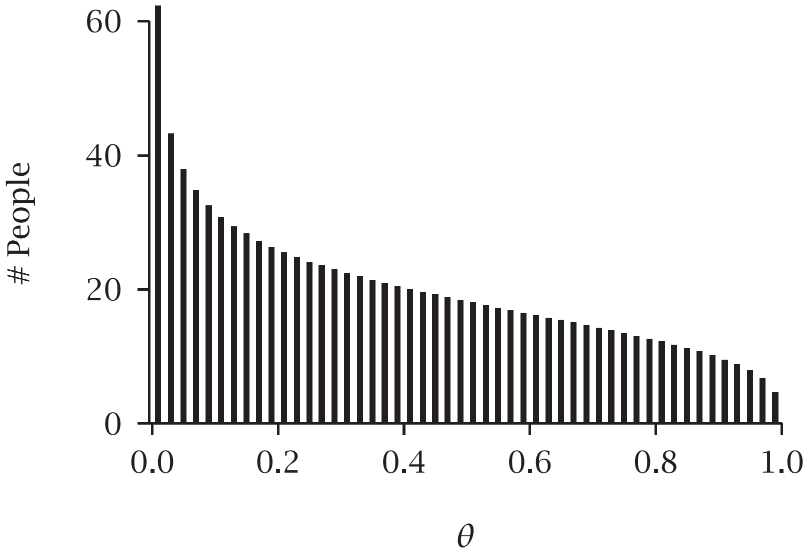
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Estimated Distribution of Churn Probabilities



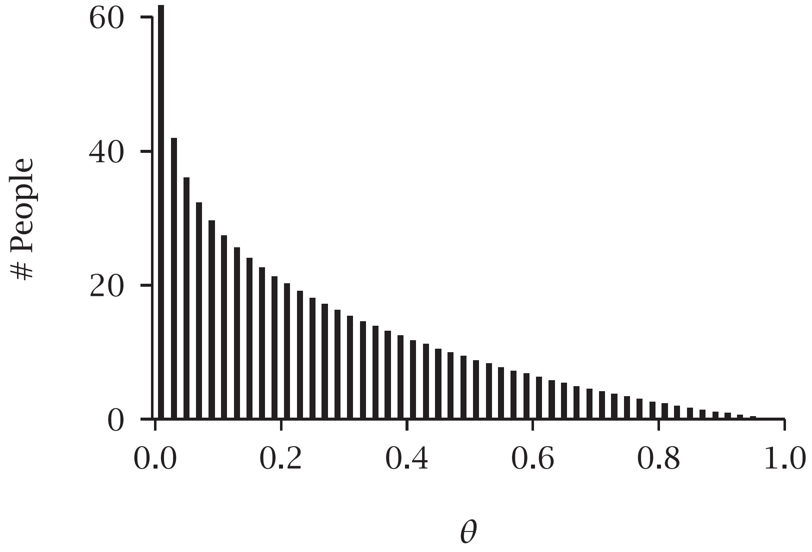
$$\hat{\gamma} = 0.764, \hat{\delta} = 1.296, \widehat{E(\Theta)} = 0.371$$

Year 1



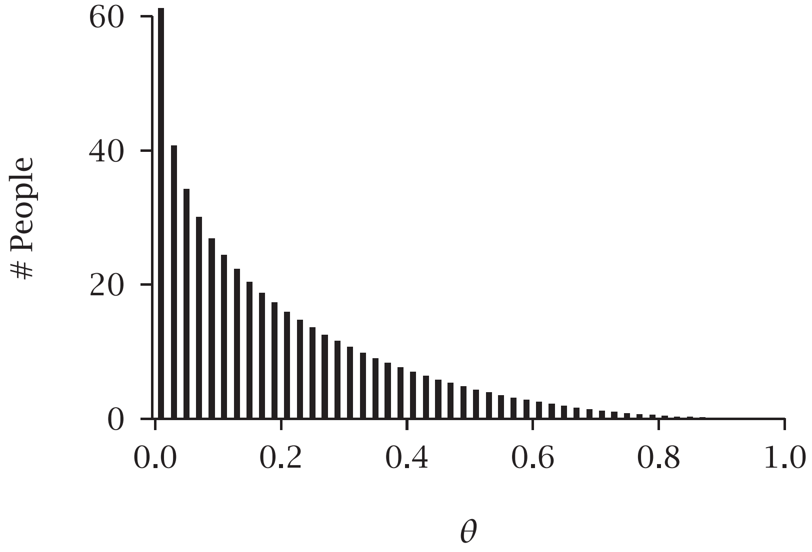
$E(\Theta) = 0.371 \longrightarrow$ expect $1000 \times (1 - 0.371) = 629$ customers to renew at the end of Year 1.

Year 2



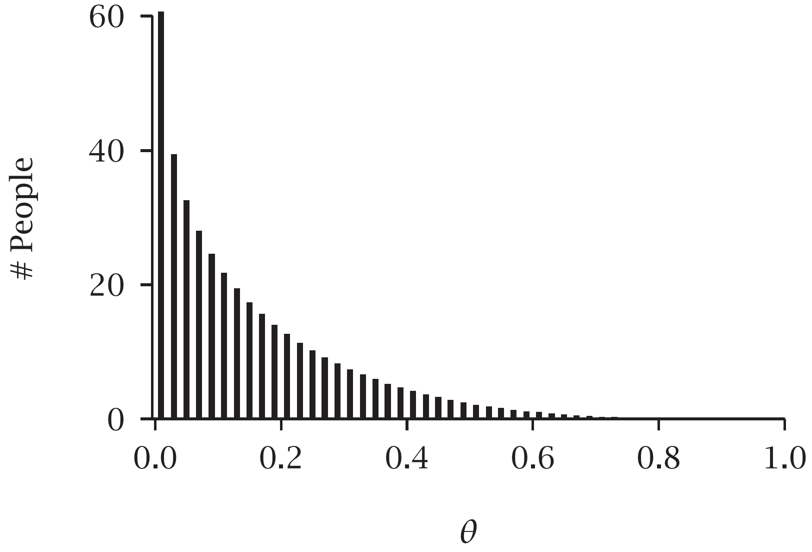
$E(\Theta) = 0.250 \longrightarrow$ expect $629 \times (1 - 0.250) = 472$ customers to renew at the end of Year 2.

Year 3

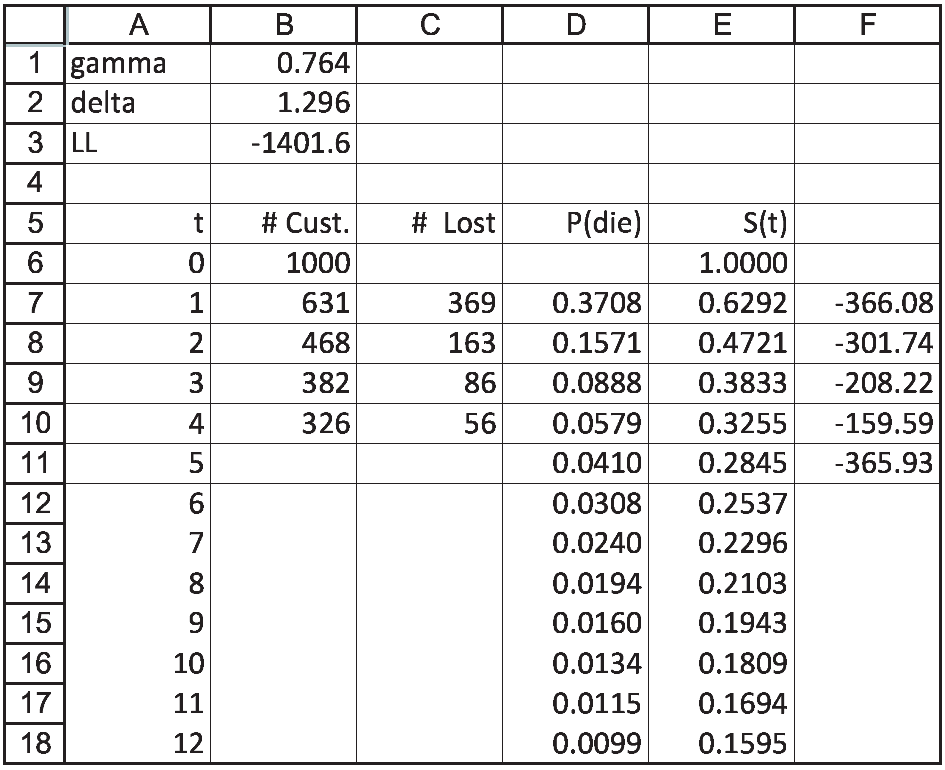


$E(\Theta) = 0.188 \longrightarrow$ expect $472 \times (1 - 0.188) = 383$ customers to renew at the end of Year 3.

Year 4

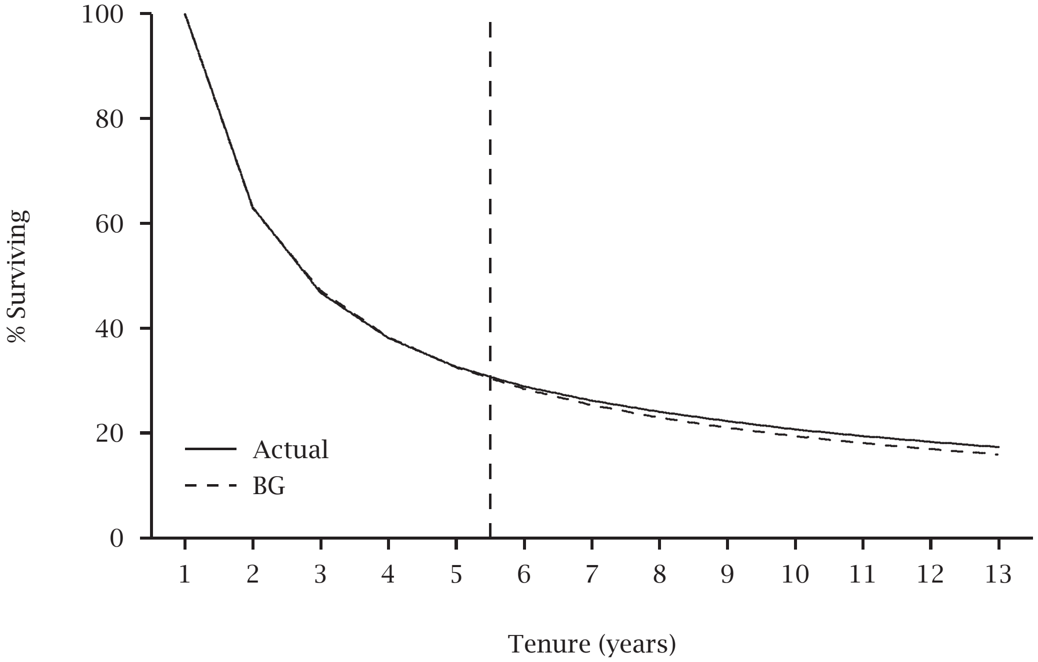


$E(\Theta) = 0.151 \longrightarrow$ expect $383 \times (1 - 0.151) = 325$ customers to renew at the end of Year 4.



	A	B	C	D	E	F
1	gamma	0.764				
2	delta	1.296				
3	LL	-1401.6				
4						
5	t	# Cust.	# Lost	P(die)	S(t)	
6	0	1000			1.0000	
7	1	631	369	0.3708	0.6292	-366.08
8	2	468	163	0.1571	0.4721	-301.74
9	3	382	86	0.0888	0.3833	-208.22
10	4	326	56	0.0579	0.3255	-159.59
11	5			0.0410	0.2845	-365.93
12	6			0.0308	0.2537	
13	7			0.0240	0.2296	
14	8			0.0194	0.2103	
15	9			0.0160	0.1943	
16	10			0.0134	0.1809	
17	11			0.0115	0.1694	
18	12			0.0099	0.1595	

Survival Curve Projection



Implied Retention Rates

- Recall that

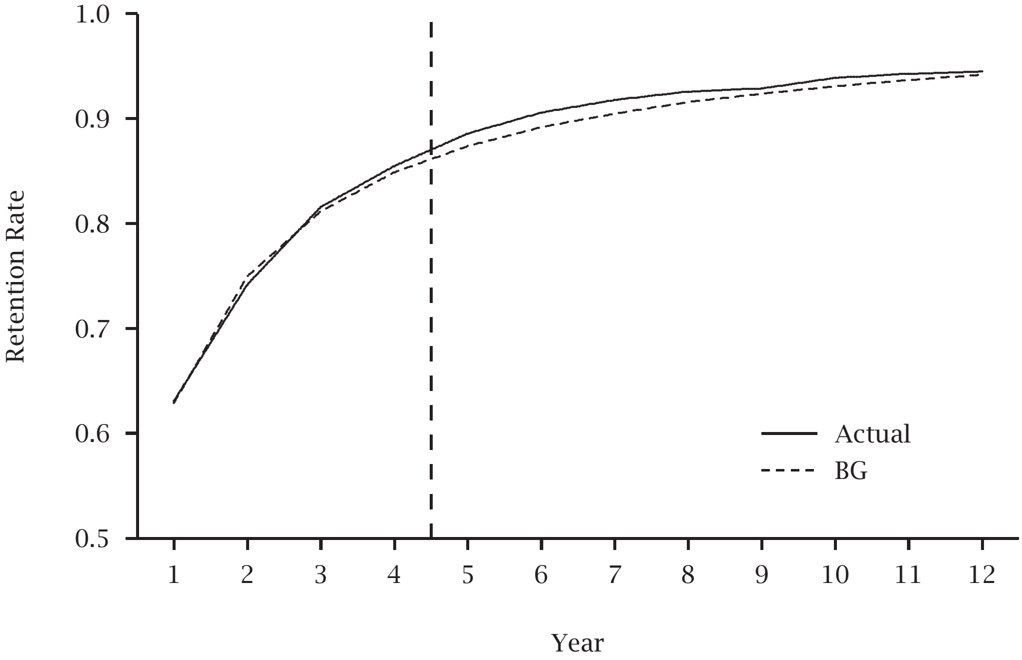
$$r(t) = \frac{S(t)}{S(t-1)}, \; t = 1, 2, 3, \ldots$$

- Given the expression BG survivor function,

$$r(t \mid \gamma, \delta) = \frac{\delta + t - 1}{\gamma + \delta + t - 1}.$$

- An increasing function of time, even though the individual-level retention probability is constant.
- A sorting effect in a heterogeneous population.

Projecting Retention Rates



Concepts and Tools Introduced

- The concept of duration-time data, with a specific focus on single-event discrete-time data.
- The idea of building a “probability model” to characterize the observed behavior of interest.
- The method of maximum-likelihood as a means of estimating model parameters.
- The beta-geometric (BG) distribution as a model of contract renewal behavior.
- Retention rate “dynamics.”

Further Reading

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

Fader, Peter S. and Bruce G.S. Hardie (2014), "A Spreadsheet-Literate Non-Statistician's Guide to the Beta-Geometric Model." (`http://brucehardie.com/notes/032/`)

Fader, Peter S. and Bruce G.S. Hardie (2007), "How Not to Project Customer Retention." (`http://brucehardie.com/notes/016/`)

Lee, Ka Lok, Peter S. Fader, and Bruce G.S. Hardie (2007), "How to Project Patient Persistency," *FORESIGHT*, Issue 8, Fall, 31–35.

Buchanan, Bruce and Donald G. Morrison (1988), "A Stochastic Model of List Falloff with Implications for Repeat Mailings," *Journal of Direct Marketing*, **2** (Summer), 7–15.

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

From Discrete to Continuous Time

- We have considered a setting where the discrete contract period is annual.
- In some cases, there is a quarterly contract period, others monthly.
- In a number of cases, the contract is effectively "renewed" on a daily basis $\Longrightarrow$ "continuous" time.

From Discrete to Continuous Time

As the number of divisions of a given time period ⟶ ∞,

$$
\begin{array}{lcl}
\text{geometric} & \longrightarrow & \text{exponential} \\
\text{BG} & \longrightarrow & \text{gamma mixture of exponentials} \\
 & = & \text{Pareto Type II}
\end{array}
$$

$$S(t \mid r, \alpha) = \left(\frac{\alpha}{\alpha + t}\right)^r$$

From Discrete to Continuous Time

- A continuous-time model can be fitted to discrete-time by treating it as “interval-censored” data:

$$P(T = t) = S(t-1) - S(t)\,.$$

- The fit and associated forecasts of the Pareto Type II are the same as those of the BG.
- Tend to favor a discrete-time model given ease of story telling.
- We use a continuous-time model when we wish to incorporate the effects of covariates.

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.

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

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.

Schweidel, David A., Peter S. Fader, and Eric T. Bradlow (2008), "Understanding Service Retention Within and Across Cohorts Using Limited Information," *Journal of Marketing*, **72** (January), 82–94.

Introduction to Probability Models

The Logic of Probability Models

- The actual data-generating process that lies behind any given data on buyer behavior embodies a huge number of factors.
- Even if the actual process were completely deterministic, it would be impossible to measure all the variables that determine an individual's buying behavior in any setting.

$\Longrightarrow$ Any account of buyer behavior must be expressed in probabilistic/random/stochastic terms so as to account for our ignorance regarding (and/or lack of data on) all the determinants.

The Logic of Probability Models

- Rather than try to tease out the effects of various marketing, personal, and situational variables, we embrace the notion of randomness and view the behavior of interest as the outcome of some probabilistic process.
- We propose a model of individual-level behavior that is "summed" across individuals (taking individual differences into account) to obtain a model of aggregate behavior.

“Winwood Reade is good upon the subject,” said Holmes. “He remarks that, while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to.”

Sir Arthur Conan Doyle, *The Sign of the Four*, 1890.

Applications of Probability Models

- Summarize and interpret patterns of market-level behavior
- Predict behavior in future periods, be it in the aggregate or at a more granular level (e.g., conditional on past behavior)
- Make inferences about behavior given summary measures
- Profile behavioral propensities of individuals
- Generate 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.

A probability model is an "as-if" story about the data-generating process.

Further Reading

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

Outline

Problem 1: Projecting Customer Retention Rates
(Modelling Discrete-Time Duration Data)

Problem 2: Estimating Concentration in Champagne Purchasing
(Modelling Count Data)

Problem 3: Test/Roll Decisions in Segmentation-based Direct Marketing
(Modelling "Choice" Data)

Problem 2:

Estimating Concentration in Champagne Purchasing

(Modelling Count Data)

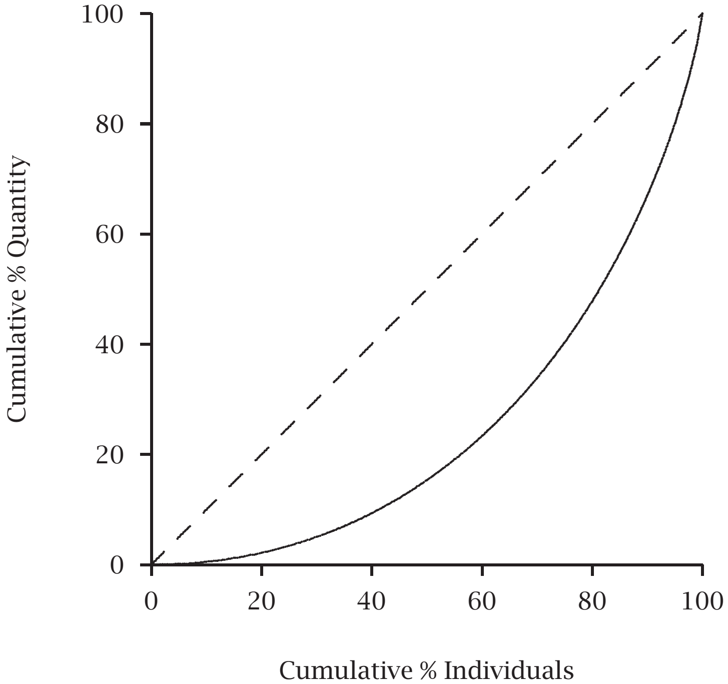
Concentration 101

- Concentration in customer purchasing means that a small proportion of customers make a large proportion of the total purchases of the product (e.g., “80/20”).

  higher concentration ⇔ greater inequality

- The *Lorenz curve* is used to illustrate the degree of inequality in the distribution of a quantity of interest (e.g., purchasing, income, wealth).
- The greater the curvature of the Lorenz Curve, the greater the concentration/inequality.

Concentration 101



Concentration 101

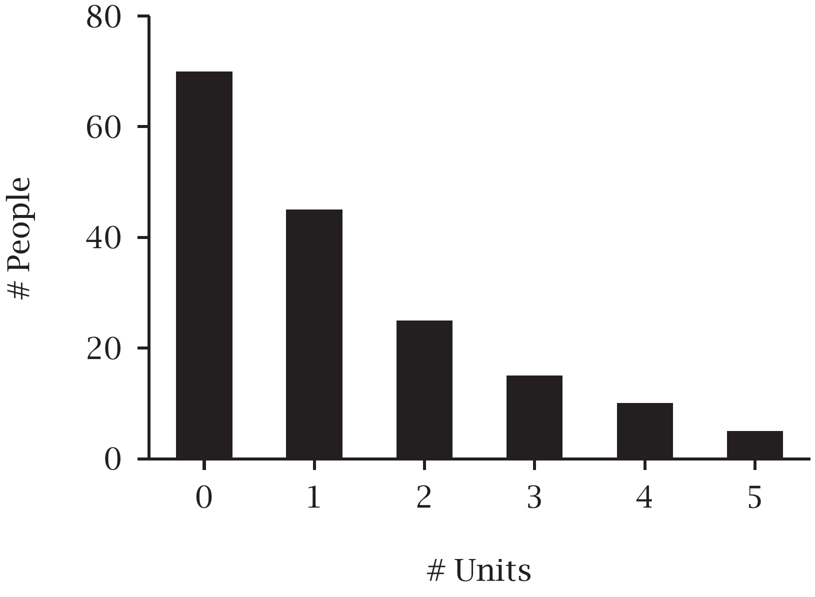
- Every point on the Lorenz curve represents the y% of the quantity of interest accounted for by the bottom x% of all relevant individuals:

$$y = L(x)$$

- 80/20 represents a specific point on the Lorenz curve: $20 = L(80)$.
- The *Gini coefficient* is the ratio of the area between the 45° line ("line of perfect equality") and the Lorenz curve to the area under the line of perfect equality.

Concentration 101

Hypothetical distribution of purchases (n = 170 people):

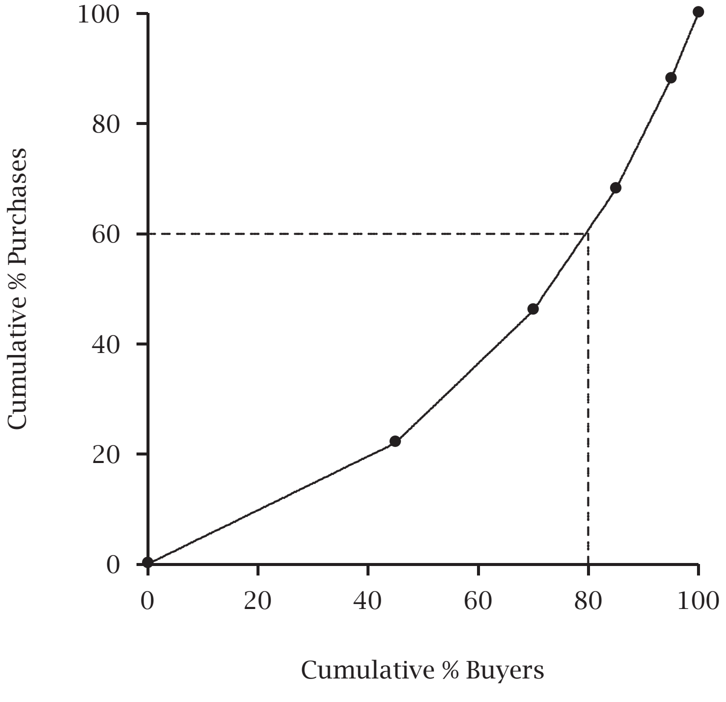


Concentration 101

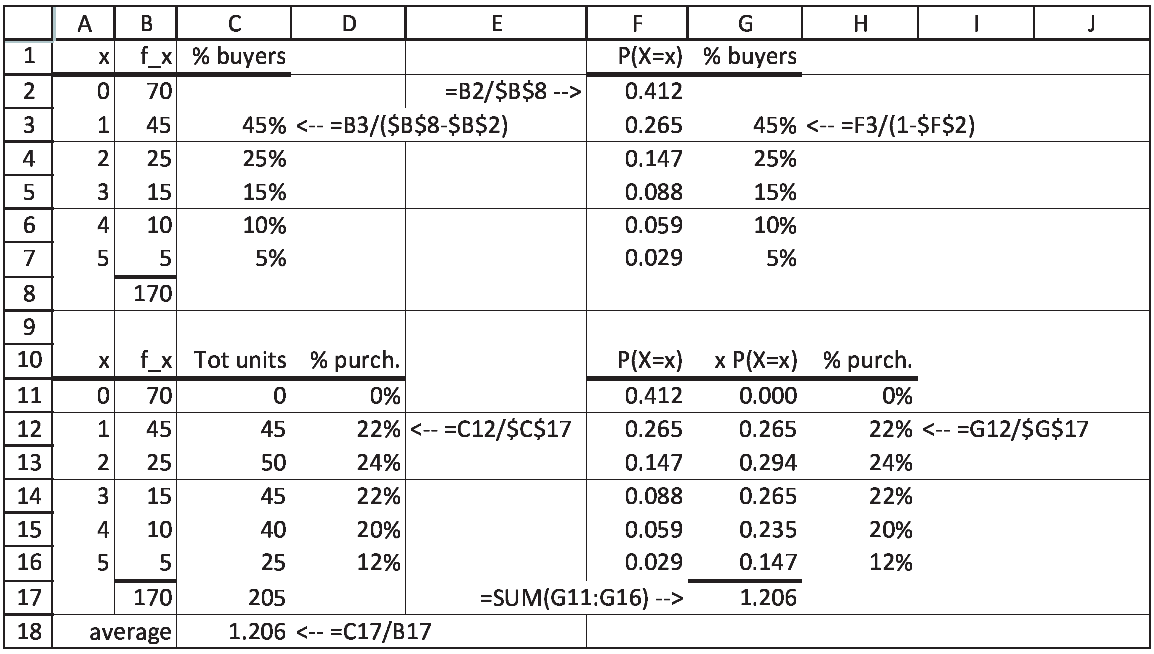
# Units	# People	Total Units	% Buyers	% Purchases	Cum. % Buyers	Cum. % Purchases
0	70	0	0%	0%	0%	0%
1	45	45	45%	22%	45%	22%
2	25	50	25%	24%	70%	46%
3	15	45	15%	22%	85%	68%
4	10	40	10%	20%	95%	88%
5	5	25	5%	12%	100%	100%

Total units:	205
Total buyers:	100

Lorenz Curve



Calculations Revisited



	A	B	C	D	E	F	G	H	I	J
1	x	f_x	% buyers			P(X=x)	% buyers			
2	0	70			=B2/$B$8 -->	0.412				
3	1	45	45%	<-- =B3/($B$8-$B$2)		0.265	45%	<-- =F3/(1-$F$2)		
4	2	25	25%			0.147	25%			
5	3	15	15%			0.088	15%			
6	4	10	10%			0.059	10%			
7	5	5	5%			0.029	5%			
8		170								
9										
10	x	f_x	Tot units	% purch.		P(X=x)	x P(X=x)	% purch.		
11	0	70	0	0%		0.412	0.000	0%		
12	1	45	45	22%	<-- =C12/$C$17	0.265	0.265	22%	<-- =G12/$G$17	
13	2	25	50	24%		0.147	0.294	24%		
14	3	15	45	22%		0.088	0.265	22%		
15	4	10	40	20%		0.059	0.235	20%		
16	5	5	25	12%		0.029	0.147	12%		
17		170	205			=SUM(G11:G16) -->	1.206			
18		average	1.206	<-- =C17/B17						

Calculations Revisited

Consider those buyers that purchased x times ($x \geq 1$) :

- What proportion of total buyers are they?

$$\frac{P(X = x)}{1 - P(X = 0)}$$

- What proportion of total purchasing do they account for?

$$\frac{x\,P(X = x)}{E(X)}$$

Problem

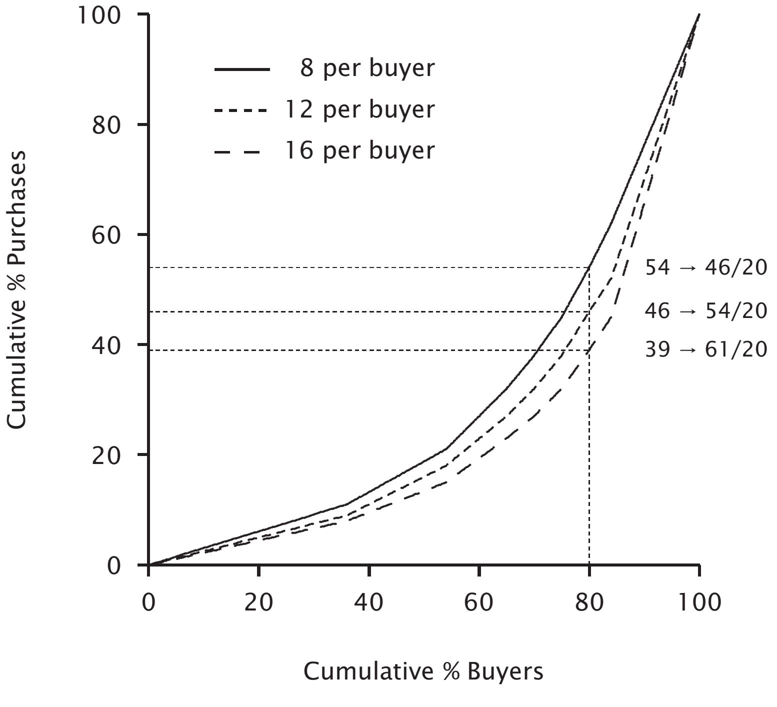
Consider the following data on the number of bottles of champagne purchased in a year by a sample of 568 French households:

# Bottles	0	1	2	3	4	5	6	7	8+
Frequency	400	60	30	20	8	8	9	6	27

Data source: Gourieroux and Visser (*Journal of Econometrics*, 1997)

- What percentage of champagne purchasing is accounted for by the top 20% of buyers?
- What percentage of buyers account for 50% of champagne purchasing?

Associated Lorenz Curves



Modelling Objective

We need to infer the full distribution from the right-censored data … from which we can create the Lorenz curve.

⟶ Develop a model that enables us to estimate the number of people purchasing $0, 1, 2, \ldots, 7, 8, 9, \ldots$ bottles of champagne in a year.

Model Development

- Let the random variable X denote the number of bottles purchased in a year.
- At the individual-level, X is assumed to be Poisson distributed with (purchase) rate parameter λ:

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

The mean and variance of the Poisson are

$$E(X \mid \lambda) = \lambda \text{ and } \operatorname{var}(X \mid \lambda) = \lambda\,.$$

Accounting for Heterogeneity

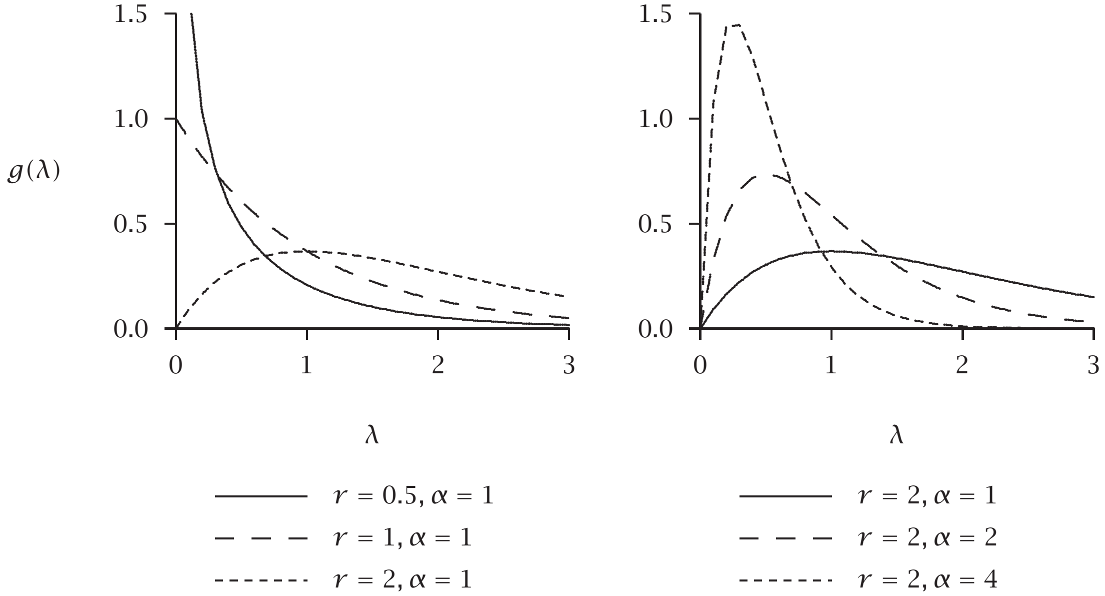
- Assume purchase rates are distributed across the population according to a gamma distribution:

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

  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



Model Development

- For a randomly chosen individual,

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

- This *gamma mixture of Poissons* is called the Negative Binomial Distribution (NBD).
- The mean and variance of the NBD are

$$E(X \mid r, \alpha) = \frac{r}{\alpha} \text{ and } \operatorname{var}(X \mid r, \alpha) = \frac{r}{\alpha} + \frac{r}{\alpha^2}\,.$$

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:

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

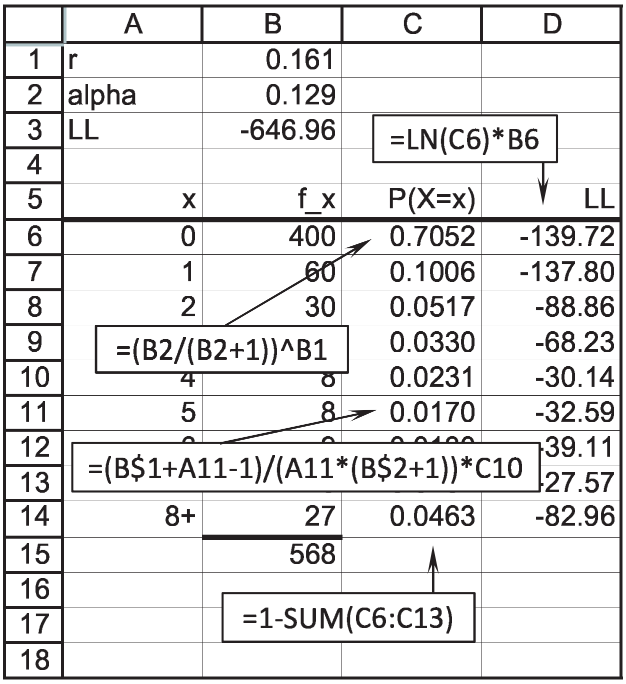
Estimating Model Parameters

The log-likelihood function is defined as:

$$\begin{aligned} LL(r,\alpha|\text{data}) = 400 \times \ln[P(X=0)]+ \\ 60 \times \ln[P(X=1)]+ \\ \ldots \qquad\qquad + \\ 6 \times \ln[P(X=7)]+ \\ 27 \times \ln[P(X \geq 8)] \end{aligned}$$

The maximum value of the log-likelihood function is $LL = -646.96$, which occurs at $\hat{r} = 0.161$ and $\hat{\alpha} = 0.129$.

Estimating Model Parameters



	A	B	C	D
1	r	0.161		
2	alpha	0.129		
3	LL	-646.96		
4				
5	x	f_x	P(X=x)	LL
6	0	400	0.7052	-139.72
7	1	60	0.1006	-137.80
8	2	30	0.0517	-88.86
9			0.0330	-68.23
10	4	8	0.0231	-30.14
11	5	8	0.0170	-32.59
12				39.11
13				27.57
14	8+	27	0.0463	-82.96
15		568		
16				
17				
18				

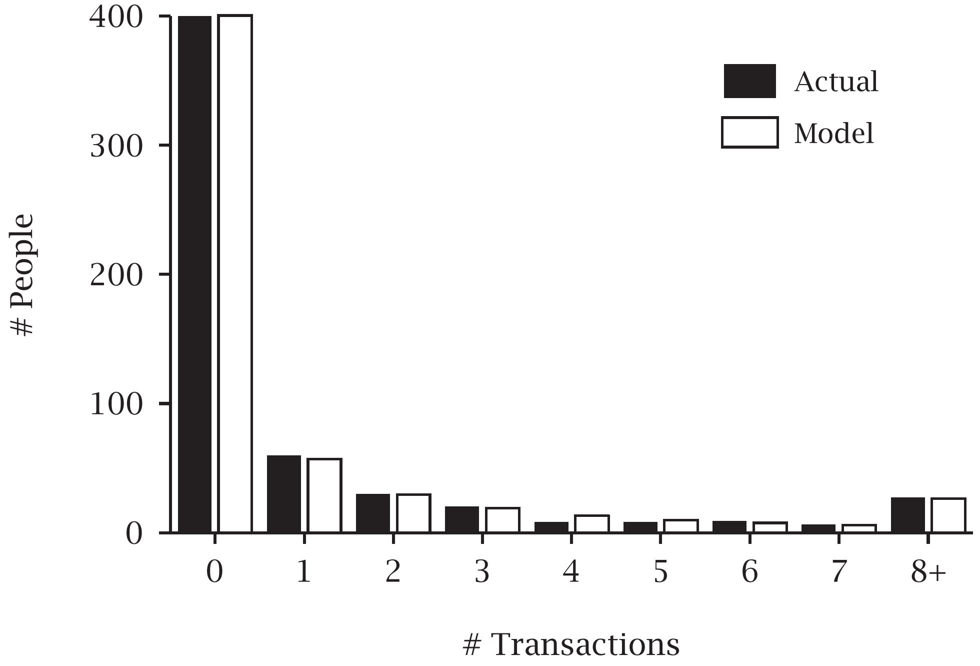
=LN(C6)*B6

=(B2/(B2+1))^B1

=(B$1+A11-1)/(A11*(B$2+1))*C10

=1-SUM(C6:C13)

Model Fit



Chi-square Goodness-of-Fit Statistic

Does the distribution $F(x|\boldsymbol{\theta})$, with s model parameters denoted by $\boldsymbol{\theta}$, provide a good fit to the sample data?

- Divide the sample into k mutually exclusive and collectively exhaustive groups.
- Let f_i $(i = 1, \ldots, k)$ be the number of sample observations in group i, p_i the probability of belonging to group i, and n the sample size.

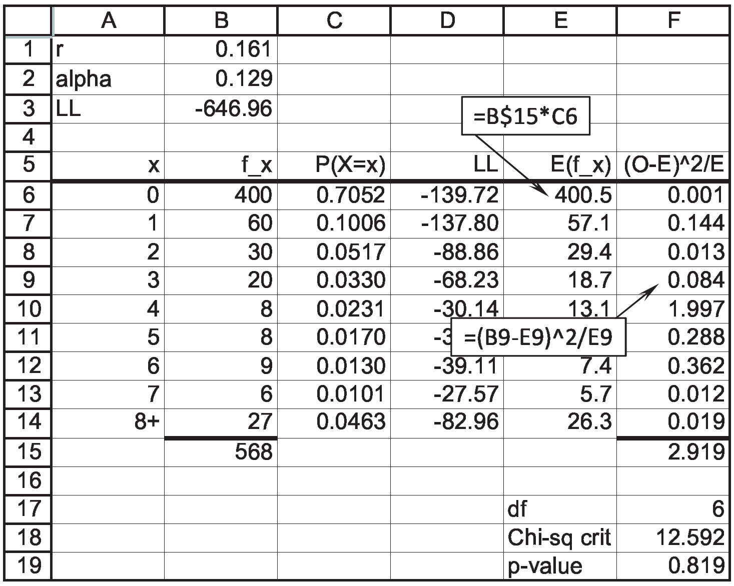
Chi-square Goodness-of-Fit Statistic

- Compute the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$$

- Reject the null hypothesis that the observed data come from $F(x|\boldsymbol{\theta})$ if the test statistic is greater than the critical value (i.e., $\chi^2 > \chi^2_{.05,k-s-1}$).
- The critical value can be computed in Excel 2010 (and above) using the `CHISQ.INV.RT` function (and the corresponding p-value using the `CHISQ.DIST.RT` function).

Model Fit

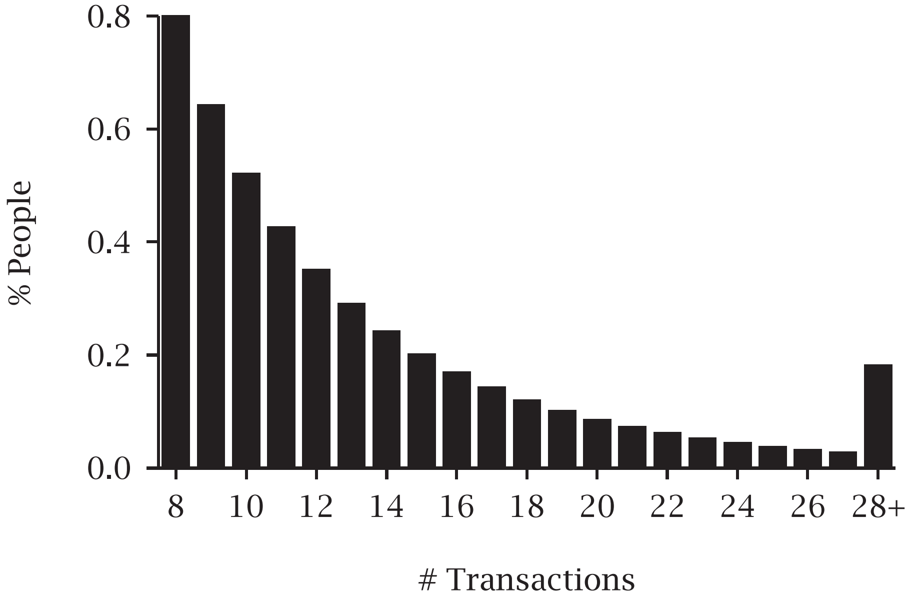


	A	B	C	D	E	F
1	r	0.161				
2	alpha	0.129				
3	LL	-646.96				
4						
5	x	f_x	P(X=x)	LL	E(f_x)	(O-E)^2/E
6	0	400	0.7052	-139.72	400.5	0.001
7	1	60	0.1006	-137.80	57.1	0.144
8	2	30	0.0517	-88.86	29.4	0.013
9	3	20	0.0330	-68.23	18.7	0.084
10	4	8	0.0231	-30.14	13.1	1.997
11	5	8	0.0170	-3	[illegible]	0.288
12	6	9	0.0130	-39.11	7.4	0.362
13	7	6	0.0101	-27.57	5.7	0.012
14	8+	27	0.0463	-82.96	26.3	0.019
15		568				2.919
16						
17					df	6
18					Chi-sq crit	12.592
19					p-value	0.819

=B$15*C6

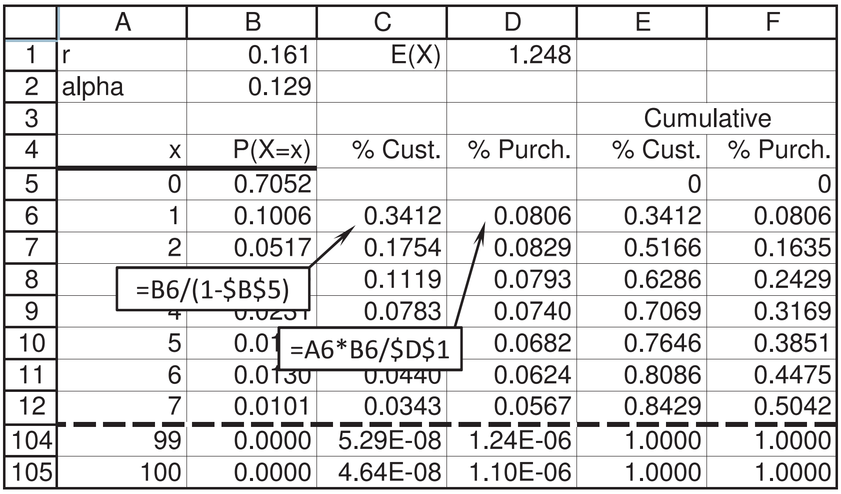
=(B9-E9)^2/E9

Decomposing the 8+ Cell



The mean for this group of people is 13.36 purchases per buyer ... but with great variability.

Creating the Lorenz Curve

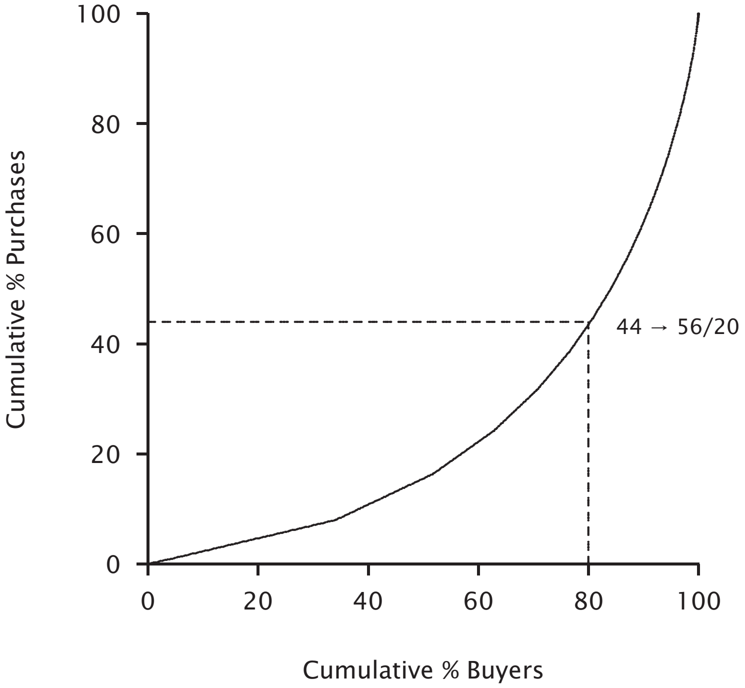


	A	B	C	D	E	F
1	r	0.161	E(X)	1.248		
2	alpha	0.129				
3					Cumulative	
4	x	P(X=x)	% Cust.	% Purch.	% Cust.	% Purch.
5	0	0.7052			0	0
6	1	0.1006	0.3412	0.0806	0.3412	0.0806
7	2	0.0517	0.1754	0.0829	0.5166	0.1635
8	[illegible]	[illegible]	0.1119	0.0793	0.6286	0.2429
9	4	0.0231	0.0783	0.0740	0.7069	0.3169
10	5	[illegible]	[illegible]	0.0682	0.7646	0.3851
11	6	0.0130	0.0440	0.0624	0.8086	0.4475
12	7	0.0101	0.0343	0.0567	0.8429	0.5042
104	99	0.0000	5.29E-08	1.24E-06	1.0000	1.0000
105	100	0.0000	4.64E-08	1.10E-06	1.0000	1.0000

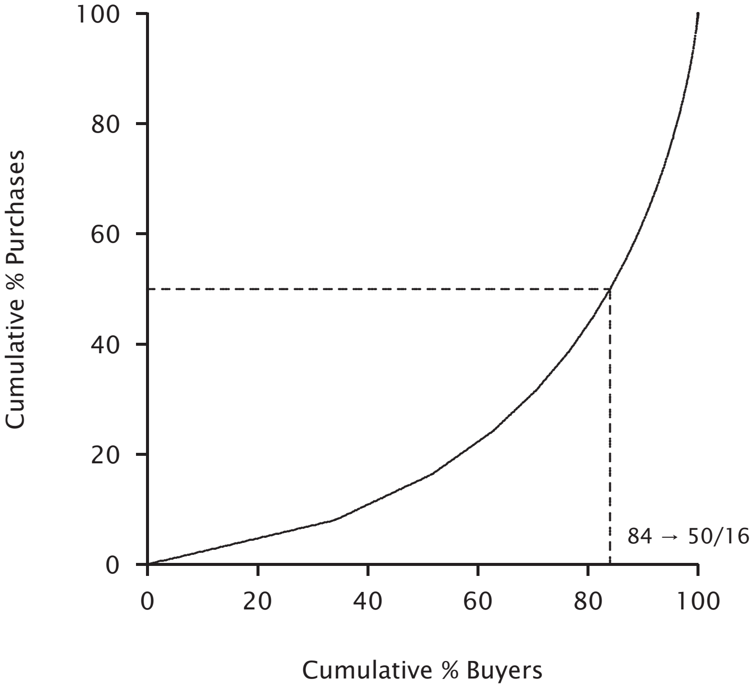
=B6/(1-$B$5)

=A6*B6/$D$1

Lorenz Curve for Champagne Purchasing



Lorenz Curve for Champagne Purchasing



Concepts and Tools Introduced

- The concept of count data.
- The NBD as a model for count data.
- The notion of concentration, and the Lorenz curve as a means of illustrating the level of “inequality” in the quantity of interest.
- Using a probability model infer a full distribution given right-censored data.

Further Reading

Ehrenberg, A. S. C. (1959), “The Pattern of Consumer Purchases,” *Applied Statistics*, **8** (March), 26–41.

Greenwood, Major and G. Udny Yule (1920), “An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents,” *Journal of the Royal Statistical Society*, **83** (March), 255–279.

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

Further Reading

Ehrenberg, A. S. C. (1988), *Repeat-Buying*, 2nd edn, London: Charles Griffin & Company, Ltd. (Available at `http://www.empgens.com/ArticlesHome/Volume5/RepeatBuying.html`)

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.

Schmittlein, David C., Lee G. Cooper, and Donald G. Morrison (1993), "Truth in Concentration in the Land of (80/20) Laws," *Marketing Science*, **12** (Spring), 167–183.

Problem 3:

Test/Roll Decisions in Segmentation-based Direct Marketing

(Modelling "Choice" Data)

The “Segmentation” Approach

i) Divide the customer list into a set of (homogeneous) segments.

ii) Test customer response by mailing to a random sample of each segment.

iii) Roll out to those segments that were profitable in the test:

$$\#\,\text{orders} \times \text{unit margin} > \#\,\text{mailings} \times \text{cost of each mailing}$$

$$\Leftrightarrow \quad \underbrace{\frac{\#\,\text{orders}}{\#\,\text{mailings}}}_{\text{response rate (RR)}} > \frac{\text{cost of each mailing}}{\text{unit margin}}$$

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.

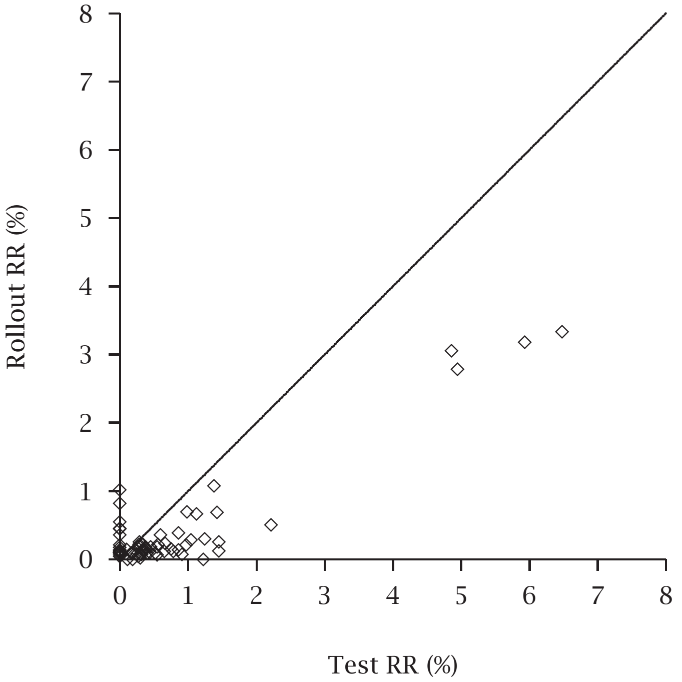
Standard approach:

- Rollout to all segments with

$$\text{Test RR} > \frac{3{,}343/10{,}000}{161.50} = 0.00207$$

- 51 segments pass this hurdle

Test vs. Actual Response Rate



Modelling Objective

Develop a model to help the manager estimate each segment's "true" response rate given the (limited) test data.

Model Development

Notation

$$N_s = \text{size of segment } s \ (s = 1, \ldots, S)$$

$$m_s = \text{\# members of segment } s \text{ tested}$$

$$X_s = \text{\# responses to test in segment } s$$

Assumptions

i) All members of segment s have the same (unknown) response probability $\theta_s \implies X_s$ is a binomial random variable:

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

Distribution of Response Probabilities

ii) Heterogeneity in θ_s is captured by a beta distribution:

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

It follows that the aggregate distribution of responses to a mailing of size m_s is given by

$$\begin{aligned} P(X_s = x_s \mid m_s\ \alpha, \beta) &= \int_0^1 P(X_s = x_s \mid m_s, \theta_s)\, g(\theta_s \mid \alpha, \beta)\, d\theta_s \\ &= \binom{m_s}{x_s} \frac{B(\alpha + x_s, \beta + m_s - x_s)}{B(\alpha, \beta)}\,. \end{aligned}$$

This is known as the beta-binomial (BB) distribution.

Numerical Evaluation of the Beta Function

- Not all computing environments have a `beta` function.
- Recall

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

- We typically have a function that evaluates $\ln(\Gamma(\cdot))$.
- In Excel we have `gammaln`:

$$\begin{aligned} \Gamma(\alpha) &= \texttt{exp(gammaln(}\alpha\texttt{))} \\ B(\alpha, \beta) &= \texttt{exp(gammaln(}\alpha\texttt{)+gammaln(}\beta\texttt{)} \\ &\qquad\quad \texttt{-gammaln(}\alpha\texttt{+}\beta\texttt{))} \end{aligned}$$

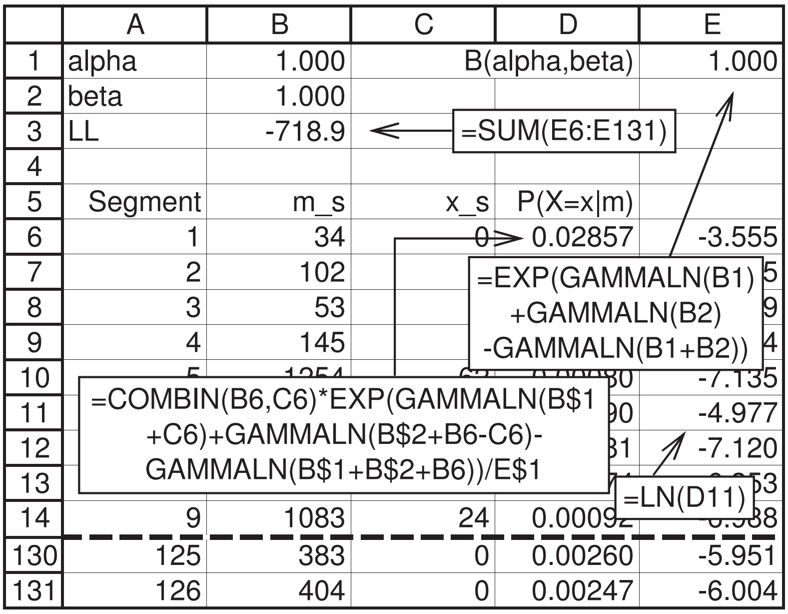
Estimating Model Parameters

The log-likelihood function is defined as:

$$
\begin{aligned}
LL(\alpha, \beta \mid \text{data}) \\
&= \sum_{s=1}^{126} \ln \{P(X_s = x_s \mid m_s, \alpha, \beta)\} \\
&= \sum_{s=1}^{126} \ln \left\{ \binom{m_s}{x_s} \frac{B(\alpha + x_s, \beta + m_s - x_s)}{B(\alpha, \beta)} \right\} \\
&= \sum_{s=1}^{126} \ln \left\{ \frac{m_s!}{(m_s - x_s)!\, x_s!} \frac{\Gamma(\alpha + x_s)\Gamma(\beta + m_s - x_s)}{\Gamma(\alpha + \beta + m_s)} \bigg/ \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \right\}
\end{aligned}
$$

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

Estimating Model Parameters



	A	B	C	D	E
1	alpha	1.000		B(alpha,beta)	1.000
2	beta	1.000			
3	LL	-718.9			
4					
5	Segment	m_s	x_s	P(X=x|m)	
6	1	34	0	0.02857	-3.555
7	2	102	[illegible]	[illegible]	[illegible]
8	3	53	[illegible]	[illegible]	[illegible]
9	4	145	[illegible]	[illegible]	[illegible]
10	5	1254	[illegible]	0.00080	-7.135
11	[illegible]	[illegible]	[illegible]	[illegible]	-4.977
12	[illegible]	[illegible]	[illegible]	[illegible]	-7.120
13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
14	9	1083	24	0.00092	[illegible]
130	125	383	0	0.00260	-5.951
131	126	404	0	0.00247	-6.004

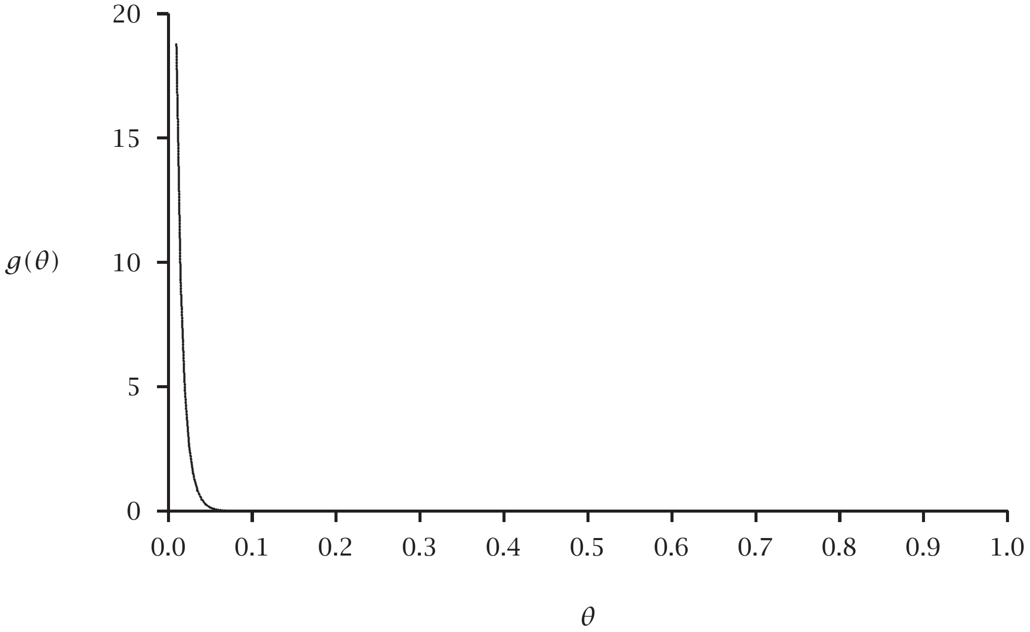
=SUM(E6:E131)

=EXP(GAMMALN(B1)
+GAMMALN(B2)
-GAMMALN(B1+B2))

=COMBIN(B6,C6)*EXP(GAMMALN(B$1
+C6)+GAMMALN(B$2+B6-C6)-
GAMMALN(B$1+B$2+B6))/E$1

=LN(D11)

Estimated Distribution of Θ



$$\hat{\alpha} = 0.439, \hat{\beta} = 95.411, \widehat{E(\Theta)} = 0.0046$$

Applying the Model

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

Intuitively, we would expect

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

Bayes' Theorem

- The *prior distribution* $g(\theta)$ captures the possible values θ can take on, prior to collecting any information about the specific individual.
- 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:

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

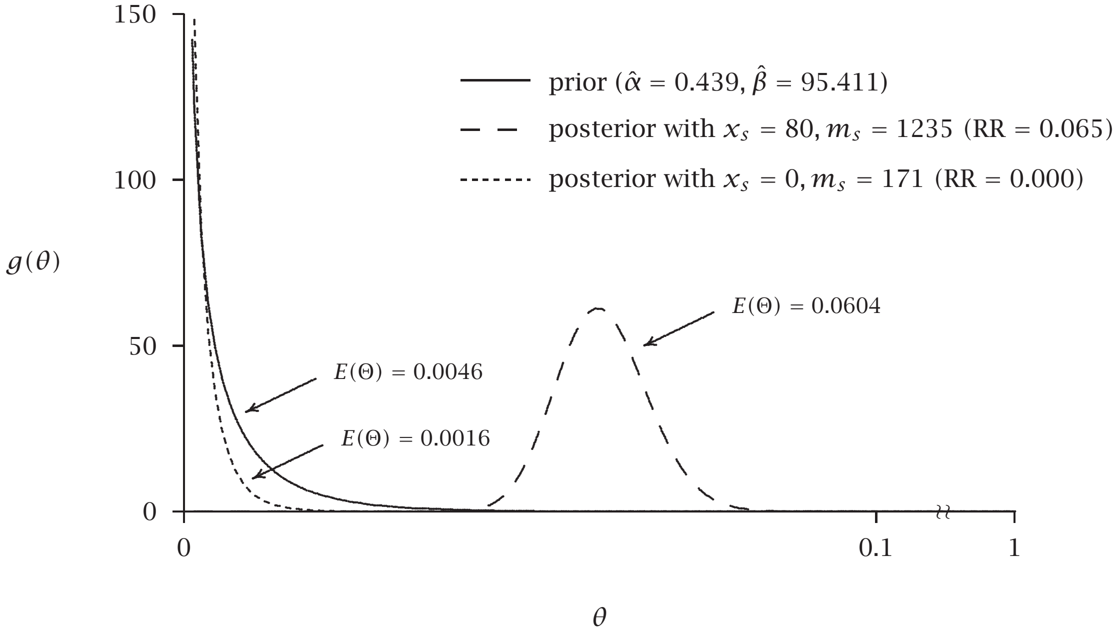
Bayes' Theorem

For the beta-binomial model, we have:

$$\begin{aligned} g(\theta_s|X_s = x_s, m_s) &= \frac{\overbrace{P(X_s = x_s|m_s, \theta_s)}^{\text{binomial}}\,\overbrace{g(\theta_s)}^{\text{beta}}}{\underbrace{\int_0^1 P(X_s = x_s|m_s, \theta_s)\,g(\theta_s)\,d\theta_s}_{\text{beta-binomial}}} \\ &= \frac{1}{B(\alpha + x_s, \beta + m_s - x_s)}\theta_s^{\alpha + x_s - 1}(1 - \theta_s)^{\beta + m_s - x_s - 1} \end{aligned}$$

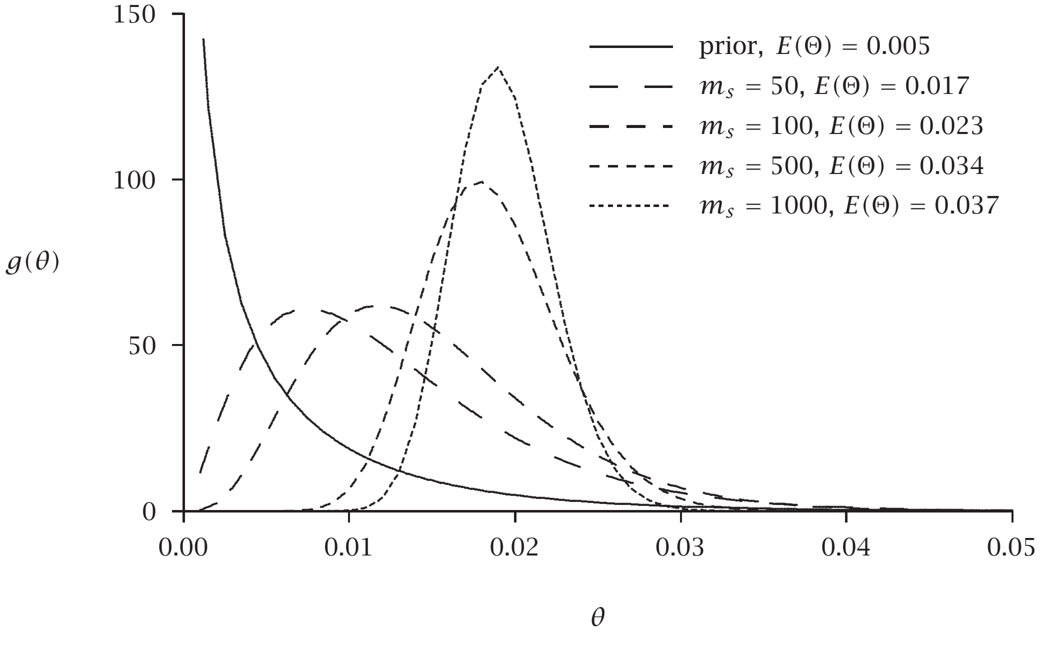
which is a beta distribution with parameters $\alpha + x_s$ and $\beta + m_s - x_s$.

Distribution of Θ



Impact of Sample Size on the Posterior

Four segments, each with a response rate of 0.04:



Applying the Model

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

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

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:

  $$E(\Theta_s|X_s = x_s, m_s) > \frac{3{,}343/10{,}000}{161.5} = 0.00207$$

  (Note that $E(\Theta_s|X_s = 0, m_s) > 0.00207$ for $m_s < 117$.)
- 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 \$6,287; 126,053 fewer contacts, saved for another offering.

Empirical Bayes Methods

- Bayesian analysis methods see us fixing the prior distribution before any data are observed.
- Empirical Bayes methods see us estimating the prior distribution from the data.
- When this prior has a parametric form, we are using parametric empirical Bayes methods.

> “There is no one less Bayesian than an empirical Bayesian.”
>
> Dennis Lindley

Conjugate Priors

- When the posterior distribution comes from the same family as the prior distribution, the prior and posterior are called *conjugate distributions* and the prior is called the *conjugate prior* ($\Longrightarrow$ a closed-form expression for the posterior, which is mathematically convenient.)
- A distribution is a conjugate prior when its kernel is the same as that of the likelihood:

prior	likelihood
$\theta^{\alpha-1}(1-\theta)^{\beta-1}$	$\theta^{x}(1-\theta)^{n-x}$

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 conjugate priors.
- The notion of (parametric) 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.

Maritz, J.S. and T. Lwin (1989), *Empirical Bayes Methods*, 2nd edn, London: Chapman and Hall.

Discussion

Recap

The preceding problems introduce simple models for three behavioral processes:

- Timing ⟶ “when / how long”
- Counting ⟶ “how many”
- “Choice” ⟶ “whether / which”

Phenomenon	Individual-level	Heterogeneity	Model
Timing (discrete) (or counting)	geometric	beta	BG
Timing (continuous)	exponential	gamma	Pareto Type II
Counting	Poisson	gamma	NBD
Choice	binomial	beta	BB

Further Applications: Timing Models

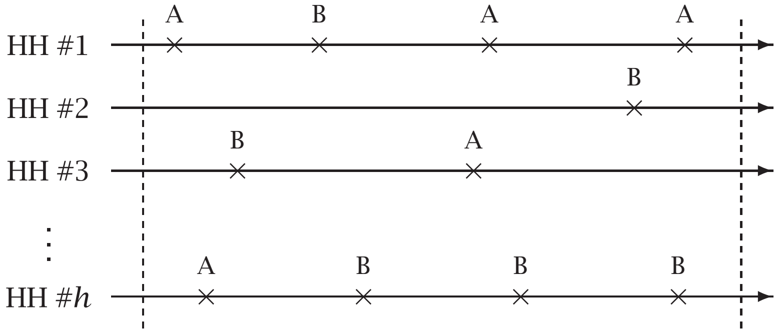
- Response times:
  - Coupon redemptions
  - Survey response
  - Direct mail (response, returns, repeat sales)
- Other durations:
  - Salesforce job tenure
  - Length of website browsing session
- Other positive “continuous” quantities (e.g., spend)

Further Applications: Count Models

- Media exposure (e.g., banner ads, billboards)
- Number of page views during a website browsing session

Further Applications: “Choice” Models

- Brand choice



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

Integrated Models

More complex behavioral phenomena can be captured by combining models from each of these processes:

- Counting + Timing
  - catalog purchases (purchasing|“alive” & “death” process)
  - “engagement” (# 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
	Counting			
Stage 1	Timing			
	Choice			

Further Issues

- Relaxing usual assumptions:
-
  - Non-exponential purchasing (greater regularity) $\longrightarrow$ 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 (HB) methods

The Excel spreadsheets associated with this tutorial, along with electronic copies of the tutorial materials, 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/`