Hierarchical Bayes

Peter Lenk

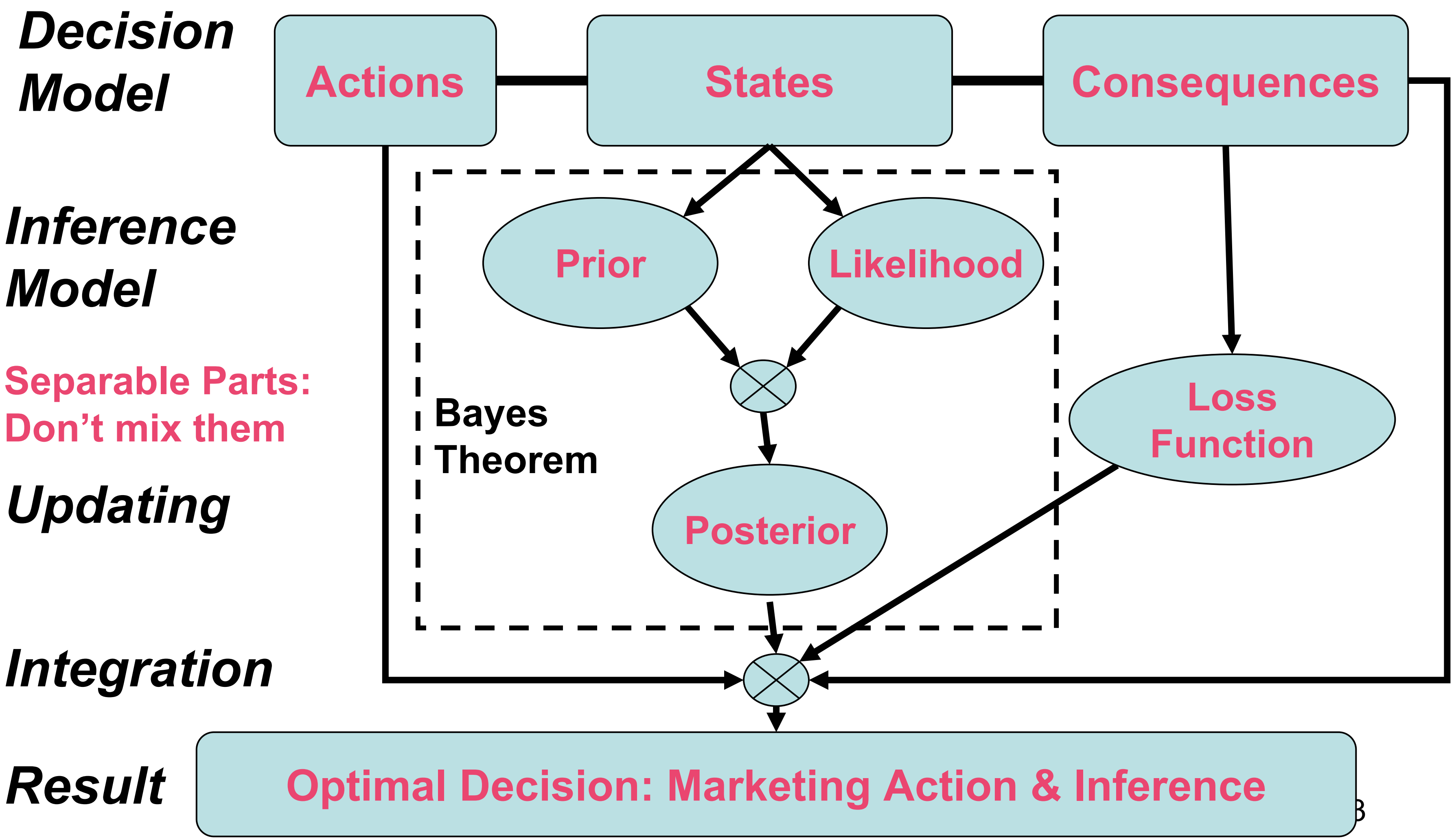
Stephen M Ross School of Business at the University of Michigan

September 2004

Outline

- Bayesian Decision Theory
- Simple Bayes and Shrinkage Estimates
- Hierarchical Bayes
- Numerical Methods
- Batting Averages
- HB Interaction Model

Bayesian Decision Model



Bayes Theorem

- Model for the data given parameters
  - $f(y\,|\theta)$ were θ = unknown parameters
  - E.g. $Y_i = \mu + \varepsilon_i$ and $\theta = (\mu,\sigma)$
  - Likelihood $l(\theta) = f(y_1\,|\theta)\, f(y_2\,|\theta)\ldots\, f(y_n\,|\theta)$
- Prior distribution of parameters $p(\theta)$
- Update prior
  - $p(\theta|\text{Data}\,) = l(\theta)p(\theta)/f(y)$
  - $f(y)$ = marginal distribution of data

Easy Example:

- Estimate mean from a normal distribution.
- $Y_i = \mu + \varepsilon_i$
- Error terms $\{\varepsilon_i\}$ are iid normal
  - Mean is zero
  - Standard deviation of error terms is σ.
  - Assume that σ is known

Conjugate Prior for Mean

- Prior distribution for μ is normal
  - Prior mean is $\mathbf{m_0}$
  - Prior variance is $\mathbf{v_0}^2$
  - Precision is $\mathbf{1/\ v_0}^2$

Posterior Distribution

- Observe n data points
- Posterior distribution is normal
  - Mean is m_n
  - Variance is v_n^2

$$m_n = w\bar{y} + (1-w)m_0$$

$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{v_0^2}} \text{ and } 0 < w < 1$$

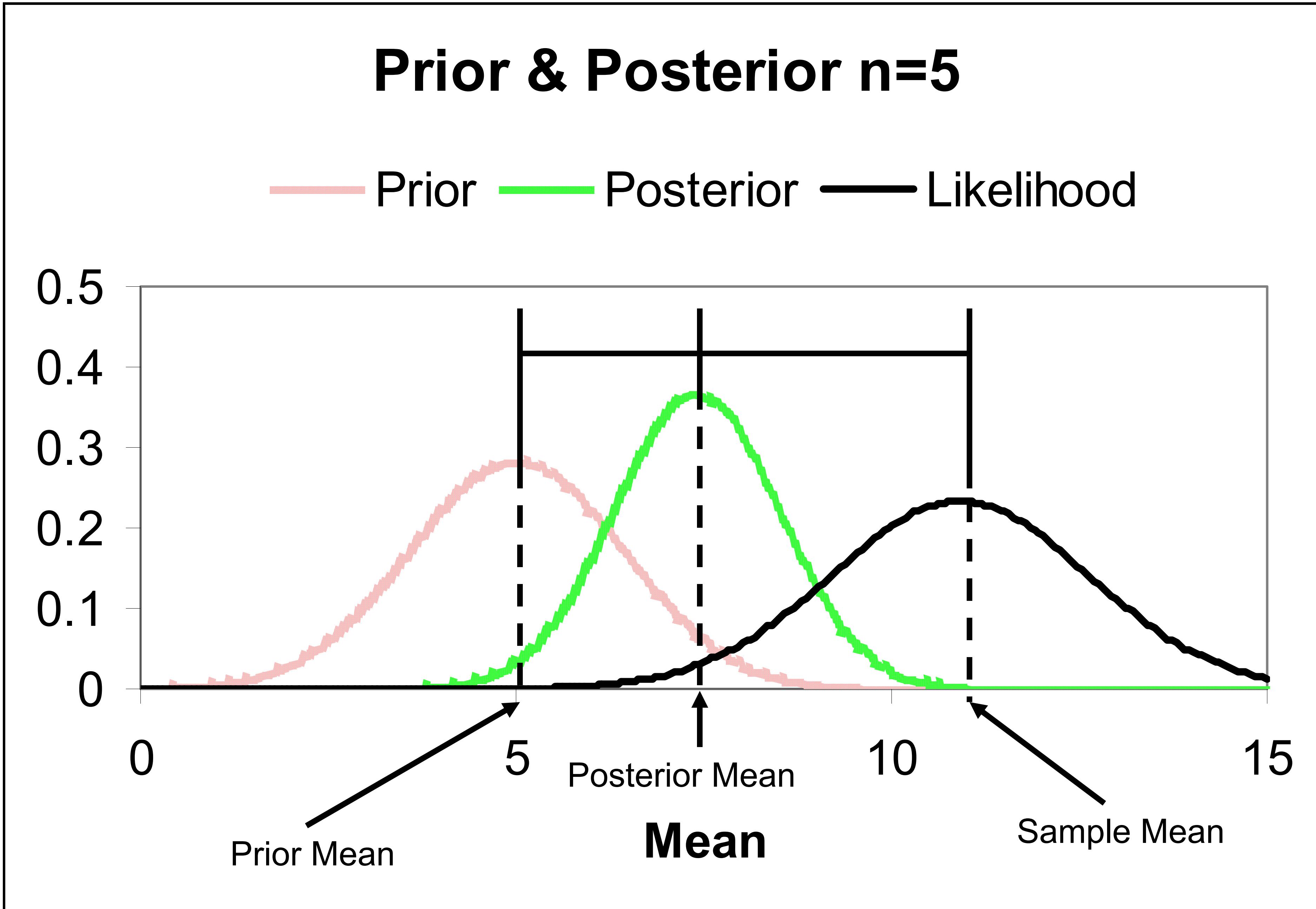
$$v_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{v_0^2}}$$

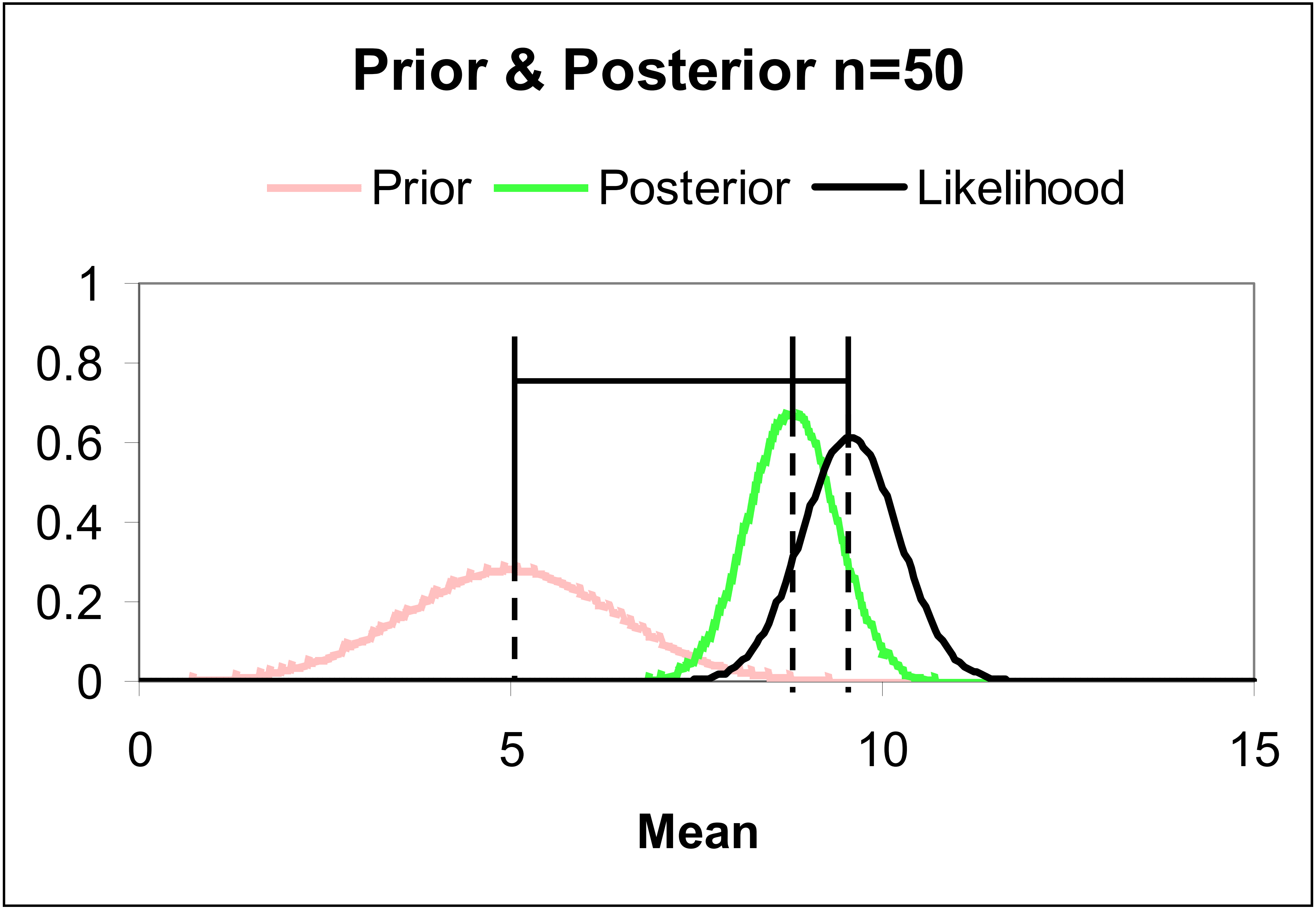
Shrinkage Estimators

- Bayes estimators combines prior guesses with sample estimates
- If the prior precision is larger than sample precision (prior has more information), then put more weight on prior mean.
- If the sample precision is larger the prior precision (sample has more information), then put more weight on sample average

Example

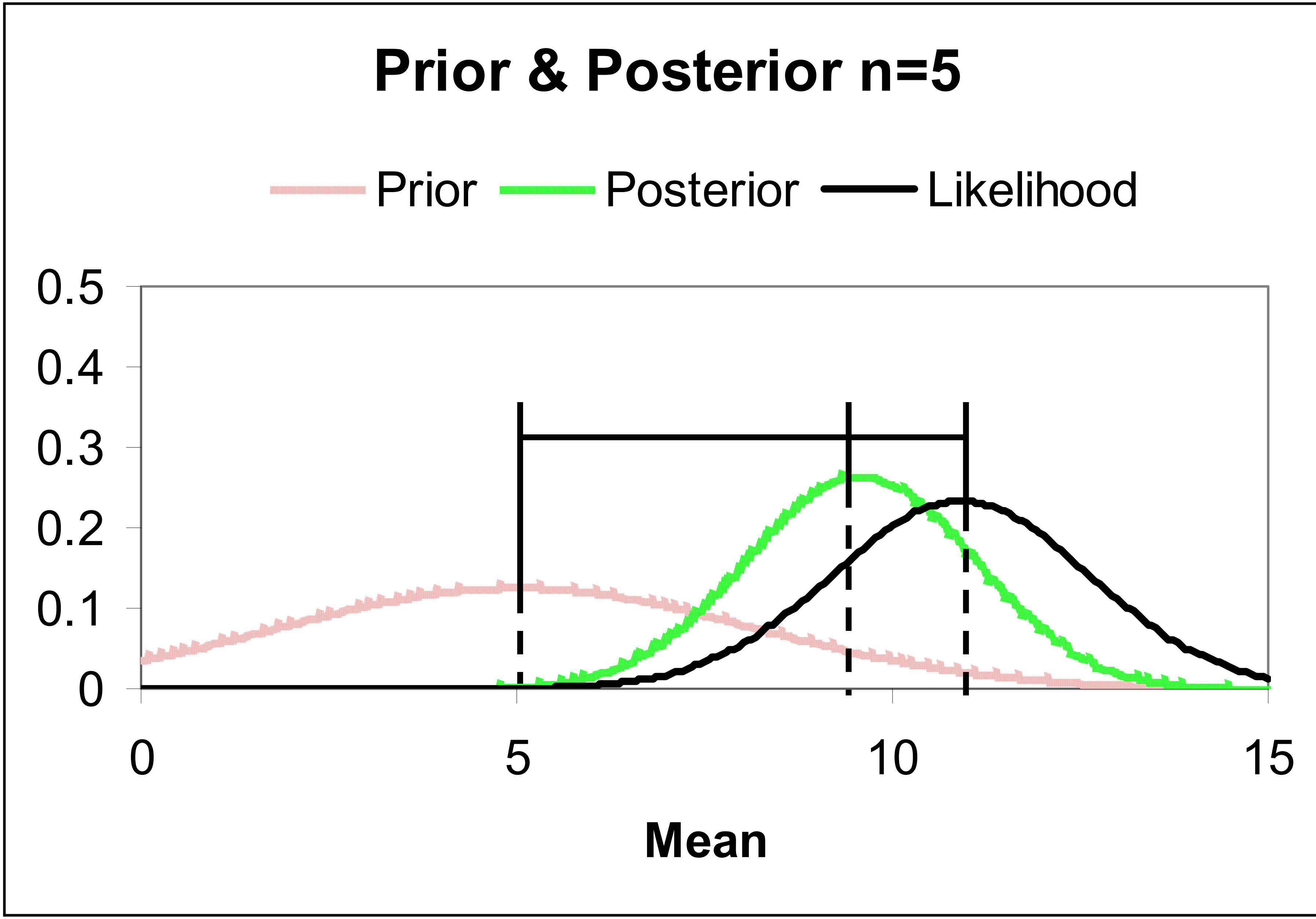
- Y is normal with mean 10 and Variance 16
- Normal prior for the population mean
  - Mean = 5 & Variance = 2
  - Prior is informative and way off
- Data
  - n = 5, Average = 10.9, Variance = 14.7
- Posterior is normal
  - Mean = 7.4 and variance is 1.2

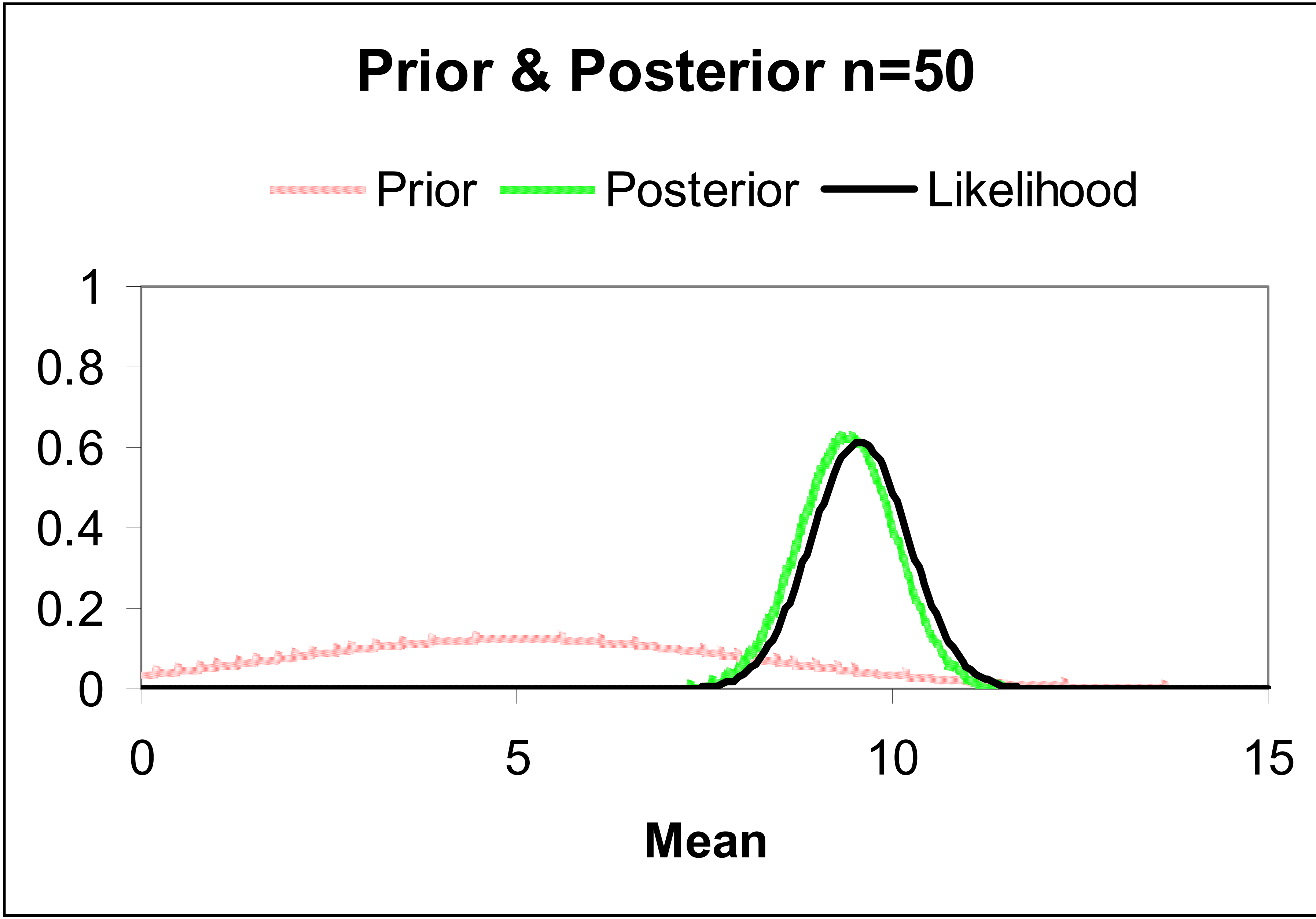

Prior & Posterior n=5
Prior
Posterior
Likelihood
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
Posterior Mean
Mean
Prior Mean
Sample Mean


Prior & Posterior n=50
Prior
Posterior
Likelihood
1
0.8
0.6
0.4
0.2
0
0
5
10
15
Mean

Use Less Informative Prior

- Y is normal with mean 10 and Variance 16
- Normal prior for the population mean
  - Mean = 5 & Variance = 10 instead of 2
  - Prior is "flatter"
- Data
  - n = 5, Average = 10.9, Variance = 14.7
- Posterior is normal
  - Mean = 9.6 and variance is 2.3


Prior & Posterior n=5
Prior
Posterior
Likelihood
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
Mean



Prior & Posterior n=50
Prior
Posterior
Likelihood
1
0.8
0.6
0.4
0.2
0
0
5
10
15
Mean

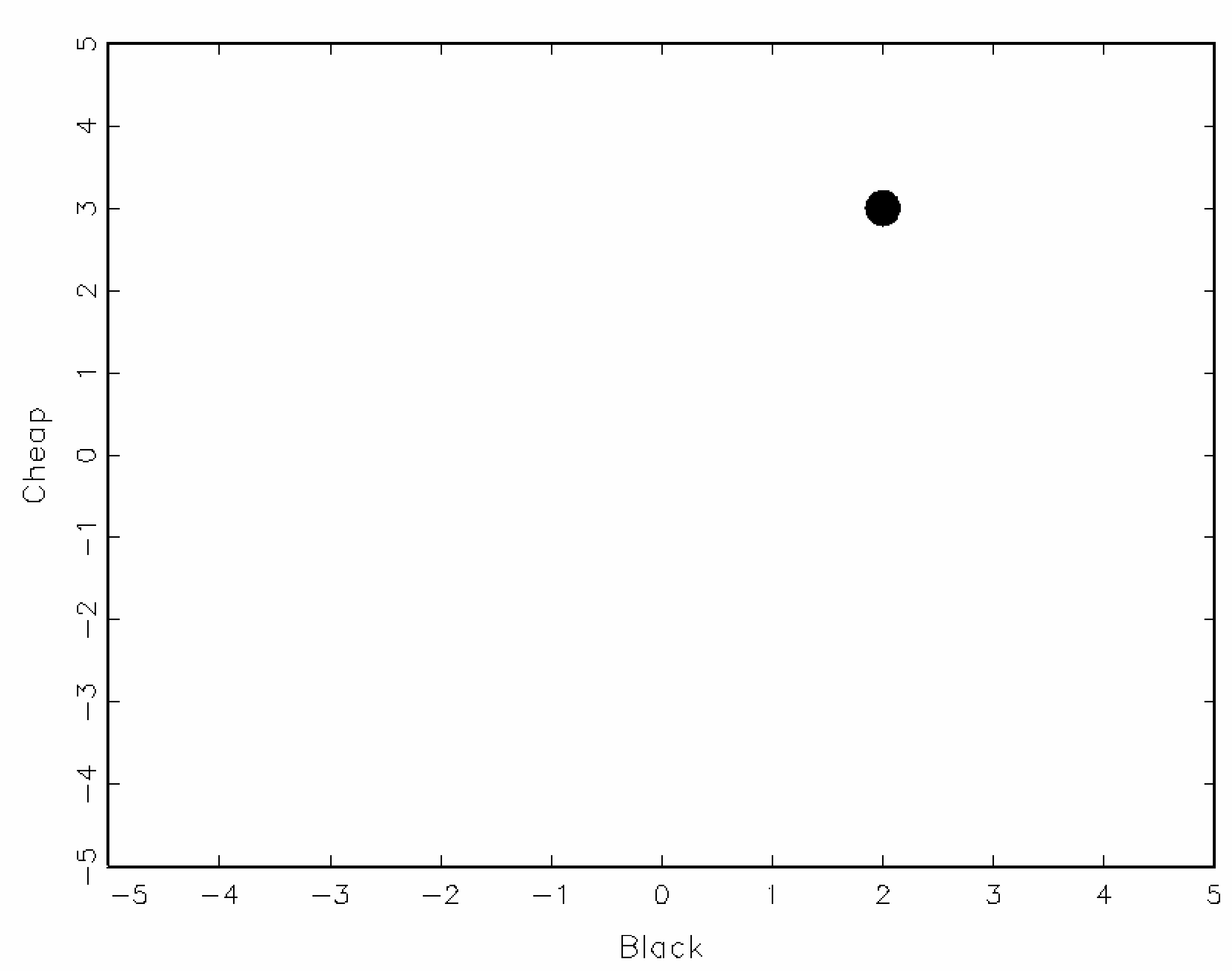
Summary

- Prior has less effect as sample size increases
- Very informative priors give good results with smaller samples if prior information is correct
- If you really don’t know, then use “flatter” or less informative priors

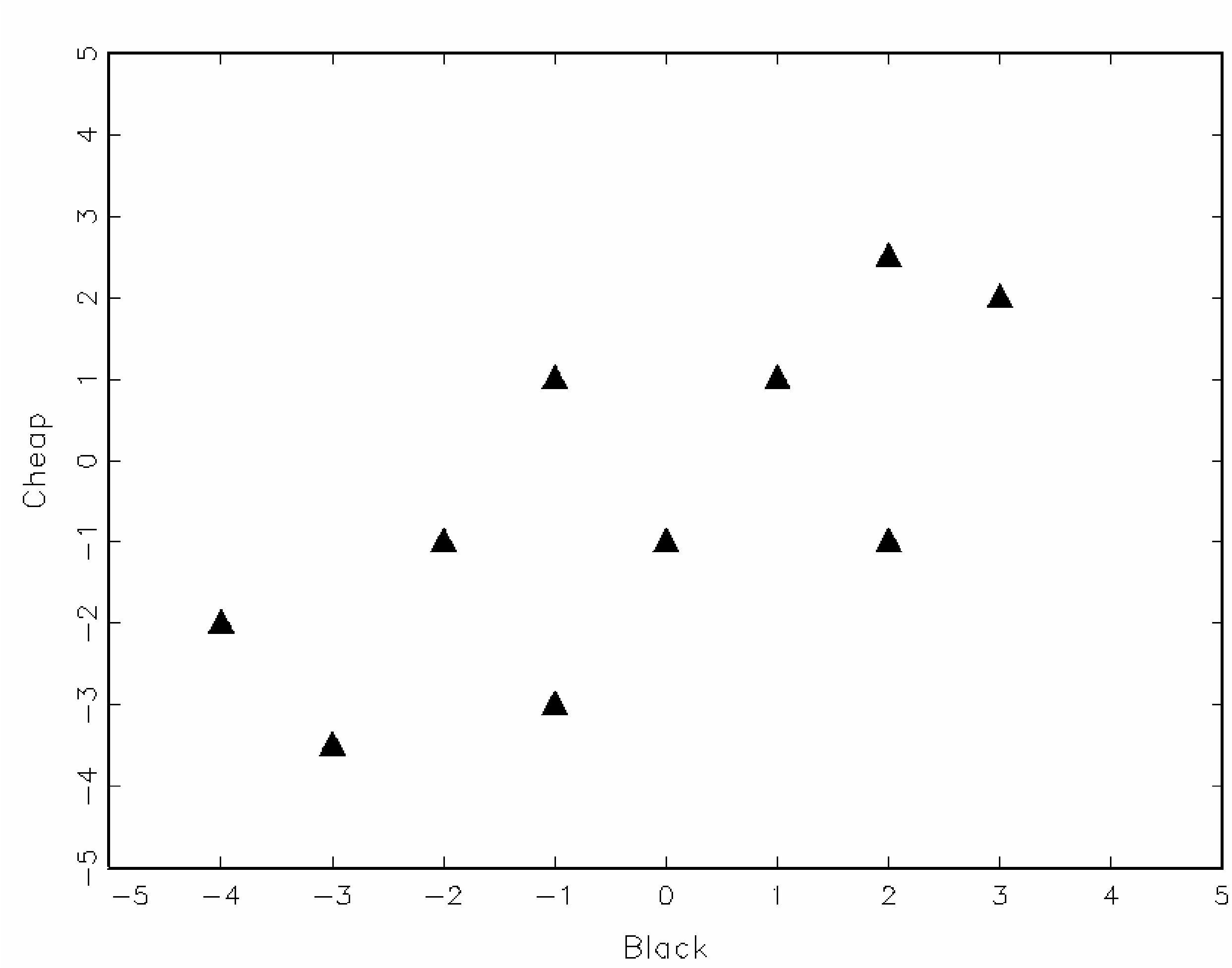
What about Marketing?

- HB revolution in how we think about customers

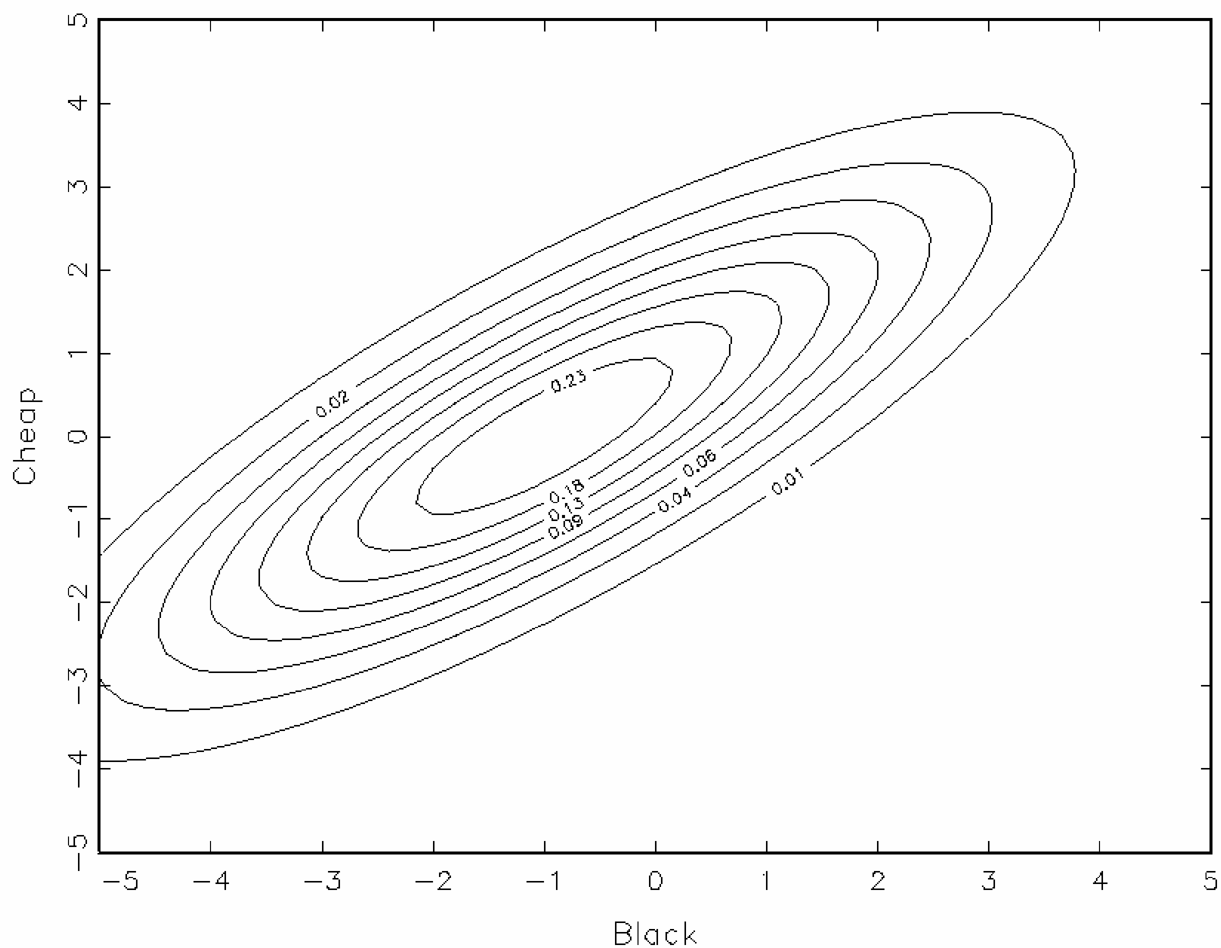
Henry Ford
All Customers are the Same



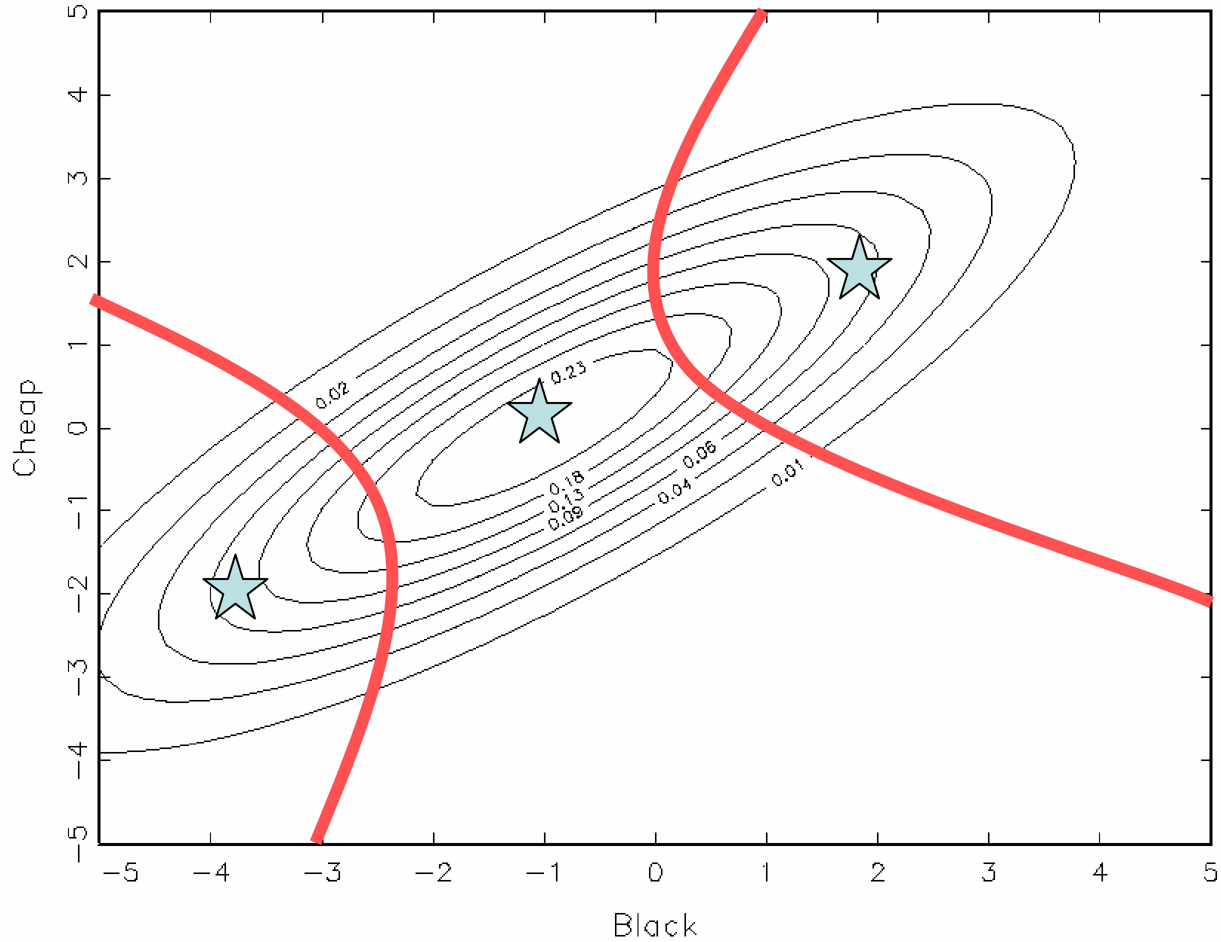
Alfred Sloan
Several Common Preferences



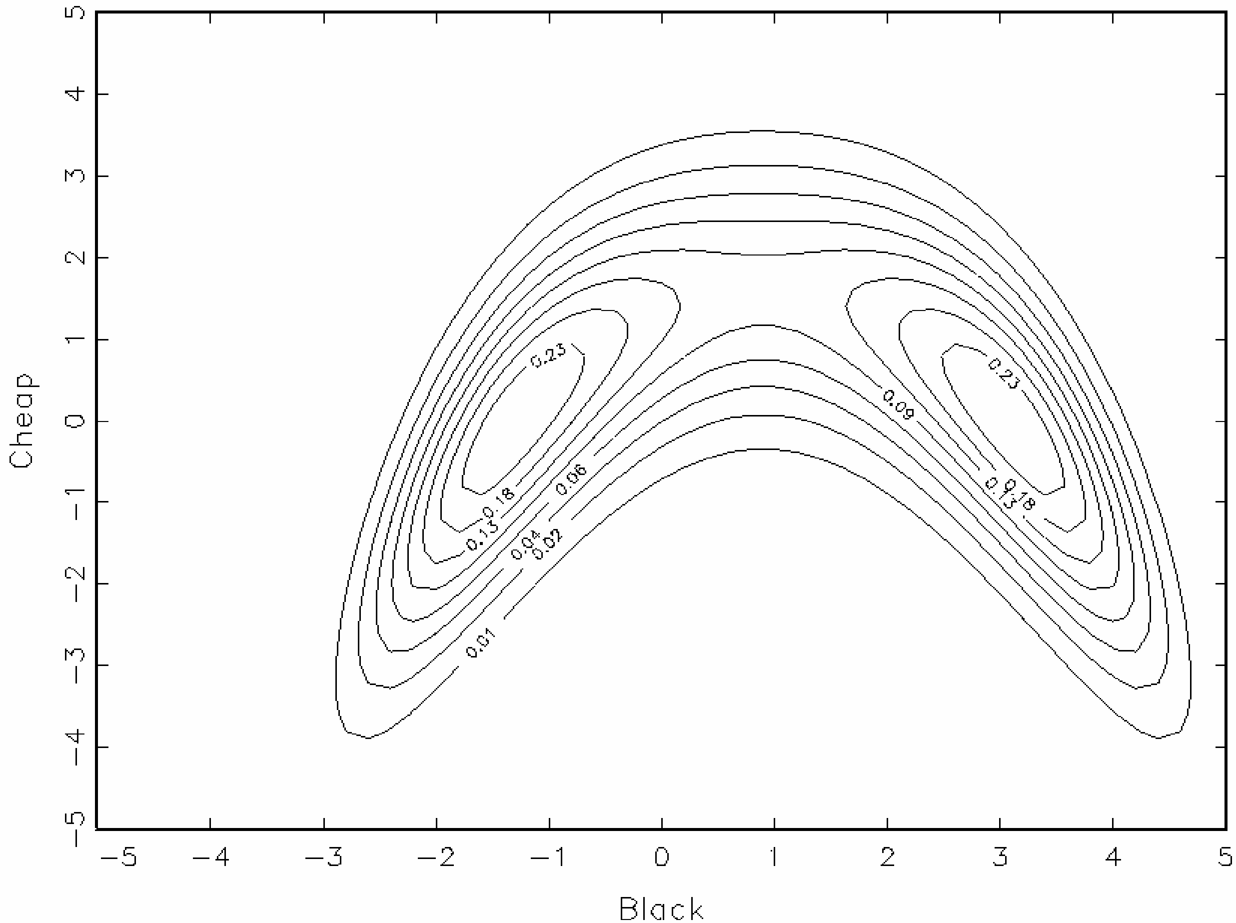
Continuous Heterogeneity



Profit Maximization



It Can Get Wild!



HB Model for Weekly Spending

- Within-subject model:
  $Y_{i,j} = \mu_i + \varepsilon_{i,j}$ and $\mathrm{var}(\varepsilon_{i,j}) = \sigma_i^2$
- Heterogeneity in mean weekly spending or between-subjects
  $\mu_i = \theta + \delta_i$ and $\mathrm{var}(\delta_i) = \tau^2$
- Prior Distribution
  θ is $N(u_0, v_0^2)$
- Variances are known

Variances & Covariances

- $Var(Y_{i,j}| \mu_i) = \sigma_i^2$ (known μ_i)
- $Var(Y_{i,j}|\theta) = \tau^2 + \sigma_i^2$ (unknown μ_i)
- $Cov(Y_{i,j}, Y_{i,k}) = \tau^2$ for j not equal to k
- Observations from different subjects are independent

Precisions = 1/Variance

$$\Pr(\theta) = \frac{1}{v_0^2} \text{ is prior precision}$$

$$\Pr(\mu_i \mid \theta) = \frac{1}{\tau^2}$$

$$\Pr(Y_{i,j} \mid \mu_i) = \frac{1}{\sigma_i^2} \text{ and } \Pr(Y_{i,j} \mid \theta) = \frac{1}{\tau^2 + \sigma_i^2}$$

$$\Pr(\overline{Y}_i \mid \mu_i) = \frac{n}{\sigma_i^2} \text{ and } \Pr(\overline{Y}_i \mid \theta) = \frac{1}{\tau^2 + \frac{\sigma_i^2}{n}}$$

Joint Distribution

$$P(Y, \mu, \theta) =$$

$$\underbrace{h\left(\theta \mid u_0, v_0^2\right)}_{\textbf{Prior}} \prod_{i=1}^{N} \underbrace{g\left(\mu_i \mid \theta, \tau^2\right)}_{\textbf{Between Subjects}} \prod_{j=1}^{n_i} \underbrace{f\left(y_{i,j} \mid \mu_i, \sigma_i^2\right)}_{\textbf{Within Subjects}}$$

Bayes Theorem

$$P(\mu, \theta \mid Y) = \frac{P(Y, \mu, \theta)}{\iint P(Y, \mu, \theta)\, d\mu\, d\theta}$$

$$P(\mu, \theta \mid Y) = \frac{P(Y, \mu, \theta)}{P(Y)}$$

Constant because Y is fixed & known

$$P(\mu, \theta \mid Y) \propto P(Y, \mu, \theta)$$

Bayes Estimator

- Posterior means are optimal under squared error loss

$$E(\mu_i|Y) \text{ and } E(\theta|Y)$$

- Measure of accuracy is posterior variance

$$\text{var}(\mu_i|Y) \text{ and } \text{var}(\theta|Y)$$

Posterior Distribution of θ

- Normal distribution
- Posterior mean is u_N
- Posterior variance is v_N^2
- Posterior precision is $Pr(\theta|Y) = 1/v_N^2$

Posterior Precision of θ

"Pr" = Precision = 1/Variance

$$\Pr\left(\theta \mid Y\right) = \Pr(\theta) + \sum_{i=1}^{N} \Pr\left(\overline{Y}_i \mid \theta\right)$$

$$\Pr\left(\theta\right) = \frac{1}{v_0^2} \text{ and } \Pr\left(\overline{Y}_i \mid \theta\right) = \frac{1}{\tau^2 + \dfrac{\sigma_i^2}{n_i}}$$

Posterior Mean of θ

$$u_N = w_0 u_0 + \sum_{i=1}^{N} w_i \overline{Y}_i$$

$$w_0 = \frac{\Pr(\theta)}{\Pr(\theta \mid Y)} \text{ and } w_i = \frac{\Pr\left(\overline{Y}_i \mid \theta\right)}{\Pr(\theta \mid Y)}$$

Updating of θ

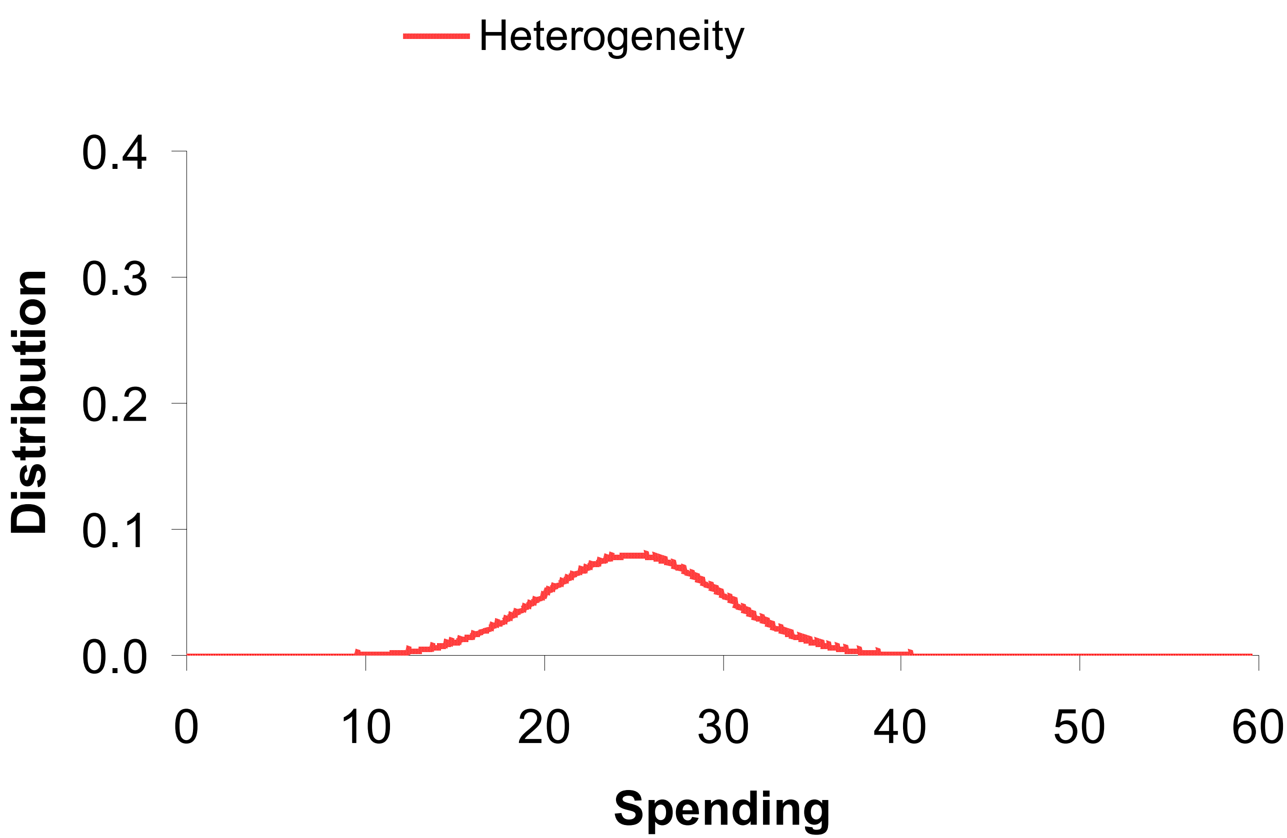
- Prior Mean ⇨ Posterior Mean
  u_0 ⇨ u_N
- Prior Var ⇨ Posterior Var
  v_0^2 ⇨ v_N^2

Posterior Mean of μ_i

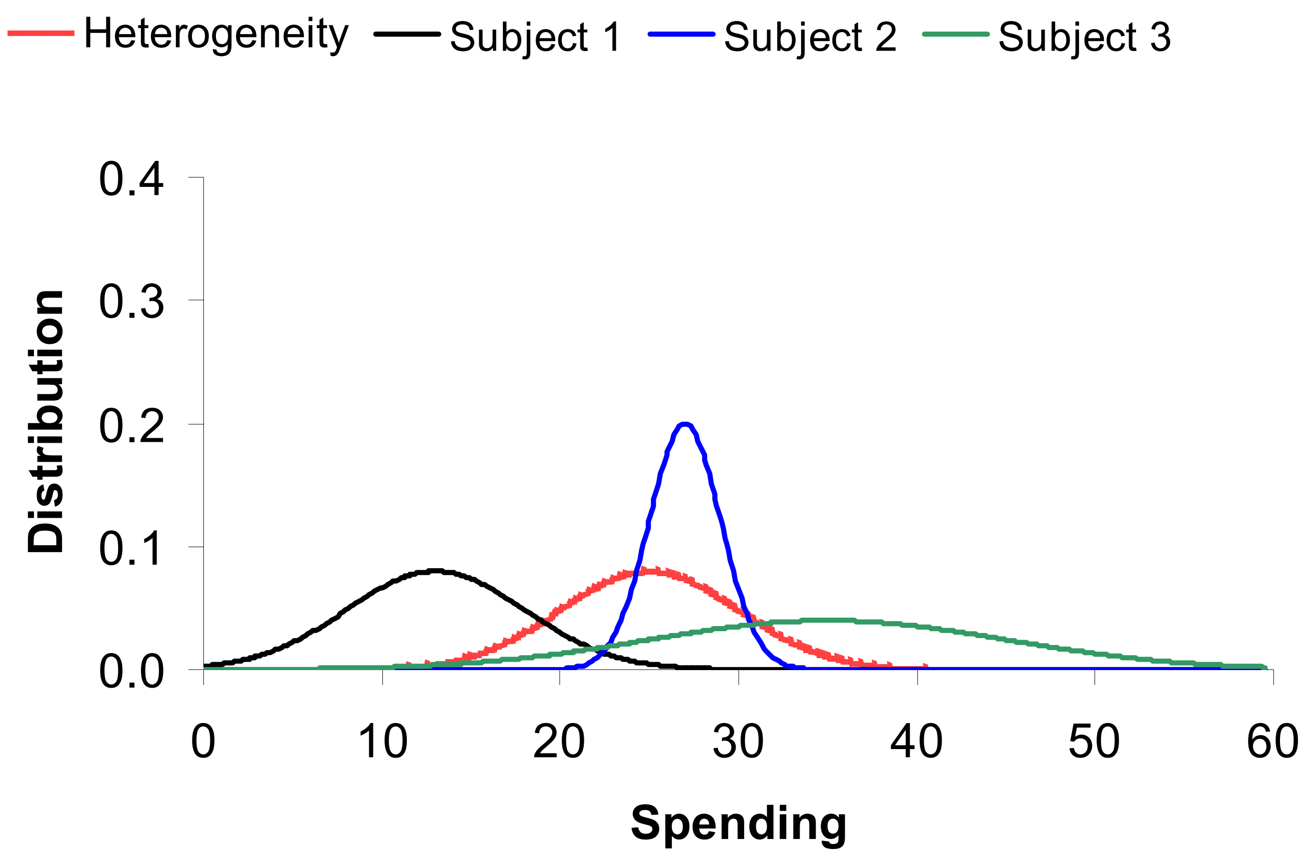
$$E\left[\mu_i \mid Y\right] = \alpha_i \overline{Y}_i + \left(1 - \alpha_i\right) u_N$$

$$\alpha_i = \frac{\Pr\left(\overline{Y}_i \mid \mu_i\right)}{\Pr\left(\mu_i \mid \theta\right) + \Pr\left(\overline{Y}_i \mid \mu_i\right)} = \frac{\dfrac{n_i}{\sigma_i^2}}{\dfrac{1}{\tau^2} + \dfrac{n_i}{\sigma_i^2}}$$

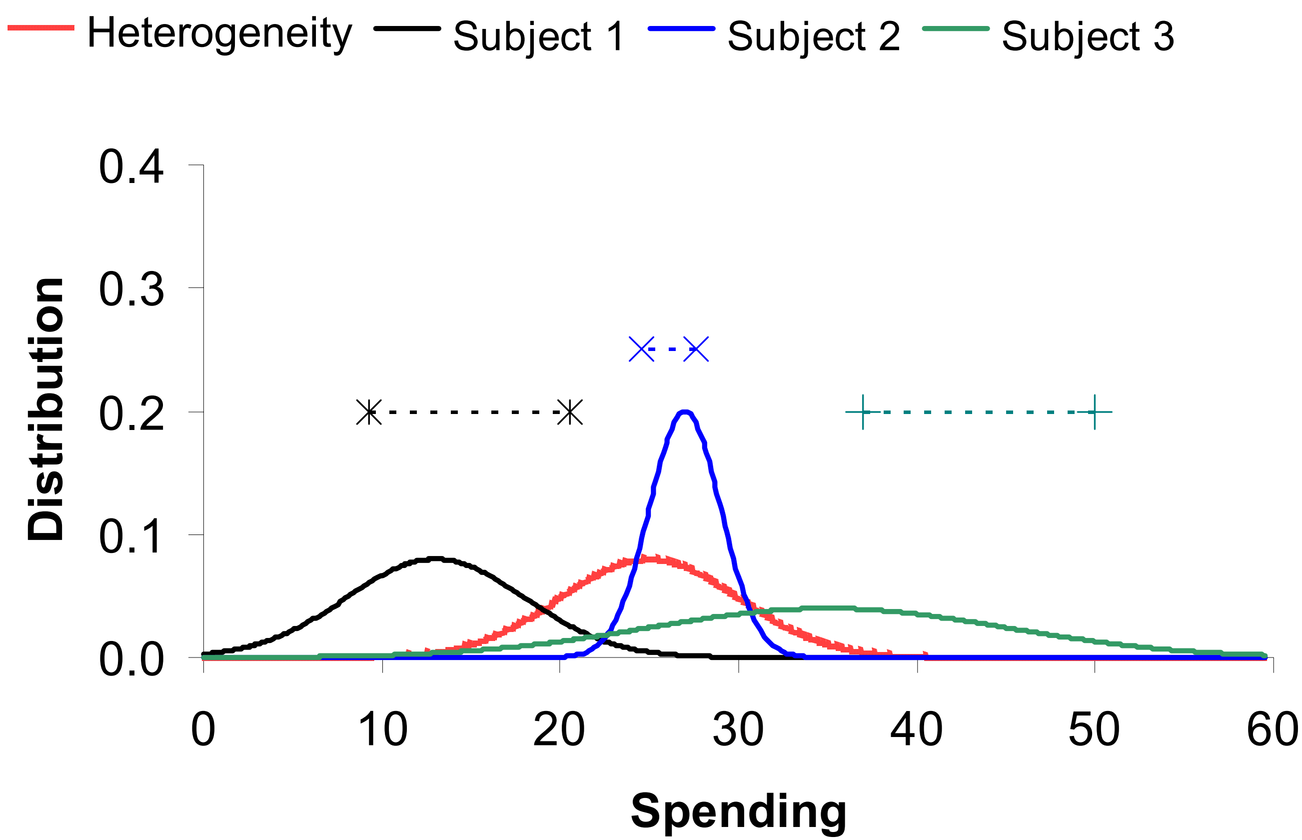
Between-Subject Heterogeneity in Mean Household Spending



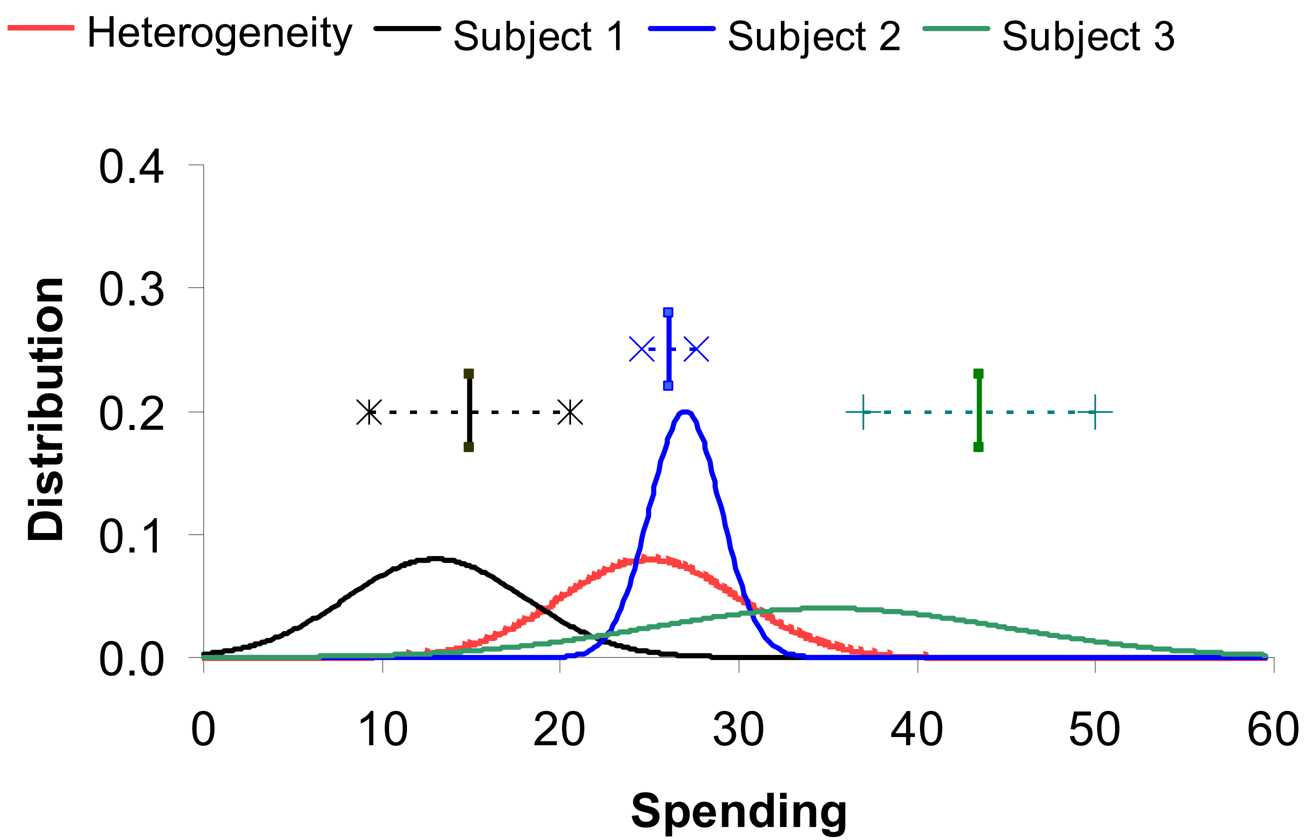
Between & Within Subjects Distributions



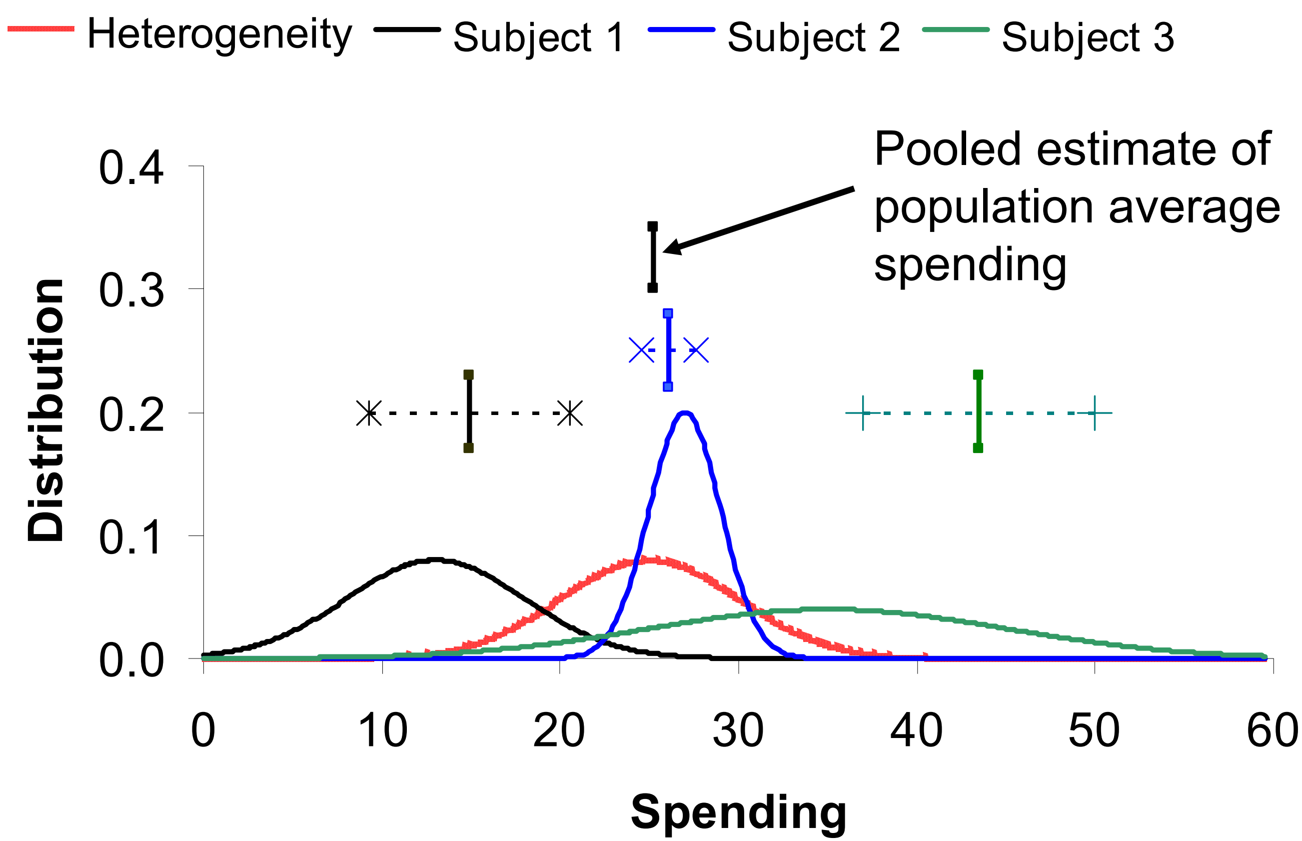
2 Observations per Subject



Subject Averages



Pooled Estimate of Mean



Sample Estimates

- Disaggregate estimate $\bar{Y}_i$ of μ_i only uses the observations for subject i.
  - Super if 30 or more observations per subject
- Pooled or aggregate estimator $\bar{\bar{Y}}$ of θ
  - Smaller sampling error
  - Ignores individual difference

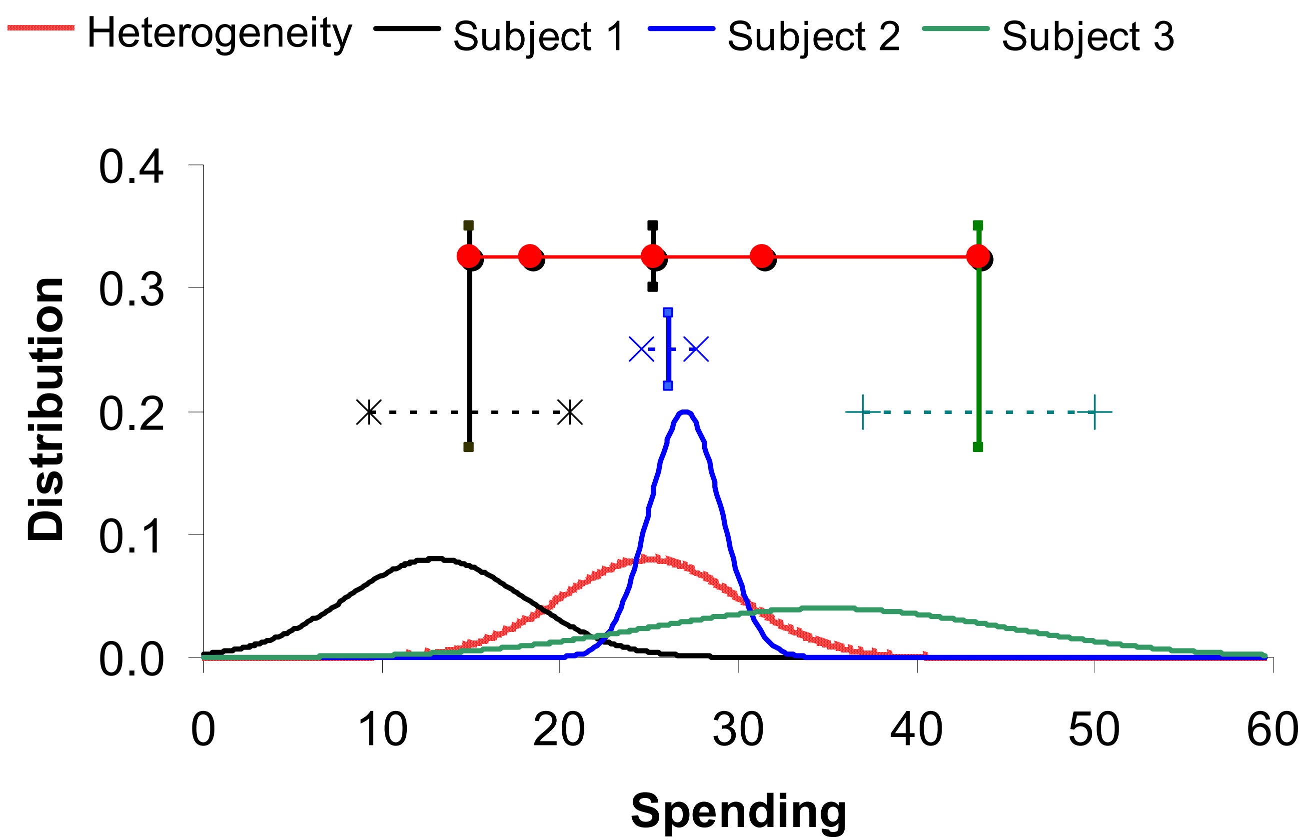
HB Shrinkage Estimator

- Take combination of individual average and pooled average

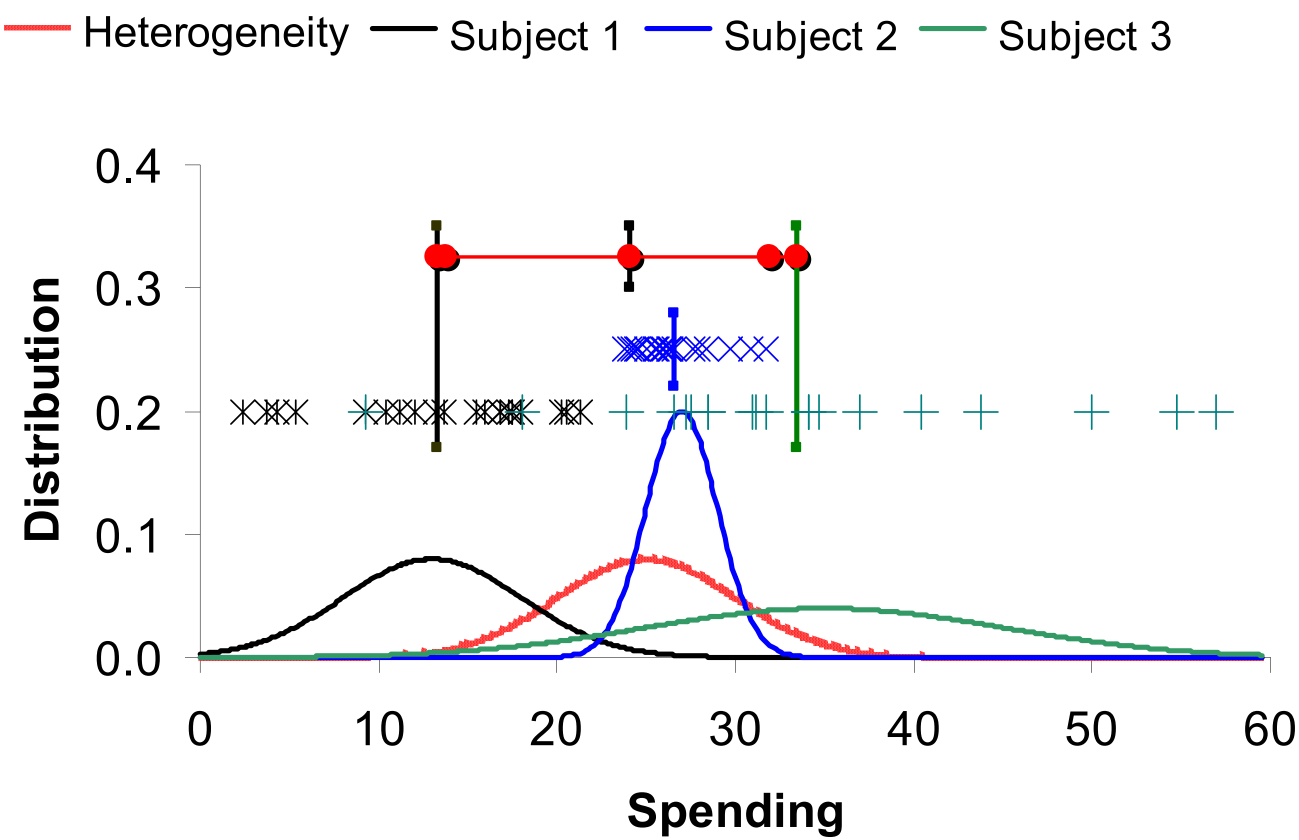
$$w_i \overline{Y}_i + (1 - w_i) \overline{\overline{Y}}$$

- What are the correct weights?
- HB automatically gives optimal weights based on
  - Prior variance of μ_i
  - Number of observations for subject i
  - Variance of past spending for subject i
  - Number of subjects
  - Amount of heterogeneity in household means

Shrinkage Estimates



20 Observations per Subject



Bayes & Shrinkage Estimates

- Bayes estimators automatically determine the optimal amount of shrinkage to minimize MSE for true parameters and predictions
- Borrows strength from all subjects
- Tradeoff some bias for variance reduction

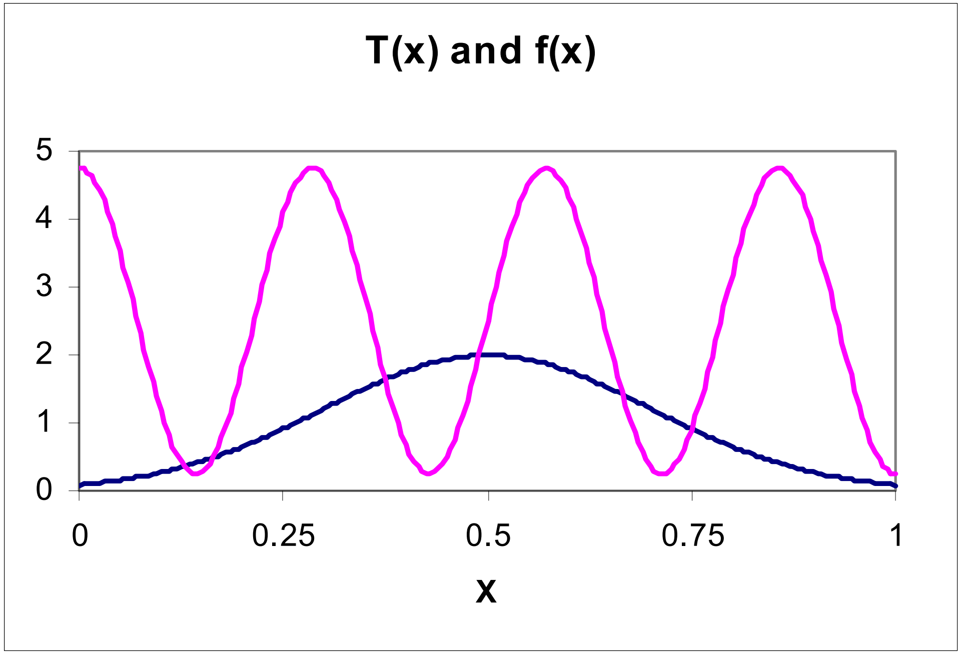
Good & Bad News

- Only simple models result in equations
- Models we use in marketing require numerical methods to compute posterior mean, posterior standard deviations, predictions and so on.

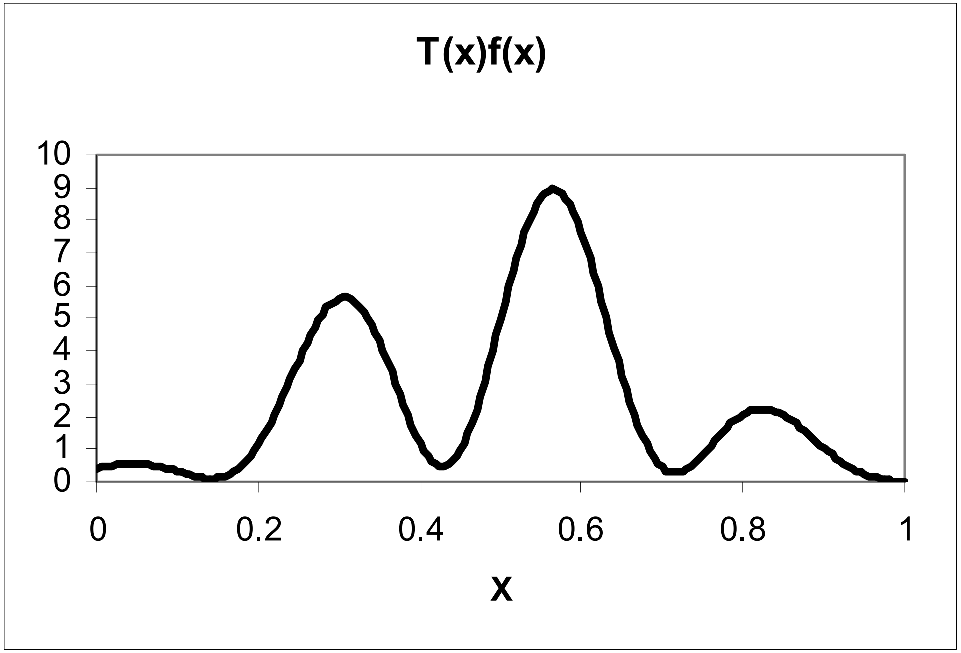
Numerical Integration

- Compute posterior mean of function T(θ).

$$E\left[T(\theta) \mid y\right] = \int T(\theta) p(\theta \mid y) d\theta$$

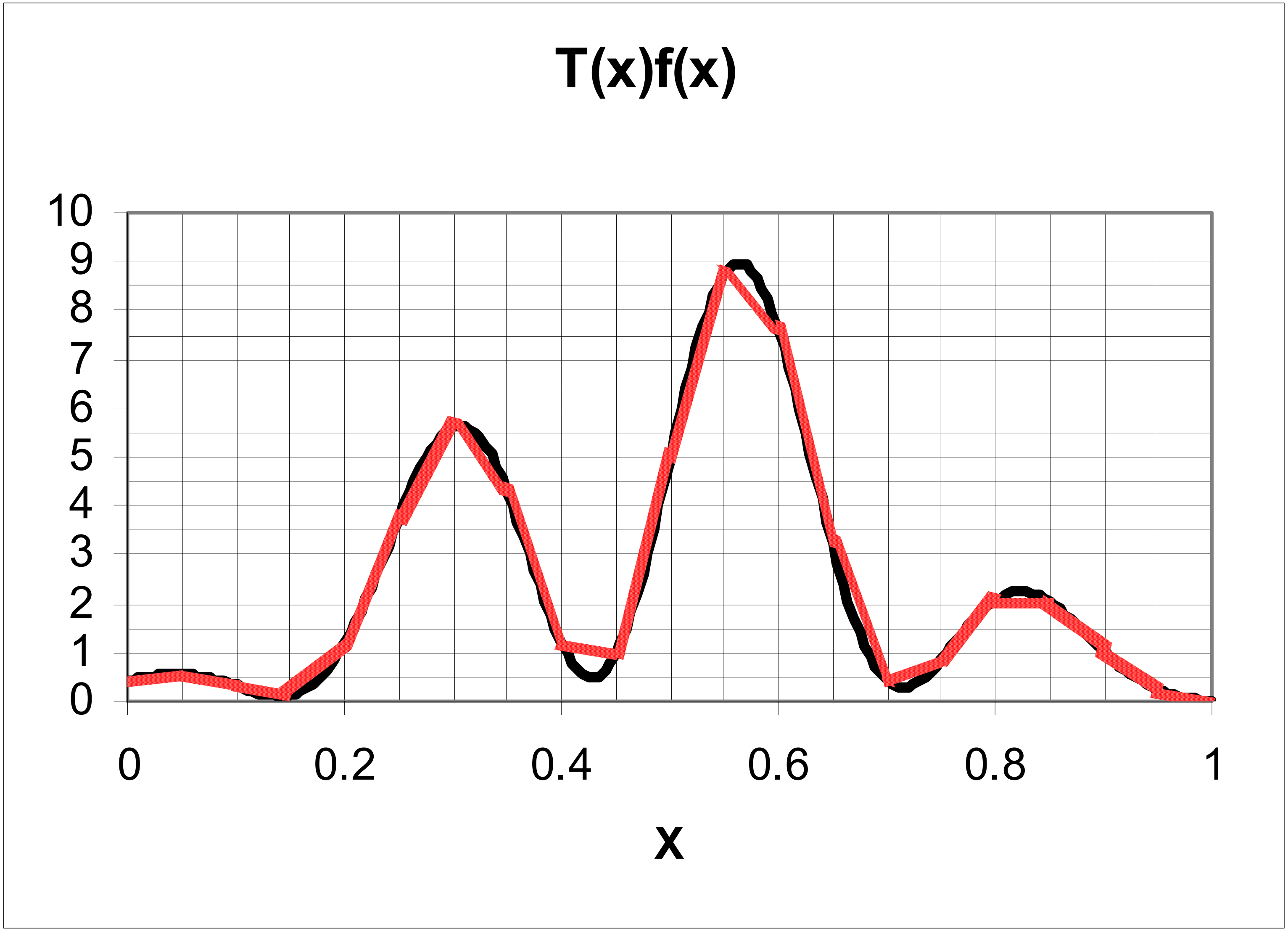


T(x) and f(x)
5
4
3
2
1
0
0
0.25
0.5
0.75
1
x



T(x)f(x)
10
9
8
7
6
5
4
3
2
1
0
0
0.2
0.4
0.6
0.8
1
x

Trapezoid Rule



Grid Methods

- Very accurate with few functional evaluations
- Need to know where the action is
- Does not scale well to higher dimensions
- You need to be very smart to make it work

Monte Carlo

- Generate random draws $\theta_1, \theta_2, \ldots, \theta_m$ from posterior distribution using a random number generator.

$$E\left[T(\theta) \mid y\right] \approx \frac{1}{m} \sum_{j=1}^{m} T\left(\theta_j\right)$$

- What happened to the density of θ?

Good & Bad News

- If your computer has a random number generator for the posterior distribution, Monte Carlo is a snap to do.
- Your computer almost never has the correct random number generator.

Importance Sampling

- Would like to sample from density f
- You have a good random number generator for the density g
- Importance sampling lets you generate random deviates from g to evaluate expectations with respect to f.
- Generate $\phi_1, \ldots, \phi_m$ from g

Importance Sampling Approximation

$$\int T(\theta)f(\theta)d\theta = \int T(\varphi)\frac{f(\varphi)}{g(\varphi)}g(\varphi)d\varphi$$

$$\approx \frac{\sum_{i=1}^{m} T(\varphi_i)r(\varphi_i)}{\sum_{i=1}^{m} r(\varphi_i)} = \sum_{i=1}^{m} T(\varphi_i)w(\varphi_i)$$

$$r(\varphi_i) \propto \frac{f(\varphi_i)}{g(\varphi_i)} \text{ and } w(\varphi_i) = \frac{r(\varphi_i)}{\sum_{j=1}^{m} r(\varphi_j)}$$

Markov Chain Monte Carlo

- Extension of Monte Carlo
- Random draws are not independent
- Joint distribution $f(\beta,\phi)$ does not have a convenient random number generator.
- Conditional distributions $g(\phi|\beta)$ and $h(\beta|\phi)$ are easy to generate from.

Iterative Generation from Full Conditionals

- Start at ϕ_0
- Generate β_1 from $h(\beta|\phi_0)$.
- Generate ϕ_1 from $g(\phi|\beta_1)$.
- …
- Generate β_{m+1} from $h(\beta|\phi_m)$
- Generate ϕ_{m+1} from $g(\phi|\beta_{m+1})$

Baseball Example

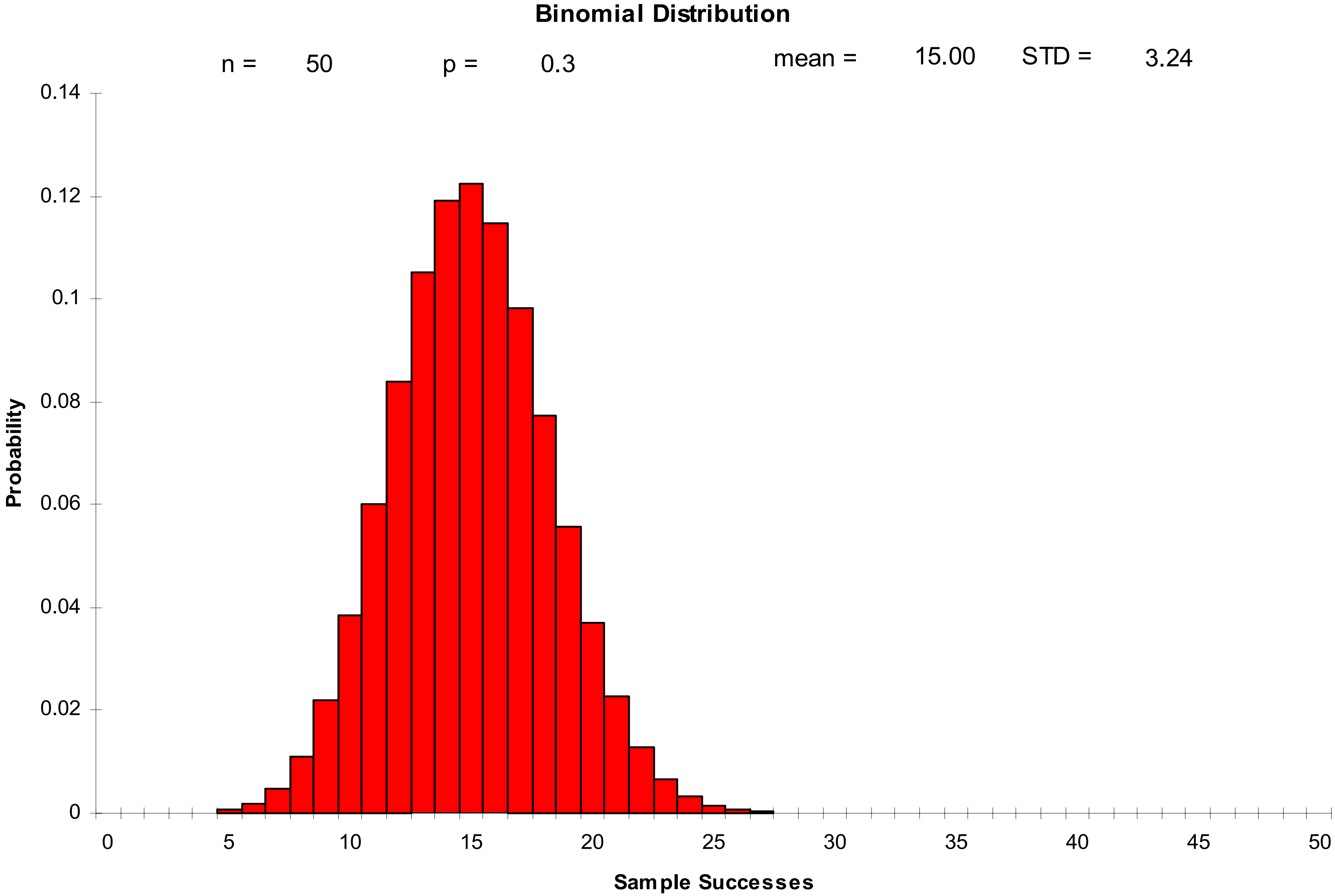
- 90 MLB Players in 2000 season.
- Observe at bats (AB) and hits (BA) in May
- Infer distribution of batting averages across players.
- Predict batting averages in October using data from May.

Baseball Batting Averages

	The Cleveland Indians - 1995			
	May		October	
	BA	AB	BA	AB
Murray	.442	43	.323	436
Belle	.400	45	.317	546
Vizquel	.204	49	.260	542
Pena	.148	27	.262	263

Estimating a Probability

- n at bats in May
- X = number of hits
- p = batting average for season
- X has a binomial distribution
  – mean np
  – variance np(1-p)


Binomial Distribution
n = 50
p = 0.3
mean = 15.00
STD = 3.24
Probability
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20
25
30
35
40
45
50
Sample Successes

Need Prior for Batting Average p

- $0 < p < 1$
- Beta distribution is popular choice
- It has two parameters: α and β
- Density is proportional to $p^{\alpha-1}(1-p)^{\beta-1}$
- Prior Mean = $\alpha/(\alpha+\beta)$

Beta Prior for p

$$f\left[p \mid \alpha, \beta\right] = Beta(p \mid \alpha, \beta)$$

$$= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad for \quad 0 \le p \le 1$$

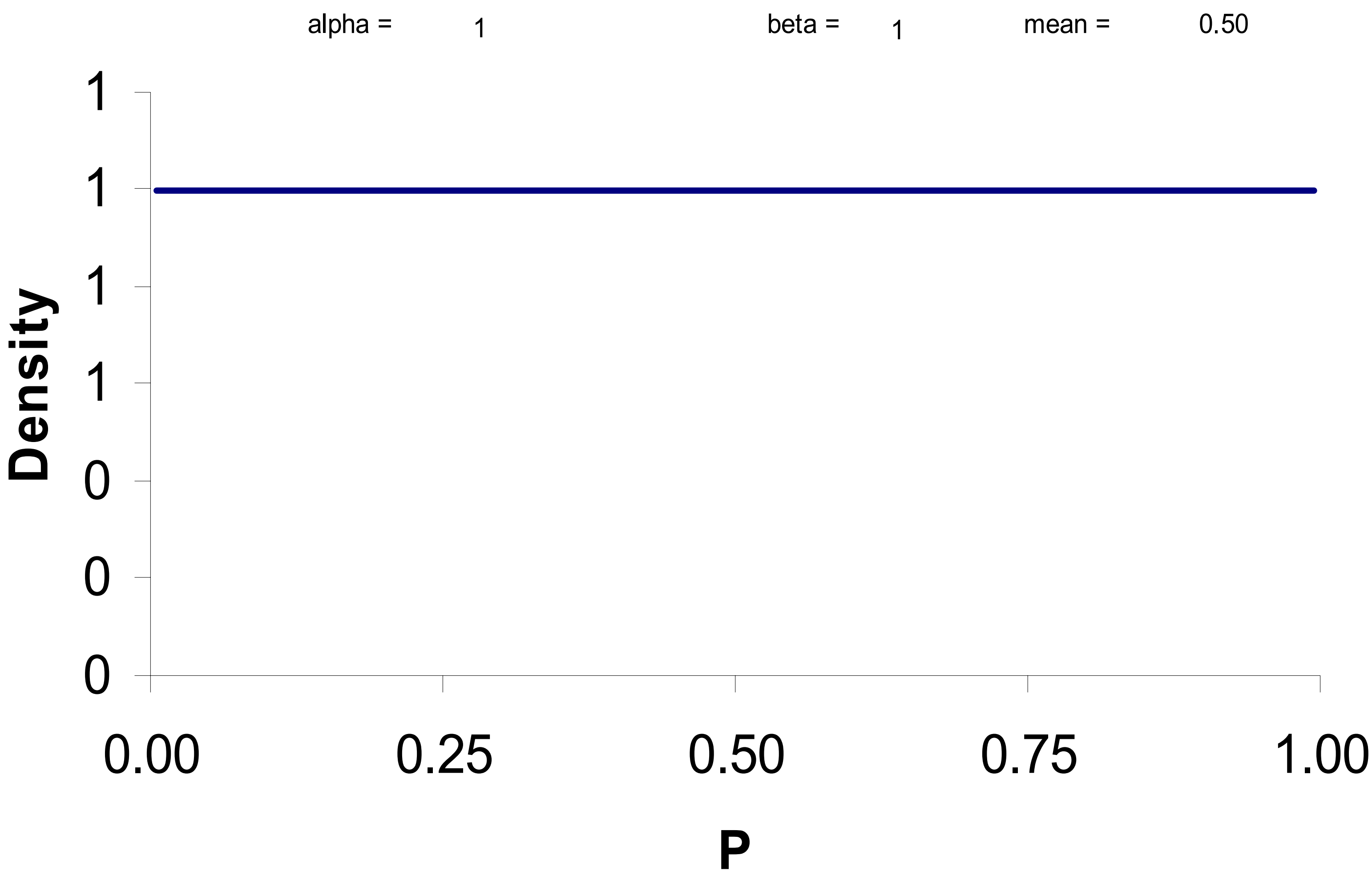
$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

Mean and Variance

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

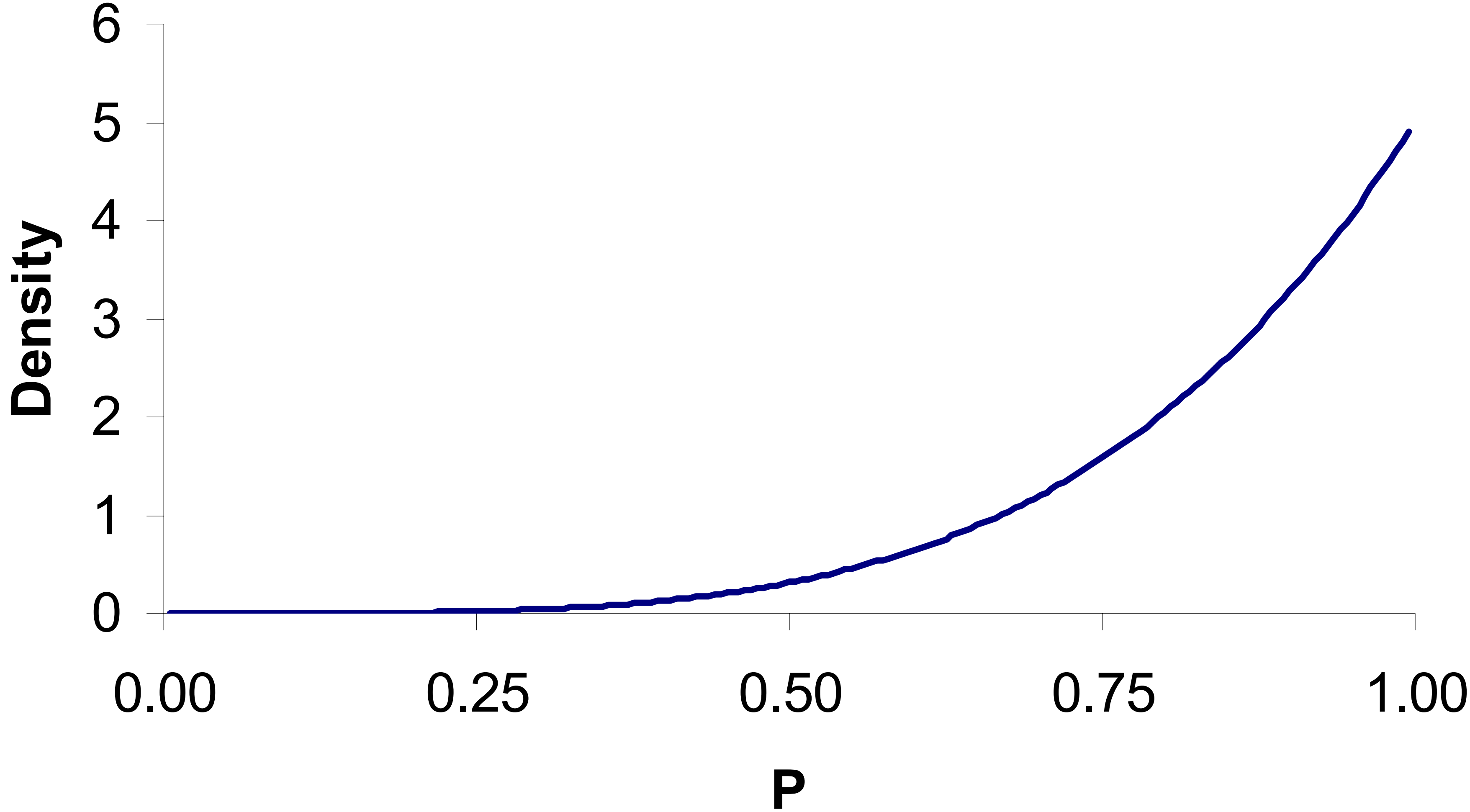
$$V(p) = \frac{E(p)\left[1 - E(p)\right]}{\alpha + \beta + 1}$$

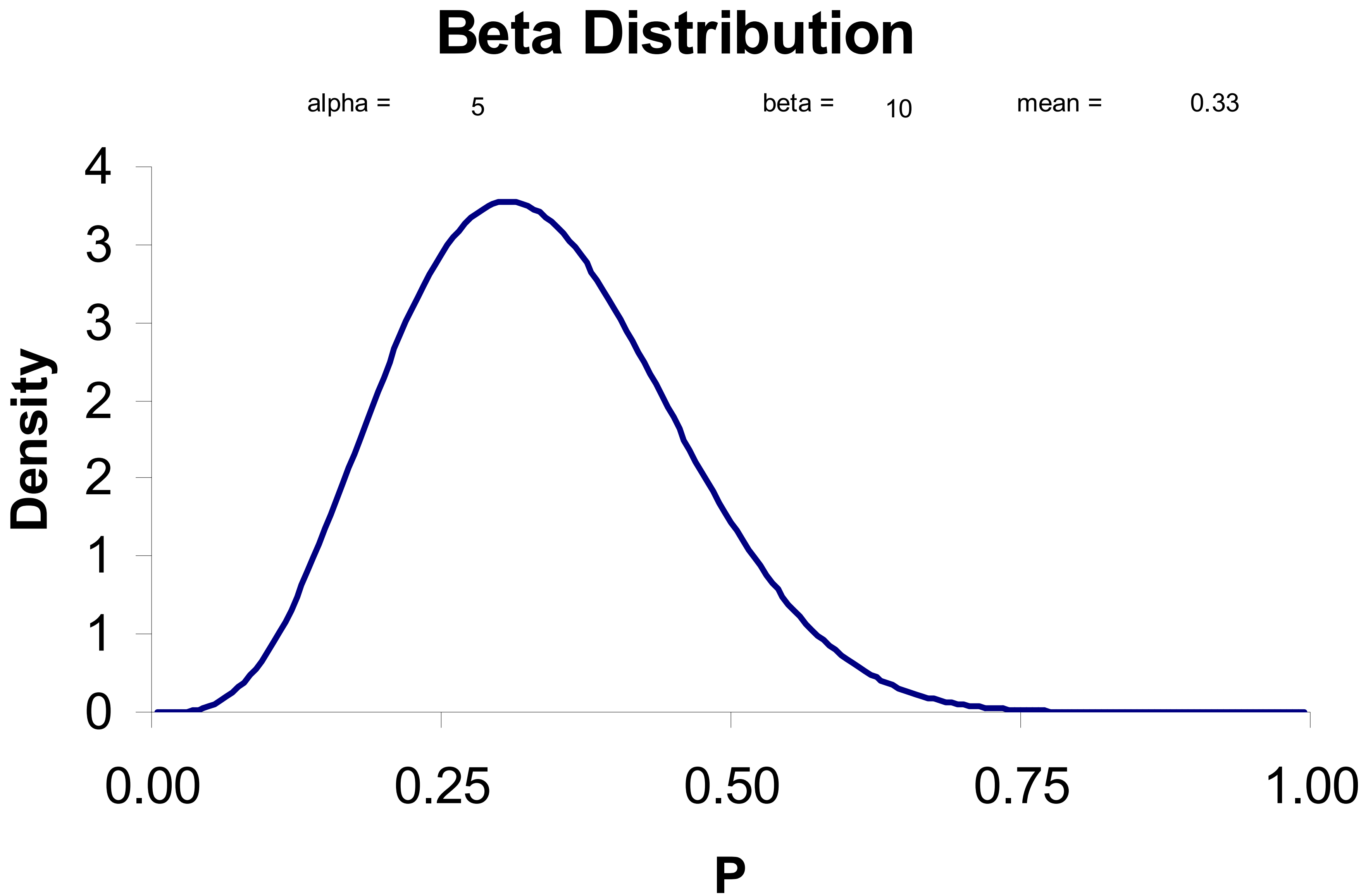
Beta Distribution



Beta Distribution

alpha = 5 beta = 1 mean = 0.83

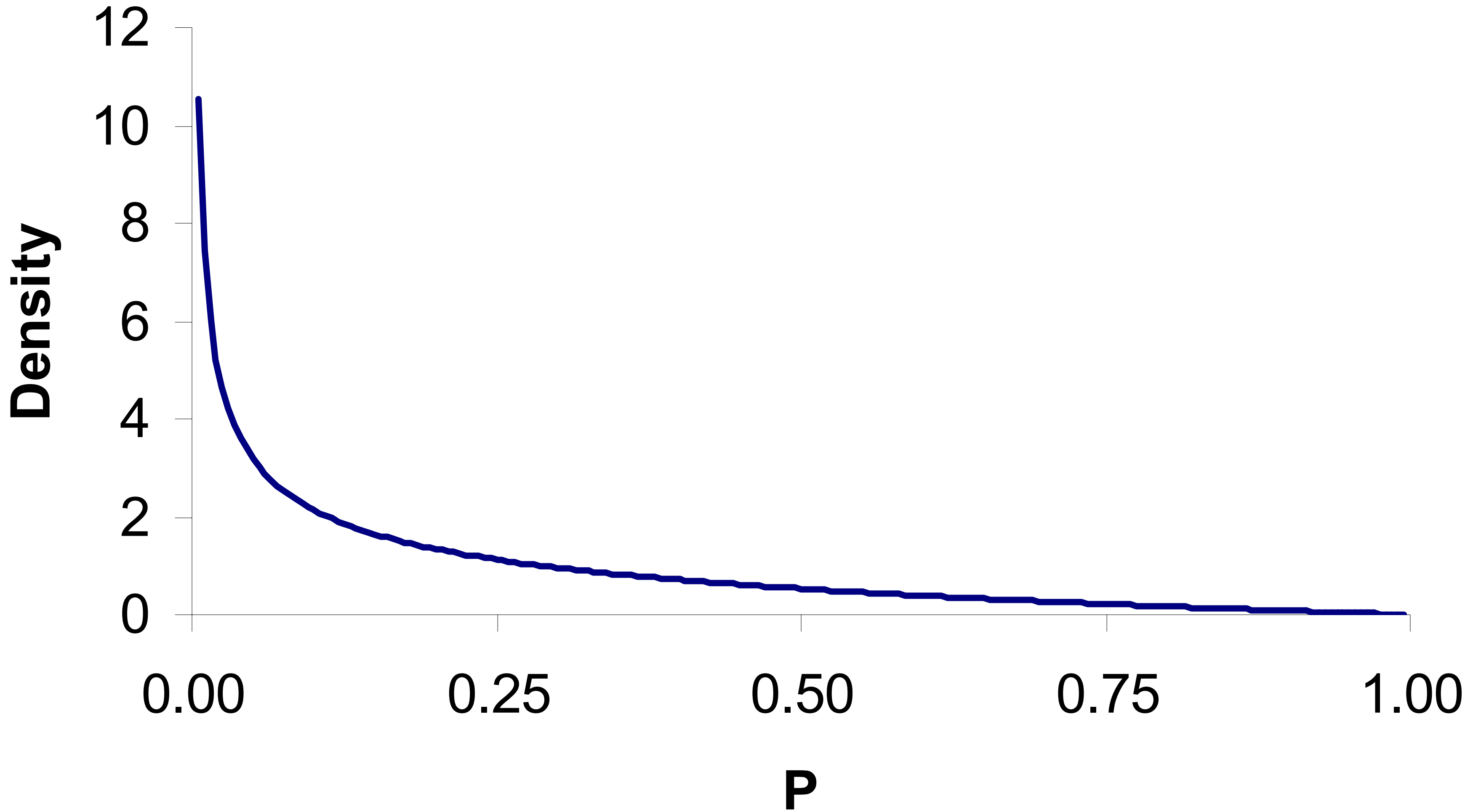




Beta Distribution
alpha = 5
beta = 10
mean = 0.33
Density
4
3
3
2
2
1
1
0
0.00
0.25
0.50
0.75
1.00
P

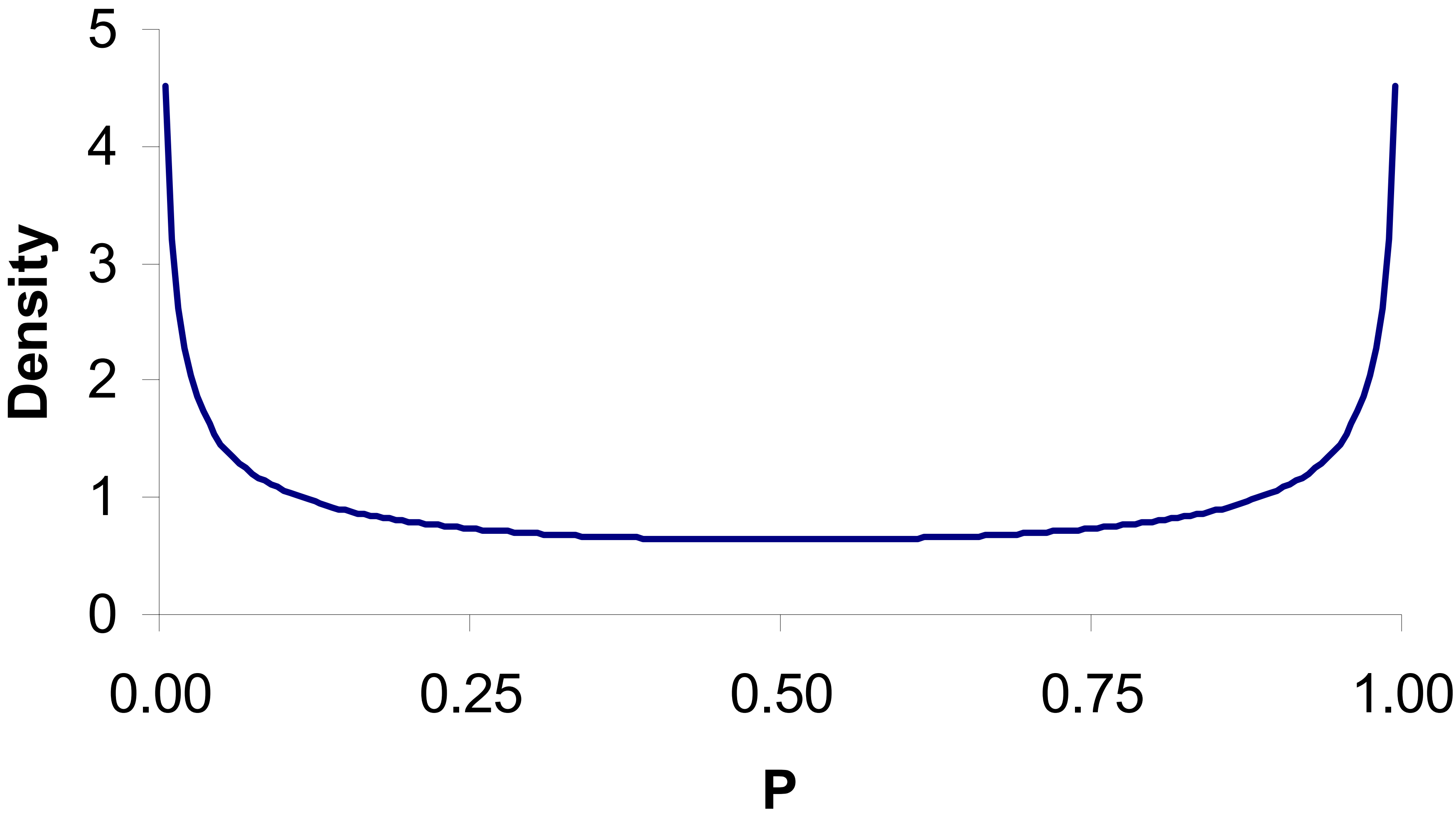
Beta Distribution

alpha = 0.5 beta = 2 mean = 0.20



Beta Distribution

alpha = 0.5 beta = 0.5 mean = 0.50



Bayes Theorem:
Update prior for p after observing n and x

$$f[p \mid x, \alpha, \beta] = \frac{\Pr[x \mid p] f[p \mid \alpha, \beta]}{\int_0^1 \Pr[x \mid q] f[q \mid \alpha, \beta] dq}$$

$$\propto \Pr[x \mid p] f[p \mid \alpha, \beta]$$

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

$$= Beta(p \mid \alpha + x, \beta + n - x)$$

Inference About P: Posterior is also Beta

Prior Parameters	Posterior Parameters
α	α+x
β	β+n-x

Posterior Mean of p:
Its another shrinkage estimator

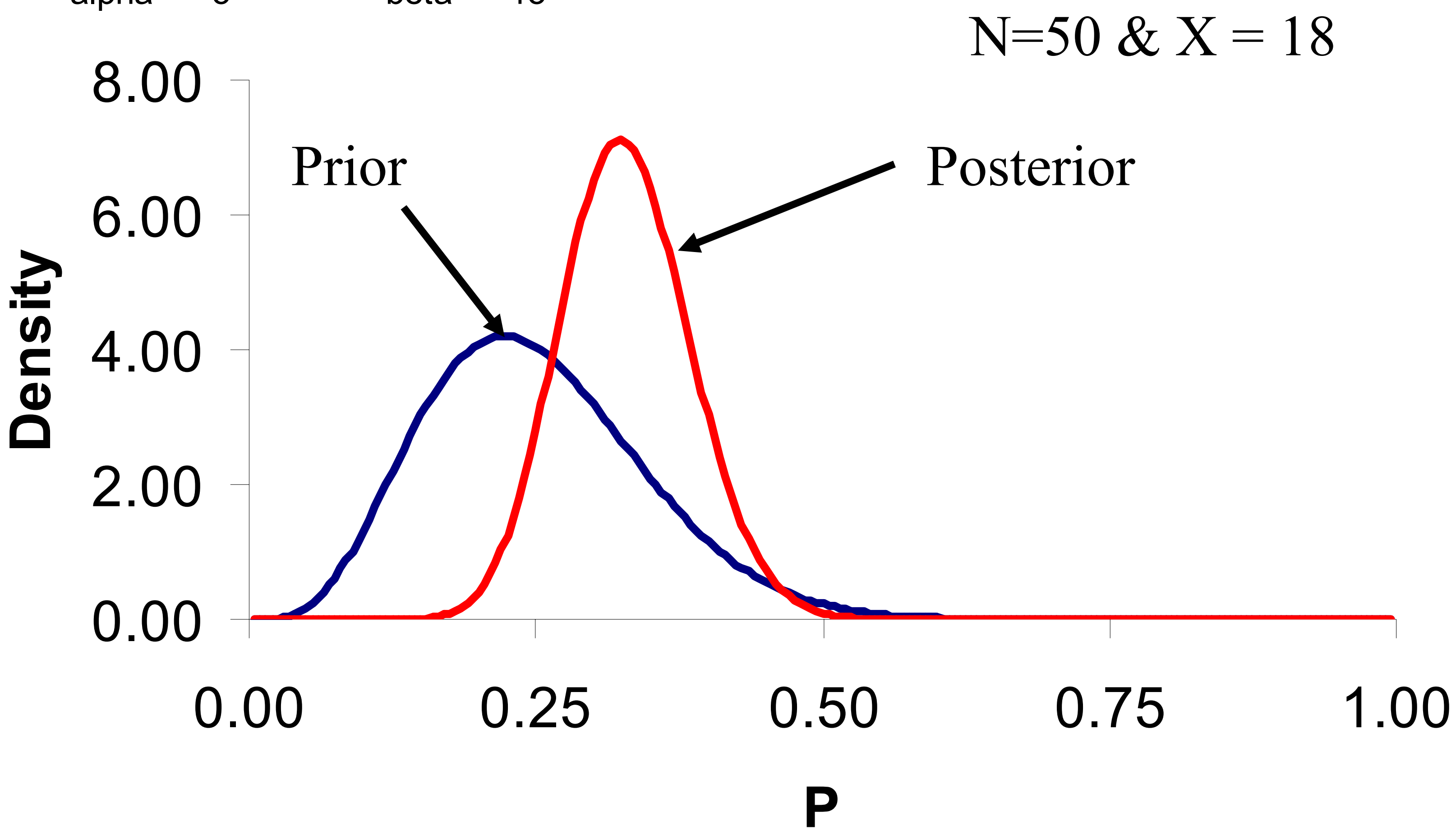
$$w_n \hat{p}_n + (1 - w_n)\left(\text{Prior Mean}\right)$$

$$\hat{p}_n = \frac{x}{n} \text{ and } w_n = \frac{n}{\alpha + \beta + n}$$

prior mean = 0.25

p=Beta & x=Binomial

alpha = 5 beta = 15



Hierarchical Bayes Model

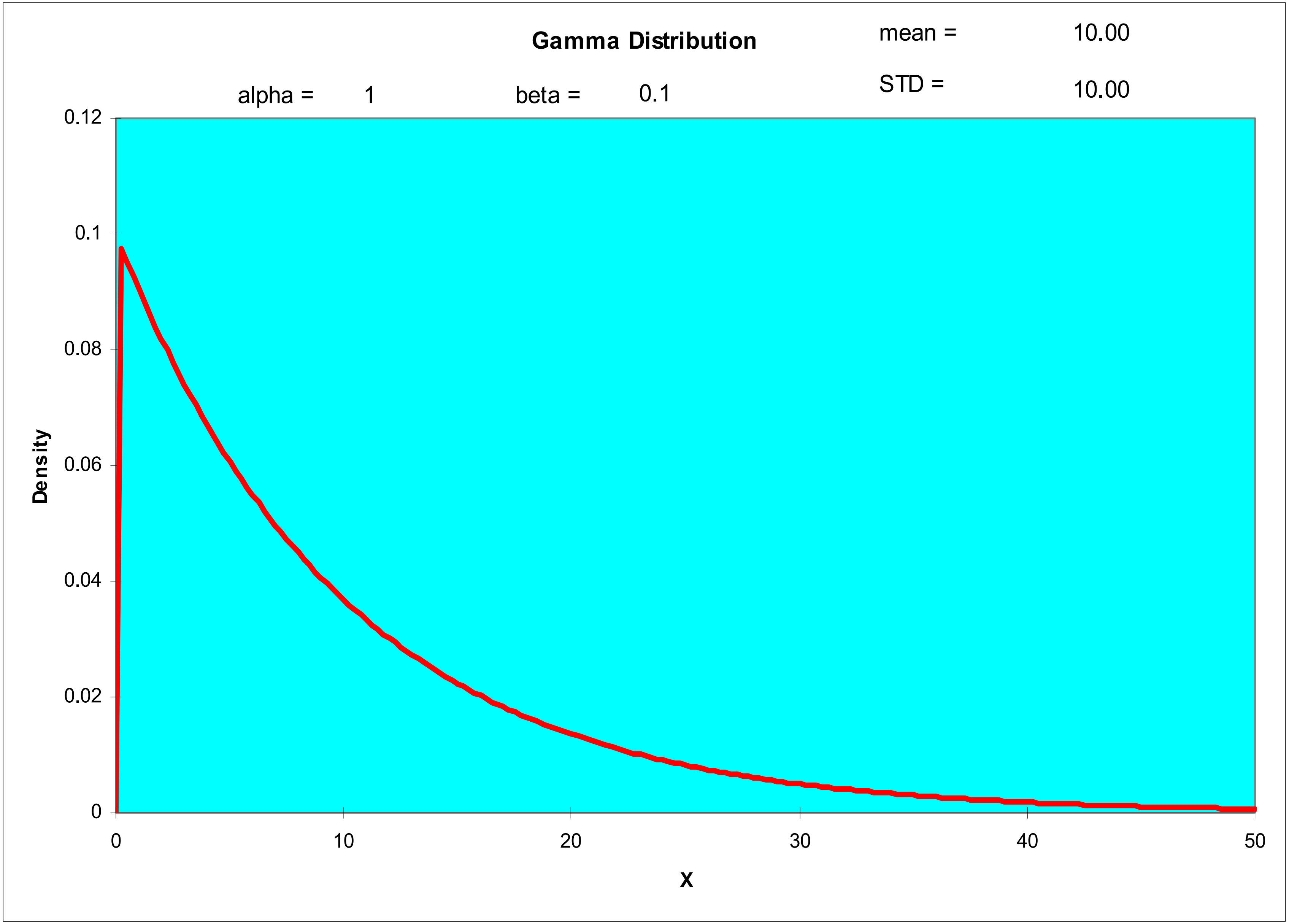
- Variation within batter i:
  - X_i given p_i has a binomial distribution
- Variation among batters:
  - p_i is a beta distribution with parameters α and β.
- Prior distribution for α and β
  - Gamma (chi-square) distribution

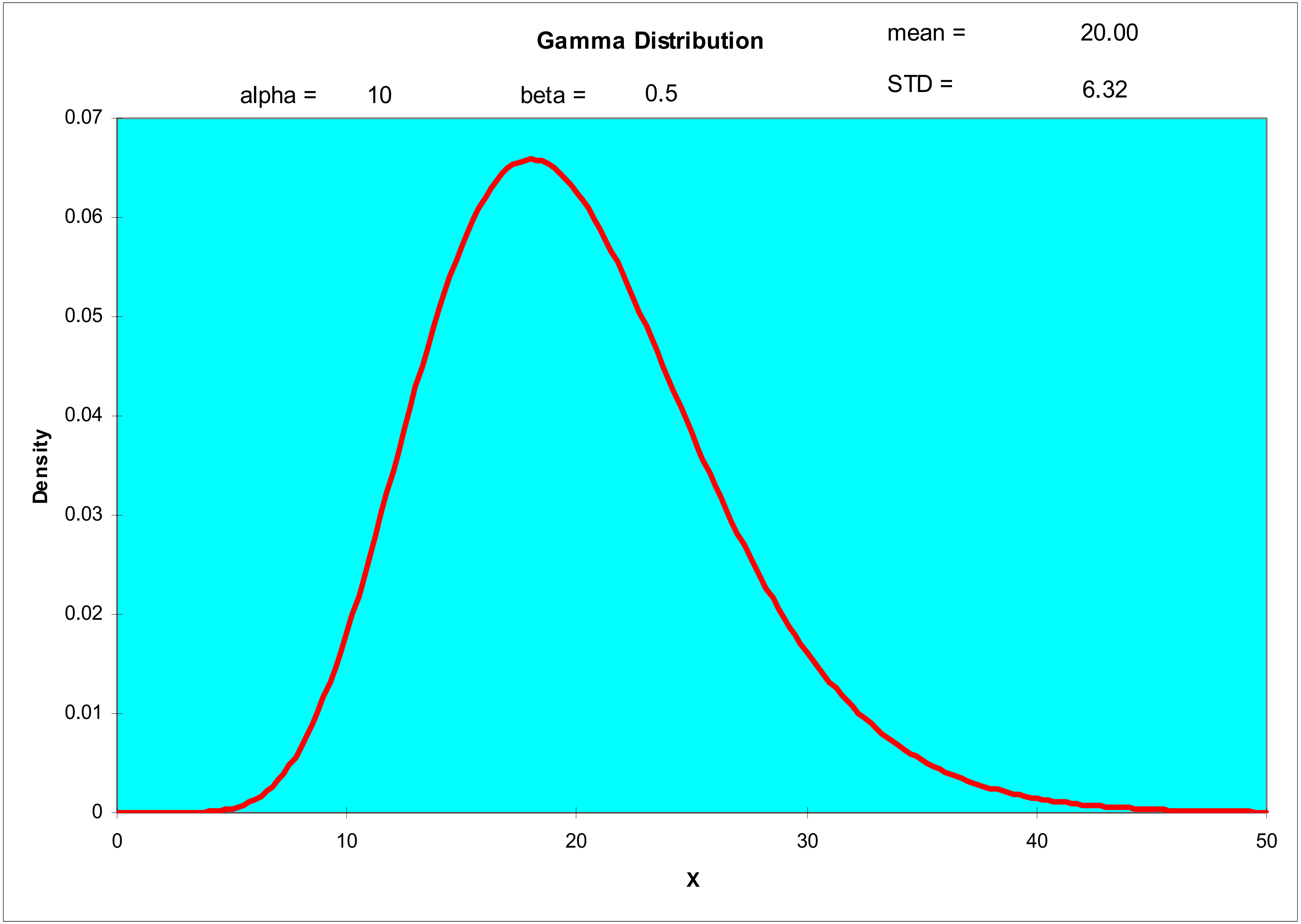
Gamma Distribution

$$g(y) = G(y \mid r, s)$$

$$= \frac{s^{r}}{\Gamma(r)} y^{r-1} e^{-sy} \quad for \quad y > 0$$

$$E(Y) = \frac{r}{s} \quad and \quad V(Y) = E(Y)\frac{1}{s}$$


Gamma Distribution
mean = 10.00
STD = 10.00
alpha = 1
beta = 0.1
Density
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
X


Gamma Distribution
mean = 20.00
STD = 6.32
alpha = 10
beta = 0.5
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Density
0
10
20
30
40
50
X

Specify Prior Parameters: r, s, u & v

- Priors: α is G(r,s) & β is G(u,v).
- $E(\alpha) = r/s$ and $V(\alpha) = E(\alpha)/s$.
- s determines variance relative to mean.
- I used s = 0.25 or the variance is four times larger than the mean.
- Same for v.

$$E(p) = E\left[E(p \mid \alpha, \beta)\right]$$
$$= \frac{r}{r+u}$$
$$= p_0$$

$$c = V\left[E(p \mid \alpha, \beta)\right] = \frac{p_0(1-p_0)}{r+u+1}$$

$$r = p_0\left(\frac{p_0(1-p_0)}{c} - 1\right) \quad and$$

$$u = (1-p_0)\left(\frac{p_0(1-p_0)}{c} - 1\right)$$

Specify Prior Parameters

- Guess a mean of all batting averages:

$$p_0 = 0.25$$

- Measure of my uncertainty of that guess:

$$c = 0.01$$

- Parameter $r = 4.4$
- Parameter $u = 13.3$

MCMC for Batting Averages

- Need full conditionals for p_i give α and β
  - Beta distribution
- Need full conditionals for α and β given p_i .
  - Unknown distribution
  - Use Metropolis algorithm

MCMC: Full Conditionals for Player i Batting Average p_i

$$f\left[p_i \mid x_i, \alpha, \beta\right] \propto \Pr\left(x_i \mid p_i\right) f\left(p_i \mid \alpha, \beta\right)$$

$$\propto p_i^{a+x_i-1}\left(1-p_i\right)^{\beta+n_i-x_i-1}$$

$$= Beta\left(p_i \mid \alpha+x_i, \beta+n_i-x_i\right)$$

MCMC: Full Conditional for α and β

$$g(\alpha, \beta \mid x_1, \ldots . x_n, p_1, \ldots . p_n)$$

$$\propto \prod_{i=1}^{n} p_i^{\alpha-1} (1 - p_i)^{\beta-1} g(\alpha \mid r, s) g(\beta \mid u, v)$$

Metropolis Algorithm

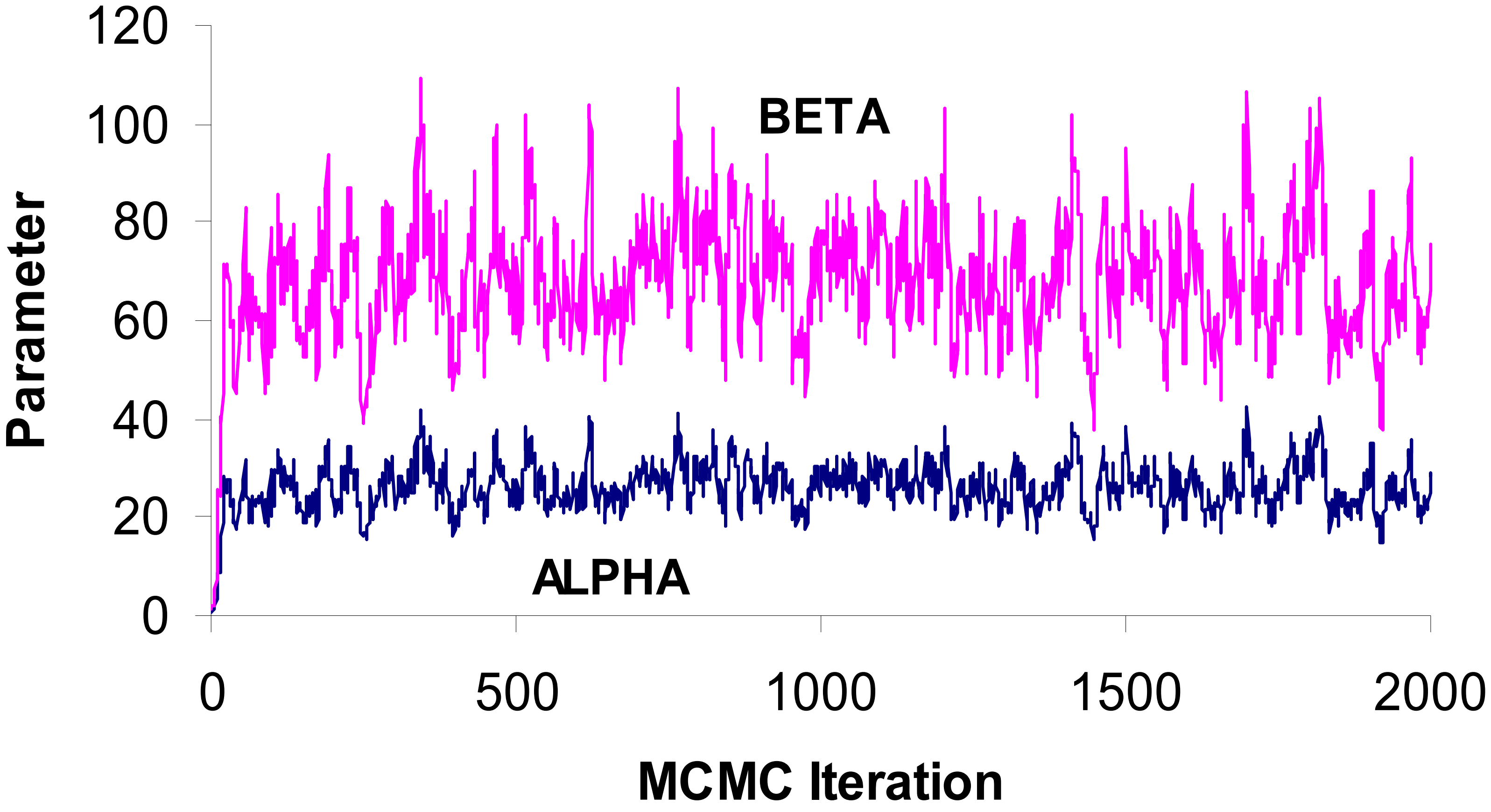
- Want to generate θ from f
- Instead, generate candidate value ϕ from g(.|θ)
  - Density g can depend on θ
  - eg Random walk: $\phi = \theta + \delta$
- With probability $\alpha(\theta,\phi)$ accept ϕ as the new value of θ
- With probability $1-\alpha(\theta,\phi)$ keep θ

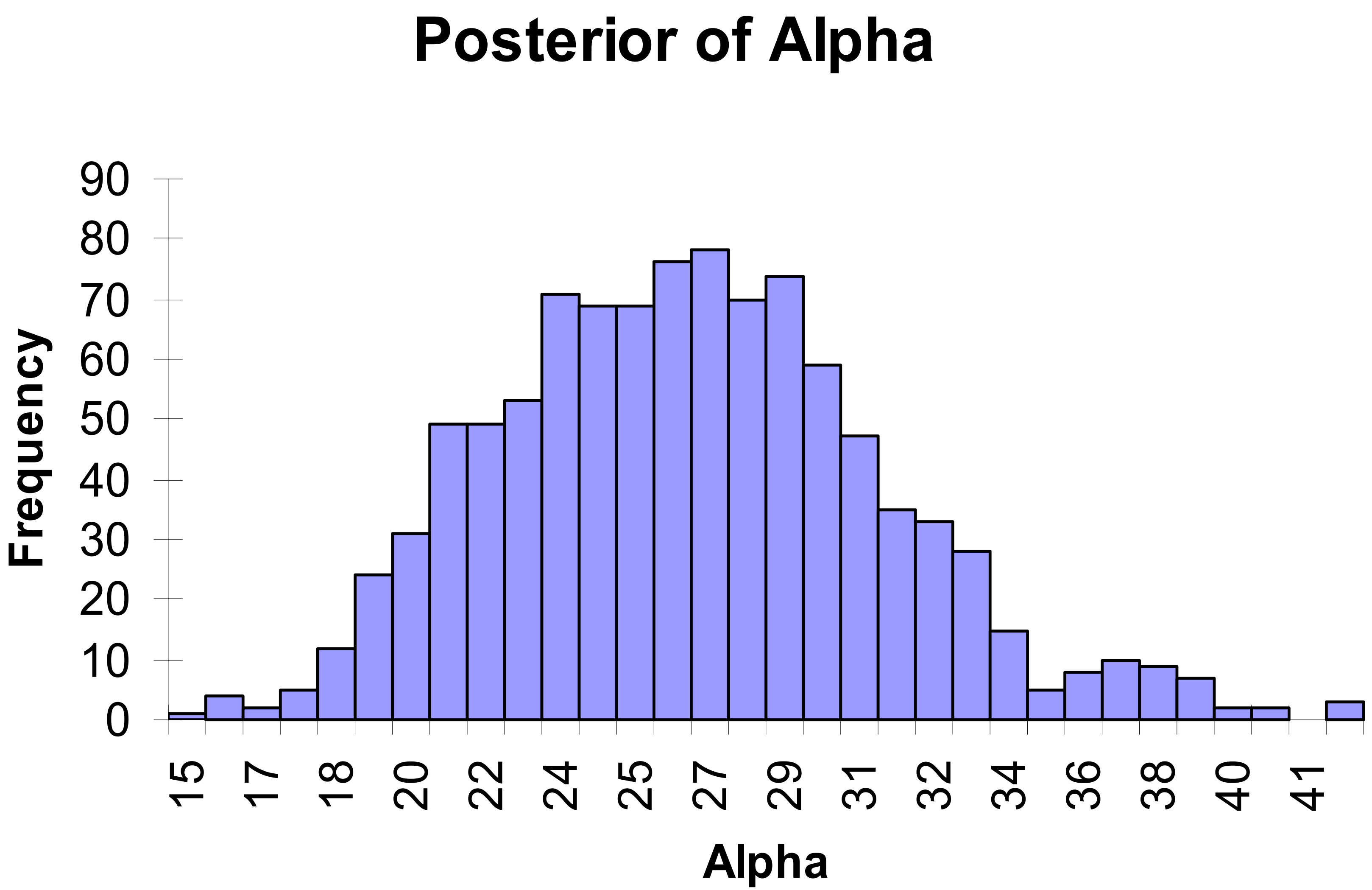
Transition Probability

$$\alpha(\theta,\varphi)=\max\left\{\frac{f(\varphi)g(\theta\mid\varphi)}{f(\theta)g(\varphi\mid\theta)},1\right\}$$

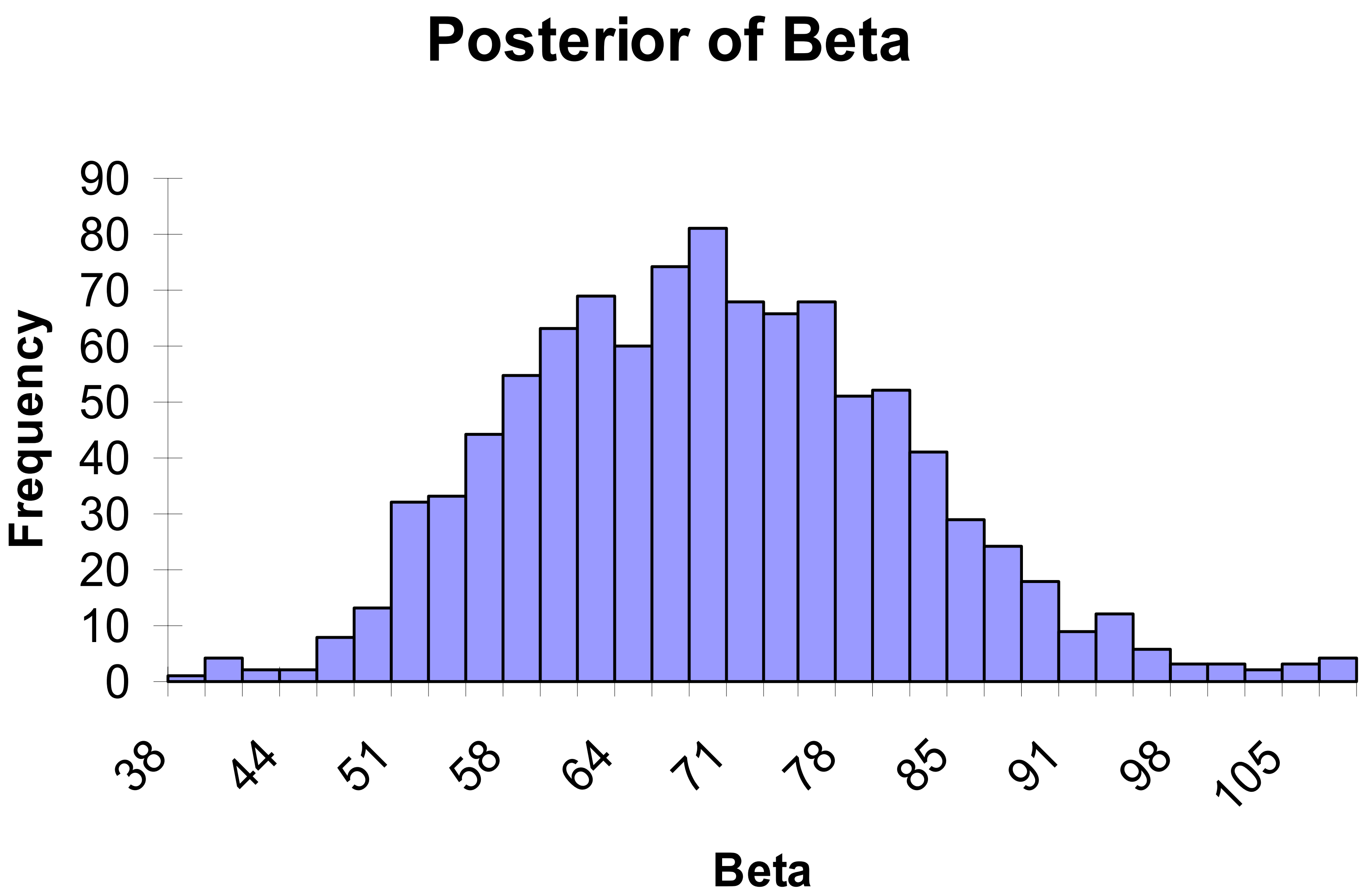
- f is the full conditional density of θ
- Ratios: do not need to know constants
- Usually compute α on log scale.
- Works if densities are not zero
- Works better if g is close to f

Alpha and Beta vs Iteration





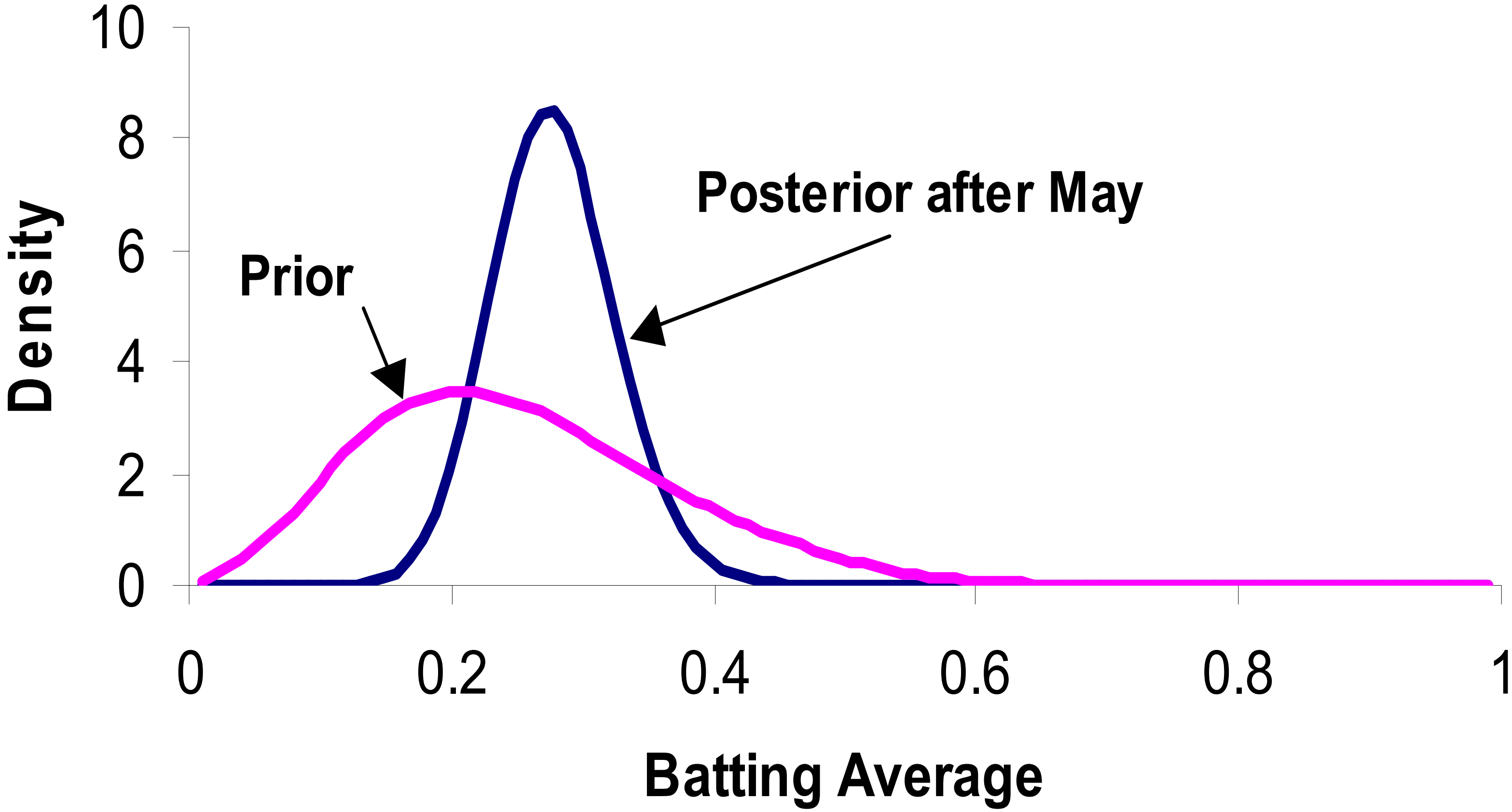
Posterior of Alpha
Frequency
90
80
70
60
50
40
30
20
10
0
15
17
18
20
22
24
25
27
29
31
32
34
36
38
40
41
Alpha


Posterior of Beta
Frequency
0
10
20
30
40
50
60
70
80
90
38
44
51
58
64
71
78
85
91
98
105
Beta

Parameters Estimates

	Prior	Posterior
α	17.8	26.2
(std)	(8.4)	(4.6)
β	53.2	68.2
(std)	(14.6)	(11.7)

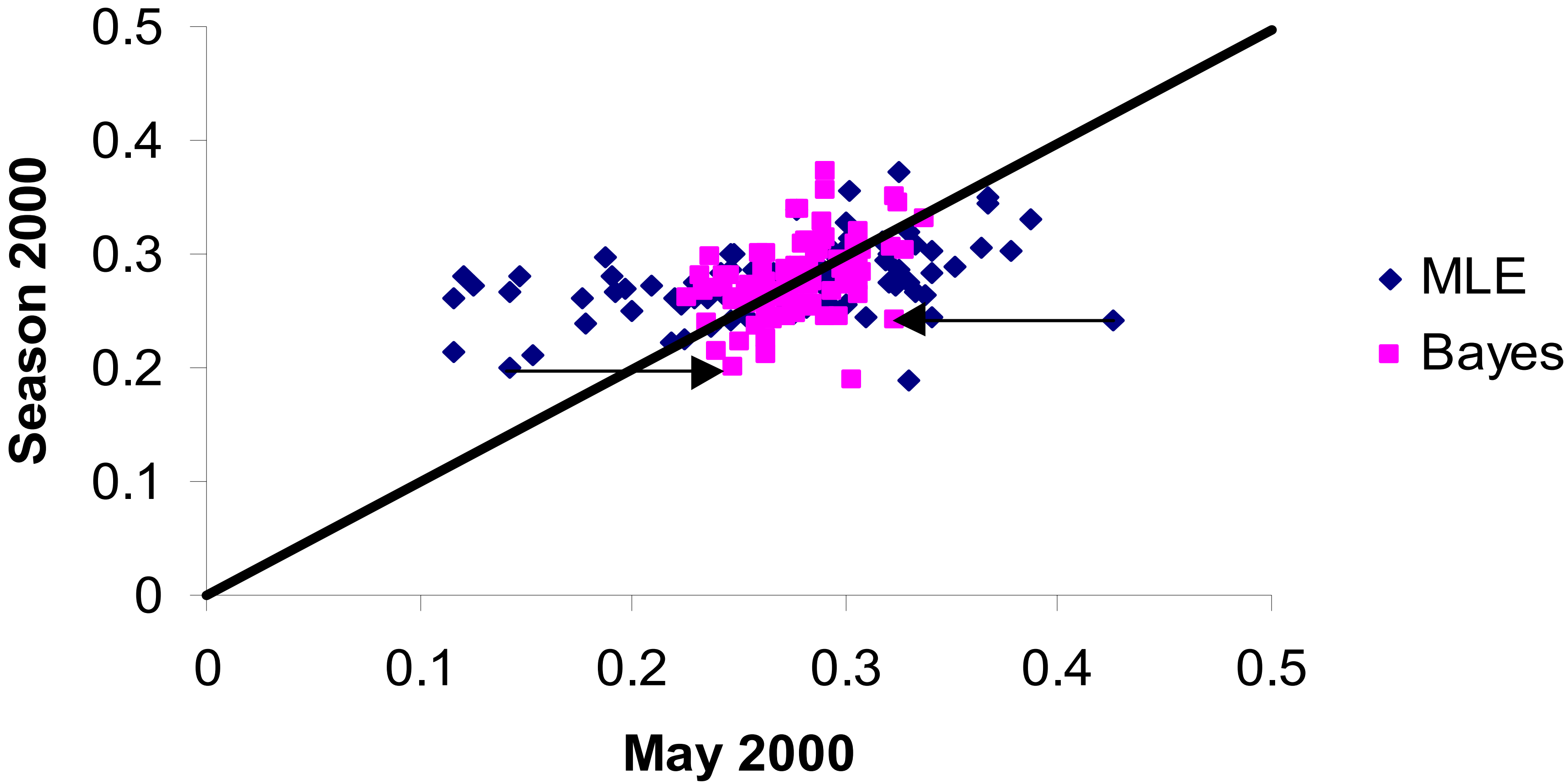
Distribution of Batting Averages



Prediction of Season Averages

	RMSE	MAPE
MLE	0.060	17.0%
Bayes	0.032	9.4%

Batting Averages
Bayes Shrinks MLE



HB Conjoint

Lenk, DeSarbo, Green, Young (1996)

- Evaluated computer profiles on a 0 to 10 scale for “likelihood to purchase”
  - 0 = Would not buy
  - 10 = Would definitely buy
- Design
  - 178 subjects
  - 13 attributes with two levels each
  - 20 profiles per subject

Attributes: Effect Coding

1. Hotline support
2. RAM
3. Screen Size
4. CPU
5. Hard Disk
6. Multimedia
7. Cache
8. Color
9. Retail Store
10. Warrantee
11. Software
12. Guarantee
13. Price

Subject-Covariates

- Female: 1 if female and 0 if male
- Years: # years of work experience
- Own: 1 if has computer & 0 else
- Nerd: 1 if technical background & 0 else
- Apply: # of software applications
- Expert: self-report expertise rating

Summary Statistics for Covariates

Variable	Mean	Std Dev	MIN	MAX
FEMALE	0.275	0.448	0	1
YEARS	4.416	2.369	1	18
OWN	0.876	0.330	0	1
NERD	0.275	0.448	0	1
APPLY	4.287	1.574	1	9
EXPERT	7.618	1.902	3	10

Interaction Model

- Within-Subjects

$$Y_i = X_i \beta_i + \varepsilon_i \text{ for } i = 1, \ldots, n$$

$$\left[\varepsilon_i\right] = N_m\left(\varepsilon_i \mid 0, \sigma^2 I_m\right)$$

- Between-Subjects Heterogeneity

$$\beta_i = \Theta' z_i + \delta_i$$

$$\left[\delta_i\right] = N_p\left(\delta_i \mid 0, \Delta\right)$$

Average over Posterior Means and Std Dev of Partworths Across Subjects

	PostMean	PostSTD		PostMean	PostSTD
CNST	4.757	1.404	Cache	0.031	0.461
HotLine	0.095	0.487	Color	0.026	0.371
RAM	0.347	0.467	Dstrbtn	0.078	0.378
ScrnSz	0.193	0.405	Wrrnty	0.124	0.392
CPU	0.392	0.646	Sftwr	0.196	0.399
HrdDsk	0.171	0.501	Grntee	0.112	0.427
MultMd	0.494	0.574	Price	-1.127	0.873

Impact of Covariates on Partworths

	CNST	RAM	CPU	Dstrbtn	Wrrnty	Grntee	Price
CNST	3.74	0.52	-0.15	0.05	-0.01	0.03	-1.55
FEMALE	-0.10			0.06	0.12		0.40
YEARS	-0.11						
OWN	-0.10	0.17	0.17	0.20		-0.12	-0.20
NERD	-0.27	0.15	0.16	-0.14			
APPLY	0.10						
EXPERT	0.17						

Summary

- HB allows individual-level coefficients
- Two level model
  - With-in subjects
  - Between subjects (heterogeneity)
- HB shrinks unstable, subject-level estimators to population mean

Summary

- BDT provides integrated framework for making decisions and inference
- Good models consider all sources of uncertainty
- Good methods keep track of all sources of uncertainty
- Bayes does both