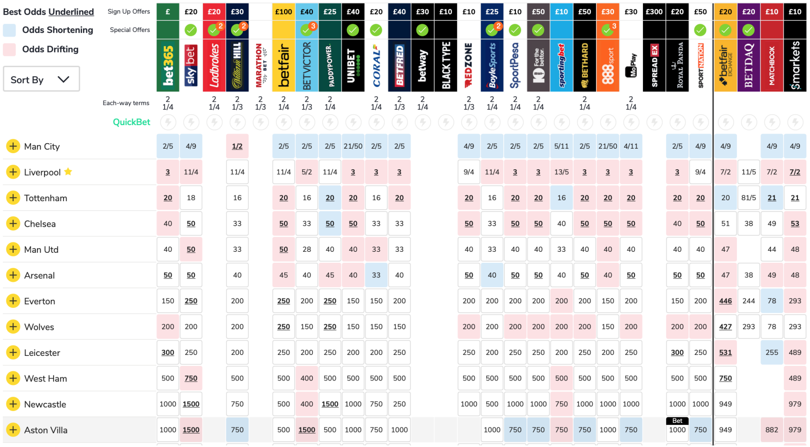
Bayes AI

Unit 3: Bayesian Inference with Conjugate Pairs

Vadim Sokolov George Mason University Spring 2025

EPL Odds



Best Odds Underlined

🟦 **Odds Shortening**

🟥 **Odds Drifting**

Sort By ⌄

	bet365	sky bet	Ladbrokes	William Hill	MARATHON BET	betfair	BETVICTOR	PADDYPOWER	UNIBET	CORAL	BETFRED	betway	BLACK TYPE	REDZONE	BoyleSports	SportPesa	10BET For the bettor.	sportingbet	BETHARD	888sport	MoPlay	SPREAD EX	ROYAL PANDA	SPORTNATION	betfair EXCHANGE	BETDAQ	MATCHBOOK	Smarkets
Sign Up Offers	£	£20	£20	£30		£100	£40	£25	£40	£20	£40	£30	£10	£10	£25	£10	£50	£10	£50	£30	£30	£300	£20	£50	£20	£20	£10	£10
Special Offers		✓	✓ 2	✓ 2			✓ 3		✓	✓		✓			✓ 2	✓	✓			✓ 3				✓	✓	✓		
Each-way terms	2 1/4		2 1/4	2 1/3	2 1/3	2 1/4	2 1/3	2 1/4		2 1/4	2 1/3	2 1/4		2 1/3	2 1/4	2 1/4	2 1/4		2 1/4		2 1/4							
QuickBet	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡	⚡
+ Man City	2/5	4/9		**1/2**		2/5	2/5	2/5	21/50	2/5	2/5			4/9	2/5	2/5	2/5	5/11	2/5	21/50	4/11		2/5	4/9	4/9		4/9	4/9
+ Liverpool ⭐	**3**	11/4		11/4		11/4	5/2	11/4	**3**	**3**	**3**			9/4	11/4	**3**	**3**	13/5	**3**	**3**	**3**		**3**	9/4	7/2	11/5	7/2	**7/2**
+ Tottenham	**20**	18		16		**20**	16	**20**	**20**	16	**20**			**20**	16	**20**	**20**	16	**20**	**20**	**20**		**20**	**20**	20	81/5	**21**	**21**
+ Chelsea	40	**50**		33		**50**	33	**50**	**50**	33	33			**50**	33	**50**	**50**	40	**50**	**50**	**50**		40	**50**	51	38	49	**53**
+ Man Utd	40	**50**		33		**50**	28	40	40	33	33			40	33	40	40	33	40	40	40		40	40	47		44	48
+ Arsenal	**50**	**50**		40		45	40	45	40	33	40			**50**	40	**50**	**50**	40	**50**	40	**50**		**50**	**50**	47	38	49	48
+ Everton	150	**250**		200		**250**	200	**250**	150	150	200			200	200	200	200	200	200	150	200		150	200	**446**	244	78	293
+ Wolves	200	200		200		**250**	150	**250**	150	150	150			200	200	200	200	200	200	150	200		200	200	**427**	293	78	293
+ Leicester	**300**	250		200		250	200	250	200	200	250			250	200	250	250	200	250	200	250		**300**	250	**531**		255	489
+ West Ham	500	**750**		500		500	400	500	500	500	500			500	500	500	500	500	500	500	500		500	500	**750**			489
+ Newcastle	1000	**1500**		750		500	400	**1500**	1000	750	250			1000	500	1000	1000	750	1000	1000	1000		1000	1000	949			979
+ Aston Villa	1000	**1500**		750		500	**1500**	500	1000	750	1000				1000	750	750	750	750	1000	750		Bet 1000	750	949		882	979

English Premier League: EPL

Calculate Odds for the possible scores in a match?

$$0-0,\ 1-0,\ 0-1,\ 1-1,\ 2-0,\ldots$$

Let $X =$ Goals scored by Arsenal

$Y =$ Goals scored by Liverpool

What's the odds of a team winning? $P\left(X > Y\right)$ Odds of a draw? $P\left(X = Y\right)$

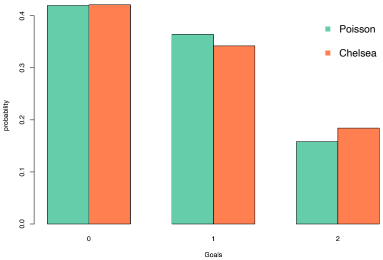
z1 = rpois(100,0.6) z2 = rpois(100,1.4) sum(z1<z2)/100 # Team 2 wins sum(z1=z2)/100 # Draw

Chelsea EPL 2017

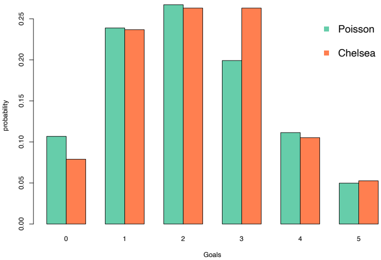
Let’s take a historical set of data on scores Then estimate λ with the sample mean of the home and away scores

home team	results		visit team
Chelsea	2	1	West Ham
Chelsea	5	1	Sunderland
Watford	1	2	Chelsea
Chelsea	3	0	Burnley
. . .			

EPL Chelsea



(a) Chelsea against



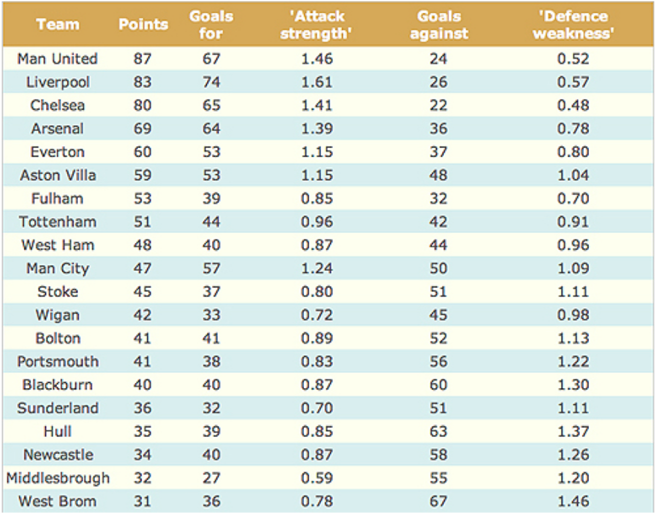
(a) Chelsea for

Our Poisson model fits the empirical data!!

EPL: Attack and Defence Strength

Each team gets an “attack” strength and “defence” weakness rating

Adjust home and away average goal estimates



Team	Points	Goals for	'Attack strength'	Goals against	'Defence weakness'
Man United	87	67	1.46	24	0.52
Liverpool	83	74	1.61	26	0.57
Chelsea	80	65	1.41	22	0.48
Arsenal	69	64	1.39	36	0.78
Everton	60	53	1.15	37	0.80
Aston Villa	59	53	1.15	48	1.04
Fulham	53	39	0.85	32	0.70
Tottenham	51	44	0.96	42	0.91
West Ham	48	40	0.87	44	0.96
Man City	47	57	1.24	50	1.09
Stoke	45	37	0.80	51	1.11
Wigan	42	33	0.72	45	0.98
Bolton	41	41	0.89	52	1.13
Portsmouth	41	38	0.83	56	1.22
Blackburn	40	40	0.87	60	1.30
Sunderland	36	32	0.70	51	1.11
Hull	35	39	0.85	63	1.37
Newcastle	34	40	0.87	58	1.26
Middlesbrough	32	27	0.59	55	1.20
West Brom	31	36	0.78	67	1.46

EPL: Hull vs ManU

Poisson Distribution

ManU Average away goals = 1.47. Prediction:
$1.47 \times 1.46 \times 1.37 = 2.95$
Attack strength times Hull's defense weakness times average
Hull Average home goals = 1.47. Prediction:
$1.47 \times 0.85 \times 0.52 = 0.65$. Simulation

Team Ex	pected Goals	0	1	2	3	4	5
Man U	2.95	7	22	26	12	11	13
Hull City	0.65	49	41	10	0	0	0

EPL Predictions

A model is only as good as its predictions

- In our simulation Man U wins 88 games out of 100, we should bet when odds ratio is below 88 to 100.
- Most likely outcome is 0-3 (12 games out of 100)
- The actual outcome was 0-1 (they played on August 27, 2016)
- In out simulation 0-1 was the fourth most probable outcome (9 games out of 100)

Hierarchical Distributions

Bayesian Methods

Modern Statistical/Machine Learning

- Bayes Rule and Probabilistic Learning
- Computationally challenging: MCMC and Particle Filtering
- Many applications in Finance:

Asset pricing and corporate finance problems.

Lindley, D.V. *Making Decisions*

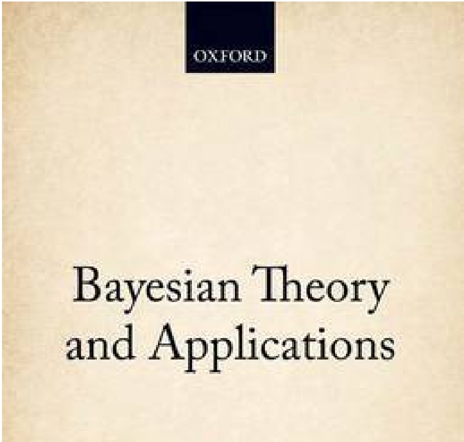
Bernardo, J. and A.F.M. Smith *Bayesian Theory*

Bayesian Books

- Hierarchical Models and MCMC
- Bayesian Nonparametrics

Machine Learning

- Dynamic State Space Models . . .



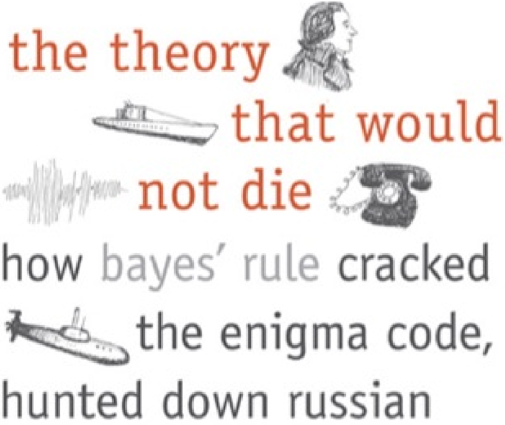
OXFORD

Bayesian Theory and Applications

Popular Books

McGrayne (2012): The Theory that would not Die

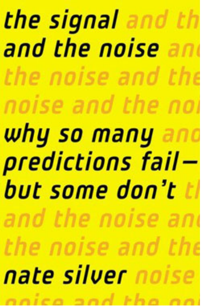
- History of Bayes-Laplace
- Code breaking
- Bayes search: Air France . . .



Nate Silver: 538 and NYT

Silver (2012): The Signal and The Noise

- Presidential Elections
- Bayes dominant methodology
- Predicting College Basketball/Oscars . . .



the signal and the noise
why so many predictions fail— but some don't
nate silver

Things to Know

Explosion of Models and Algorithms starting in 1950s

- Bayesian Regularisation and Sparsity
- Hierarchical Models and Shrinkage
- Hidden Markov Models
- Nonlinear Non-Gaussian State Space Models

Algorithms

- Monte Carlo Method (von Neumann and Ulam, 1940s)
- Metropolis-Hastings (Metropolis, 1950s)
- Gibbs Sampling (Geman and Geman, Gelfand and Smith, 1980s)
- Sequential Particle Filtering

Probabilistic Reasoning

- Bayesian Probability (Ramsey, 1926, de Finetti, 1931)
  1. Beta-Binomial Learning: Black Swans
  2. Elections: Nate Silver
  3. Baseball: Kenny Lofton and Derek Jeter
- Monte Carlo (von Neumann and Ulam, Metropolis, 1940s)
- Shrinkage Estimation (Lindley and Smith, Efron and Morris, 1970s)

Bayesian Inference

Key Idea: Explicit use of probability for summarizing uncertainty.

1. A probability distribution for data given parameters

$$f(y|\theta) \text{ Likelihood}$$

2. A probability distribution for unknown parameters

$$p(\theta) \text{ Prior}$$

3. Inference for unknowns conditional on observed data

Inverse probability (Bayes Theorem);

Formal decision making (Loss, Utility)

Posterior Inference

Bayes theorem to derive posterior distributions

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$
$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

Allows you to make probability statements

- They can be very different from p-values!

Hypothesis testing and Sequential problems

- Markov chain Monte Carlo (MCMC) and Filtering (PF)

Conjugate Priors

- Definition: Let F denote the class of distributions $f(y|\theta)$.

A class Π of prior distributions is conjugate for F if the posterior distribution is in the class Π for all $f \in F, \pi \in \Pi, y \in Y$.

- *Example: Binomial/Beta:*

Suppose that $Y_1, \ldots, Y_n \sim Ber(p)$.

Let $p \sim Beta(\alpha, \beta)$ where (α, β) are known hyper-parameters.

The beta-family is very flexible

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

Bayes Learning: Beta-Binomial

How do I update my beliefs about a coin toss?

Likelihood for Bernoulli

$$p\left(y|\theta\right)=\prod_{t=1}^{T}p\left(y_t|\theta\right)=\theta^{\sum_{t=1}^{T}y_t}\left(1-\theta\right)^{T-\sum_{t=1}^{T}y_t}.$$

Initial prior distribution $\theta\sim\mathcal{B}\left(a,A\right)$ given by

$$p\left(\theta|a,A\right)=\frac{\theta^{a-1}\left(1-\theta\right)^{A-1}}{B\left(a,A\right)}$$

Bayes Learning: Beta-Binomial

Updated posterior distribution is also Beta

$$p\left(\theta|y\right) \sim \mathcal{B}\left(a_T, A_T\right) \text{ and } a_T = a + \sum_{t=1}^{T} y_t, A_T = A + T - \sum_{t=1}^{T} y_t$$

The posterior mean and variance are

$$E\left[\theta|y\right] = \frac{a_T}{a_T + A_T} \text{ and } var\left[\theta|y\right] = \frac{a_T A_T}{\left(a_T + A_T\right)^2 \left(a_T + A_T + 1\right)}$$

Binomial-Beta

$p(p|\bar{y})$ is the posterior distribution for p

$\bar{y}$ is a sufficient statistic.

- Bayes theorem gives

$$\begin{aligned} p(p|y) &\propto f(y|p)p(p|\alpha,\beta) \\ &\propto p^{\sum y_i}(1-p)^{n-\sum y_i} \cdot p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{\alpha+\sum y_i-1}(1-p)^{n-\sum y_i+\beta-1} \\ &\sim Beta(\alpha+\sum y_i, \beta+n-\sum y_i) \end{aligned}$$

- The posterior mean is a shrinkage estimator

Combination of sample mean $\bar{y}$ and prior mean $E(p)$

$$E(p|y) = \frac{\alpha+\sum_{i=1}^{n} y_i}{\alpha+\beta+n} = \frac{n}{n+\alpha+\beta}\bar{y} + \frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta}$$

Black Swans

Taleb, The Black Swan: the Impact of the Highly Improbable

Suppose you're only see a sequence of White Swans, having never seen a Black Swan.

What's the Probability of Black Swan event *sometime* in the future?

Suppose that after T trials you have only seen successes $(y_1, \ldots, y_T) = (1, \ldots, 1)$. The next trial being a success has

$$p(y_{T+1} = 1 | y_1, \ldots, y_T) = \frac{T+1}{T+2}$$

For large T is almost certain. Here $a = A = 1$.

Black Swans

Principle of Induction (Hume)

The probability of never seeing a Black Swan is given by

$$p(y_{T+1} = 1, \ldots, y_{T+n} = 1 | y_1, \ldots, y_T) = \frac{T+1}{T+n+1} \to 0$$

Black Swan will eventually happen – don't be surprised when it actually happens.

Bayesian Learning: Poisson-Gamma

Poisson/Gamma: Suppose that $Y_1, \ldots, Y_n \mid \lambda \sim Poi(\lambda)$.

Let $\lambda \sim Gamma(\alpha, \beta)$

(α, β) are known hyper-parameters.

- The posterior distribution is

$$
\begin{aligned}
p(\lambda|y) &\propto \exp(-n\lambda)\lambda^{\sum y_i}\lambda^{\alpha-1}\exp(-\beta\lambda) \\
&\sim Gamma(\alpha + \sum y_i, n + \beta)
\end{aligned}
$$

Example: Clinical Trials

Novick and Grizzle: Bayesian Analysis of Clinical Trials

Four treatments for duodenal ulcers.

Doctors assess the state of the patient.

Sequential data

(α-spending function, can only look at prespecified times).

Treat	Excellent	Fair	Death
A	76 1	7	7
B	89 1	0	1
C	86 1	3	1
D	88 9		3

Conclusion: Cannot reject at the 5% level

Conjugate binomial/beta model+sensitivity analysis.

Binomial-Beta

Let p_i be the death rate proportion under treatment i.

- To compare treatment A to B directly compute $P(p_1 > p_2|D)$.
- Prior $beta(\alpha, \beta)$ with prior mean $E(p) = \frac{\alpha}{\alpha+\beta}$.

Posterior $beta(\alpha + \sum x_i, \beta + n - \sum x_i)$

- For A, $beta(1,1) \rightarrow beta(8,94)$

For B, $beta(1,1) \rightarrow beta(2,100)$

- Inference: $P(p_1 > p_2|D) \approx 0.98$

Sensitivity Analysis

Important to do a sensitivity analysis.

Treat	Excellent	Fair	Death
A	76 1	7	7
B	89 1	0	1
C	86 1	3	1
D	88 9		3

Poisson-Gamma, prior $\Gamma(m, z)$ and λ_i be the expected death rate.

Compute $P\left(\frac{\lambda_1}{\lambda_2} > c|D\right)$

Prob	(0 , 0)	(100, 2)	(200 , 5)
$P\left(\frac{\lambda_1}{\lambda_2} > 1.3|D\right)$	0.95	0.88	0.79
$P\left(\frac{\lambda_1}{\lambda_2} > 1.6|D\right)$	0.91	0.80	0.64

Bayesian Learning: Normal-Normal

Using Bayes rule we get

$$p(\mu|y) \propto p(y|\mu)p(\mu)$$

- Posterior is given by

$$p(\mu|y) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 - \frac{1}{2\tau^2}(\mu-\mu_0)^2\right)$$

Hence $\mu|y \sim N\left(\hat{\mu}_B, V_\mu\right)$ where

$$\hat{\mu}_B = \frac{n/\sigma^2}{n/\sigma^2+1/\tau^2}\bar{y} + \frac{1/\tau^2}{n/\sigma^2+1/\tau^2}\mu_0 \text{ and } V_\mu^{-1} = \frac{n}{\sigma^2}+\frac{1}{\tau^2}$$

A shrinkage estimator.

SAT Scores

SAT $(200 - 800)$: 8 high schools and estimate effects.

School	Estimated y_j	St. Error σ_j	Average Treatment θ_i
A	28	15	?
B	8	10	?
C	-3	16	?
D	7	11	?
E	-1	9	?
F	1	11	?
G	18	10	?
H	12	18	?

- θ_j average effects of coaching programs
- y_j estimated treatment effects, for school j, standard error σ_j.

Estimates

Two programs appear to work (improvements of 18 and 28)

- Large standard errors. Overlapping Confidence Intervals?
- Classical hypothesis test fails to reject the hypothesis that the θ_j's are equal.
- Pooled estimate has standard error of 4.2 with

$$\hat{\theta} = \frac{\sum_j (y_j/\sigma_j^2)}{\sum_j (1/\sigma_j^2)} = 7.9$$

- Neither separate or pooled seems sensible.

Bayesian shrinkage!

Hierarchical Model

Hierarchical Model (σ_j^2 known) is given by

$$\bar{y}_j | \theta_j \sim N(\theta_j, \sigma_j^2)$$

Unequal variances–differential shrinkage.

- Prior Distribution: $\theta_j \sim N(\mu, \tau^2)$ for $1 \leq j \leq 8$.

Traditional random effects model.

Exchangeable prior for the treatment effects.

As $\tau \to 0$ (complete pooling) and as $\tau \to \infty$ (separate estimates).

- Hyper-prior Distribution: $p(\mu, \tau^2) \propto 1/\tau$.

The posterior $p(\mu, \tau^2 | y)$ can be used to "estimate" (μ, τ^2).

Posterior

Joint Posterior Distribution $y = (y_1, \ldots, y_J)$

$$p(\theta, \mu, \tau | y) \propto p(y|\theta)p(\theta|\mu, \tau)p(\mu, \tau)$$

$$\propto p(\mu, \tau^2) \prod_{i=1}^{8} N(\theta_j|\mu, \tau^2) \prod_{j=1}^{8} N(y_j|\theta_j)$$

$$\propto \tau^{-9} \exp\left(-\frac{1}{2} \sum_j \frac{1}{\tau^2}(\theta_j - \mu)^2 - \frac{1}{2} \sum_j \frac{1}{\sigma_j^2}(y_j - \theta_j)^2 \right)$$

MCMC!

Posterior Inference

Report posterior quantiles

School	2.5%	25%	50%	75%	97.5%
A -	2 6	10	16	3	2
B -	5 4	8	12	2	0
C -1	2 3	7	11	2	2
D -	6 4	8	12	2	1
E -1	0 2	6	10	1	9
F -	9 2	6	10	1	9
G -	1 6	10	15	2	7
H -	7 4	8	13	2	3
μ	-2	5	8	11	18
τ	0.3	2.3	5.1	8.8	21

Schools *A* and *G* are similar!

Bayesian Shrinkage

Bayesian shrinkage provides a way of modeling complex datasets.

1. Baseball batting averages: Stein's Paradox
2. Batter-pitcher match-up: Kenny Lofton and Derek Jeter
3. Bayes Elections
4. Toxoplasmosis
5. Bayes MoneyBall
6. Bayes Portfolio Selection

Example: Baseball

Batter-pitcher match-up?

Prior information on overall ability of a player.

Small sample size, pitcher variation.

- Let p_i denote Jeter's ability. Observed number of hits y_i

$$(y_i|p_i) \sim Bin(T_i, p_i) \text{ with } p_i \sim Be(\alpha, \beta)$$

where T_i is the number of at-bats against pitcher i. A priori $E(p_i) = \alpha/(\alpha+\beta) = \bar{p}_i$.

- The extra heterogeneity leads to a prior variance $Var(p_i) = \bar{p}_i(1-\bar{p}_i)\phi$ where $\phi = (\alpha+\beta+1)^{-1}$.

Sports Data: Baseball

Kenny Lofton hitting versus individual pitchers.

Pitcher At	-bats Hi	ts Ob	sAvg
J.C. Romero	9	6	.667
S. Lewis	5 3	.	600
B. Tomko	20 1	1 .	550
T. Hoffman	6	3	.500
K. Tapani	45	22	.489
A. Cook	9 4	.	444
J. Abbott	34	14	.412
A.J. Burnett	15	6	.400
K. Rogers	43	17	.395
A. Harang	6	2	.333
K. Appier	49	15	.306
R. Clemens	62	14	.226
C. Zambrano	9	2	.222
N. Ryan	10 2	.	200
E. Hanson	41	7	.171

Baseball

Kenny Lofton

Kenny Lofton (career .299 average, and current .308 average for 2006 season) was facing the pitcher Milton (current record 1 for 19)

.

- Is putting in a weaker player really a better bet?
- Over-reaction to bad luck?

$\mathbb{P}\left(\leq 1 \text{ hit in } 19 \text{ attempts} | p = 0.3\right) = 0.01$

An unlikely 1-in-100 event.

Baseball

Kenny Lofton

Bayes solution: shrinkage. Borrow strength across pitchers

Bayes estimate: use the posterior mean

Lofton's batting estimates that vary from .265 to .340.

The lowest being against Milton.

$.265 < .275$

Conclusion: resting Lofton against Milton was justified!!

Bayes Batter-pitcher match-up

Here's our model again ...

- Small sample sizes and pitcher variation.
- Let p_i denote Lofton's ability. Observed number of hits y_i

$$(y_i|p_i) \sim Bin(T_i, p_i) \;\text{ with }\; p_i \sim Be(\alpha, \beta)$$

where T_i is the number of at-bats against pitcher i.

Estimate (α, β)

Example: Derek Jeter

Derek Jeter 2006 season versus individual pitchers.

Pitcher	At-bats	Hits	ObsAvg	EstAvg	95% Int
R. Mendoza	6	5	.833	.322	(.282, .394)
H. Nomo	20	12	.600	.326	(.289, .407)
A.J.Burnett	5	3	.600	.320	(.275, .381)
E. Milton	28	14	.500	.324	(.291, .397)
D. Cone	8	4	.500	.320	(.218, .381)
R. Lopez	45	21	.467	.326	(.291, .401)
K. Escobar	39	16	.410	.322	(.281, .386)
J. Wettland	5	2	.400	.318	(.275, .375)
T. Wakefield	81	26	.321	.318	(.279, .364)
P. Martinez	83	21	.253	.312	(.254, .347)
K. Benson	8	2	.250	.317	(.264, .368)
T. Hudson	24	6	.250	.315	(.260, .362)
J. Smoltz	5	1	.200	.314	(.253, .355)
F. Garcia	25	5	.200	.314	(.253, .355)
B. Radke	41	8	.195	.311	(.247, .347)

Bayes Estimates

Stern stimates $\hat{\phi} = (\alpha + \beta + 1)^{-1} = 0.002$ for Jeter

Doesn't vary much across the population of pitchers.

The extremes are shrunk the most also matchups with the smallest sample sizes.

Jeter had a season .308 average.

Bayes estimates vary from .311 to .327–he's very consistent.

If all players had a similar record then a constant batting average would make sense.

Bayes Elections: Nate Silver

Multinomial-Dirichlet

Predicting the Electoral Vote (EV)

- Multinomial-Dirichlet: $(\hat{p}|p) \sim Multi(p), (p|\alpha) \sim Dir(\alpha)$

$$p_{Obama} = (p_1, \ldots, p_{51}|\hat{p}) \sim Dir\left(\alpha + \hat{p}\right)$$

- Flat uninformative prior $\alpha \equiv 1$.

`http://www.electoral-vote.com/evp2012/Pres/prespolls.csv`

Bayes Elections: Nate Silver

Simulation

Calculate probabilities via simulation: `rdirichlet`

$$p\left(p_{j,O}|\mathrm{data}\right) \ \ \mathrm{and} \ \ p\left(EV > 270|\mathrm{data}\right)$$

The election vote prediction is given by the sum

$$EV = \sum_{j=1}^{51} EV(j)\mathbb{E}\left(p_j|\mathrm{data}\right)$$

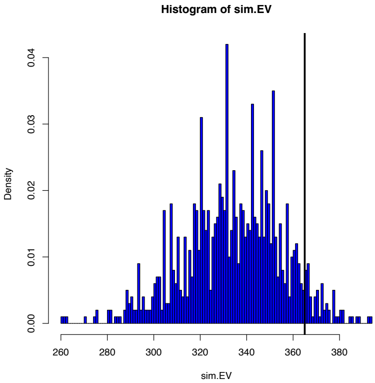
where $EV(j)$ are for individual states

Polling Data: electoral-vote.com

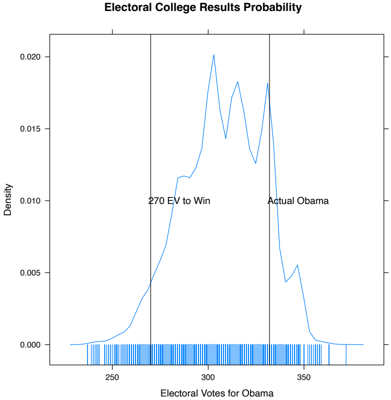
Electoral Vote (EV), Polling Data: Mitt and Obama percentages

State	M.pct	O.pct	EV
Alabama	58	36	9
Alaska	55	37	3
Arizona	50	46	10
Arkansas	51	44	6
California	33	55	55
Colorado	45	52	9
Connecticut	31	56	7
Delaware	38	56	3
D.C.	13	82	3
Florida	46	50	27
Georgia	52	47	15
Hawaii	32	63	4
Idaho	68	26	4
Illinois	35	59	21
Indiana	48	48	11

Polling Data:



(a) Election 2008 Prediction. Obama 370



(a) Election 2012 Prediction. Obama 332.

Chicago Bears 2014-2015 Season

Bayes Learning: Update our beliefs in light of new information

- In the 2014-2015 season.

The Bears suffered back-to-back 50-points defeats.

Partiots-Bears $51-23$

Packers-Bears $55-14$

- Their next game was at home against the Minnesota Vikings.

Current line against the Vikings was -3.5 points.

Slightly over a field goal

What's the Bayes approach to learning the line?

Hierarchical Model

Hierarchical model for the current average win/lose this year

$$\bar{y}|\theta \sim N\left(\theta, \frac{\sigma^2}{n}\right) \sim N\left(\theta, \frac{18.34^2}{9}\right)$$
$$\theta \sim N(0, \tau^2)$$

Here $n = 9$ games so far. With $s = 18.34$ points

Pre-season prior mean $\mu_0 = 0$, standard deviation $\tau = 4$.

Record so-far. Data $\bar{y} = -9.22$.

Chicago Bears

Bayes Shrinkage estimator

$$\mathbb{E}\left(\theta|\bar{y},\tau\right) = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}}\bar{y}$$

The Shrinkage factor is 0.3!!

That's quite a bit of shrinkage. Why?

- Our updated estimator is

$$\mathbb{E}\left(\theta|\bar{y},\tau\right) = -2.75 > -.3.5$$

where current line is -3.5.

- Based on our hierarchical model this is an over-reaction.

One point change on the line is about 3% on a probability scale.

Alternatively, calculate a market-based τ given line $= -3.5$.

Chicago Bears

Last two defeats were 50 points scored by opponent (2014-15)

```
[1] -9.222222

[1] 18.34242

[1] 0.2997225

[1] 0.4390677
```

Home advantage is worth 3 points. Vikings an average record.

Result: Bears 21, Vikings 13

Stein's Paradox

Stein paradox: possible to make a uniform improvement on the MLE in terms of MSE.

- Mistrust of the statistical interpretation of Stein's result.

In particular, the loss function.

- Difficulties in adapting the procedure to special cases
- Long familiarity with good properties for the MLE

Any gains from a "complicated" procedure could not be worth the extra trouble (Tukey, savings not more than 10 % in practice)

For $k \geq 3$, we have the remarkable inequality

$$MSE(\hat{\theta}_{JS}, \theta) < MSE(\bar{y}, \theta) \; \forall \theta$$

Bias-variance explanation! Inadmissability of the classical stats.

Baseball Batting Averages

Data: 18 major-league players after 45 at bats (1970 season)

Player	$\bar{y}_i$	$E(p_i \vert D)$	average season
Clemente	0.400	0.290	0.346
Robinson	0.378	0.286	0.298
Howard	0.356	0.281	0.276
Johnstone	0.333	0.277	0.222
Berry	0.311	0.273	0.273
Spencer	0.311	0.273	0.270
Kessinger	0.311	0.268	0.263
Alvarado	0.267	0.264	0.210
Santo	0.244	0.259	0.269
Swoboda	0.244	0.259	0.230
Unser	0.222	0.254	0.264
Williams	0.222	0.254	0.256
Scott	0.222	0.254	0.303
Petrocelli	0.222	0.254	0.264
Rodriguez	0.222	0.254	0.226

Baseball Data

First Shrinkage Estimator: Efron and Morris

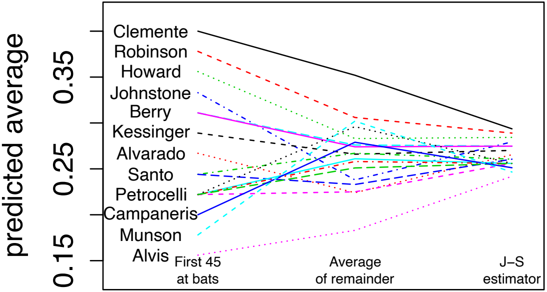


Figure 5: Baseball Shrinkage

Shrinkage

Let θ_i denote the end of season average

- Lindley: shrink to the overall grand mean

$$c = 1 - \frac{(k-3)\sigma^2}{\sum(\bar{y}_i - \bar{y})^2}$$

where $\bar{y}$ is the overall grand mean and

$$\hat{\theta} = c\bar{y}_i + (1-c)\bar{y}$$

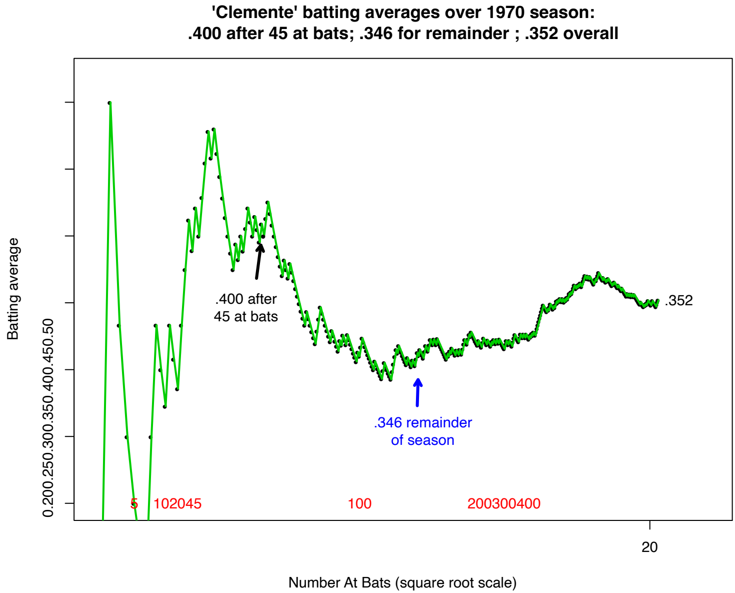
- Baseball data: $c = 0.212$ and $\bar{y} = 0.265$.

Compute $\sum(\hat{\theta}_i - \bar{y}_i^{obs})^2$ and see which is lower:

$$MLE = 0.077 \quad STEIN = 0.022$$

That's a factor of 3.5 times better!

Batting Averages



Baseball Paradoxes

Shrinkage on Clemente too severe:
$z_{Cl} = 0.265 + 0.212(0.400 - 0.265) = 0.294$.

The 0.212 seems a little severe

- Limited translation rules, maximum shrinkage eg. 80%
- Not enough shrinkage eg O'Connor ($y = 1, n = 2$).
  $z_{O'C} = 0.265 + 0.212(0.5 - 0.265) = 0.421$.

Still better than Ted Williams 0.406 in 1941.

- Foreign car sales ($k = 19$) will further improve MSE performance! It will change the shrinkage factors.
- Clearly an improvement over the Stein estimator is

$$\hat{\theta}_{S+} = \max\left(\left(1 - \frac{k-2}{\sum \bar{Y}_i^2}\right), 0\right) \bar{Y}_i$$

Baseball Prior

Include extra prior knowledge

Empirical distribution of all major league players

$$\theta_i \sim N(0.270, 0.015)$$

The 0.270 provides another origin to shrink to and the prior variance 0.015 would give a different shrinkage factor.

To fully understand maybe we should build a probabilistic model and see what the posterior mean is as our estimator for the unknown parameters.

Shrinkage: Unequal Variances

Model $Y_i|\theta_i \sim N(\theta_i, D_i)$ where $\theta_i \sim N(\theta_0, A) \sim N(0.270, 0.015)$.

- The D_i can be different – unequal variances
- Bayes posterior means are given by

$$E(\theta_i|Y) = (1 - B_i)Y_i \;\text{ where }\; B_i = \frac{D_i}{D_i + A}$$

where $\hat{A}$ is estimated from the data, see Efron and Morris (1975).

- Different shrinkage factors as different variances D_i.

$D_i \propto n_i^{-1}$ and so smaller sample sizes are shrunk more.

Makes sense.

Example: Toxoplasmosis Data

Disease of Blood that is endemic in tropical regions.

Data: 5000 people in El Salvador (varying sample sizes) from 36 cities.

- Estimate “true” prevalences θ_i for $1 \leq i \leq 36$
- Allocation of Resources: should we spend funds on the city with the highest observed occurrence of the disease? Same shrinkage factors?
- Shrinkage Procedure (Efron and Morris, p315)

$$z_i = c_i y_i$$

where y_i are the observed relative rates (normalized so $\bar{y} = 0$
The smaller sample sizes will get shrunk more.

The most gentle are in the range $0.6 \to 0.9$ but some are $0.1 \to 0.3$.

Bayes Portfolio Selection

de Finetti and Markowitz: Mean-variance portfolio shrinkage: $\frac{1}{\gamma}\Sigma^{-1}\mu$

Different shrinkage factors for different history lengths.

Portfolio Allocation in the SP500 index

Entry/exit; splits; spin-offs etc. For example, 73 replacements to the SP500 index in period 1/1/94 to 12/31/96.

Advantage: $E(\alpha|D_t) = 0.39$, that is 39 bps per month which on an annual basis is $\alpha = 468$ bps.

The posterior mean for β is $p(\beta|D_t) = 0.745$

$\bar{x}_M = 12.25\%$ and $\bar{x}_{PT} = 14.05\%$.

SP Composition

Date	Symbol	6/96	12/89	12/79	12/69
General Electric	GE	2.800	2.485	1.640	1.569
Coca Cola	KO	2.342	1.126	0.606	1.051
Exxon	XON	2.142	2.672	3.439	2.957
ATT	T	2.030	2.090	5.197	5.948
Philip Morris	MO	1.678	1.649	0.637	*****
Royal Dutch	RD	1.636	1.774	1.191	*****
Merck	MRK	1.615	1.308	0.773	0.906
Microsoft	MSFT	1.436	*****	*****	*****
Johnson/Johnson	JNJ	1.320	0.845	0.689	*****
Intel	INTC	1.262	*****	*****	*****
Procter and Gamble	PG	1.228	1.040	0.871	0.993
Walmart	WMT	1.208	1.084	*****	*****
IBM	IBM	1.181	2.327	5.341	9.231
Hewlett Packard	HWP	1.105	0.477	0.497	*****
Pepsi	PEP	1.061	0.719	*****	*****

SP Composition

Date	Symbol	6/96	12/89	12/79	12/69
Pfizer	PFE	0.918	0.491	0.408	0.486
Dupont	DD	0.910	1.229	0.837	1.101
AIG	AIG	0.910	0.723	*****	*****
Mobil	MOB	0.906	1.093	1.659	1.040
Bristol Myers Squibb	BMY	0.878	1.247	*****	0.484
GTE	GTE	0.849	0.975	0.593	0.705
General Motors	GM	0.848	1.086	2.079	4.399
Disney	DIS	0.839	0.644	*****	*****
Citicorp	CCI	0.831	0.400	0.418	*****
BellSouth	BLS	0.822	1.190	*****	*****
Motorola	MOT	0.804	*****	*****	*****
Ford	F	0.798	0.883	0.485	0.640
Chervon	CHV	0.794	0.990	1.370	0.966
Amoco	AN	0.733	1.198	1.673	0.758
Eli Lilly	LLY	0.720	0.814	*****	*****

SP Composition

Date	Symbol	6/96	12/89	12/79	12/69
Abbott Labs	ABT	0.690	0.654	*****	*****
AmerHome Products	AHP	0.686	0.716	0.606	0.793
FedNatlMortgage	FNM	0.686	*****	*****	*****
McDonald's	MCD	0.686	0.545	*****	*****
Ameritech	AIT	0.639	0.782	*****	*****
Cisco Systems	CSCO	0.633	*****	*****	*****
CMB	CMB	0.621	*****	*****	*****
SBC	SBC	0.612	0.819	*****	*****
Boeing	BA	0.598	0.584	0.462	*****
MMM	MMM	0.581	0.762	0.838	1.331
BankAmerica	BAC	0.560	*****	0.577	*****
Bell Atlantic	BEL	0.556	0.946	*****	*****
Gillette	G	0.535	*****	*****	*****
Kodak	EK	0.524	0.570	1.106	*****
Chrysler	C	0.507	*****	*****	0.367
Home Depot	HD	0.497	*****	*****	*****
Colgate	COL	0.489	0.499	*****	*****

Keynes versus Buffett: CAPM

$$\text{keynes} = 15.08 + 1.83\ \text{market}$$

$$\text{buffett} = 18.06 + 0.486\ \text{market}$$

Year	Keynes	Market
1928	-3.4	7.9
1929	0.8	6.6
1930	-32.4	-20.3
1931	-24.6	-25.0
1932	44.8	-5.8
1933	35.1	21.5
1934	33.1	-0.7
1935	44.3	5.3
1936	56.0	10.2
1937	8.5	-0.5
1938	-40.1	-16.1
1939	12.9	-7.2
1940	-15.6	-12.9
1941	33.5	12.5

SuperBowl XLVII: Ravens vs 49ers

TradeSports.com

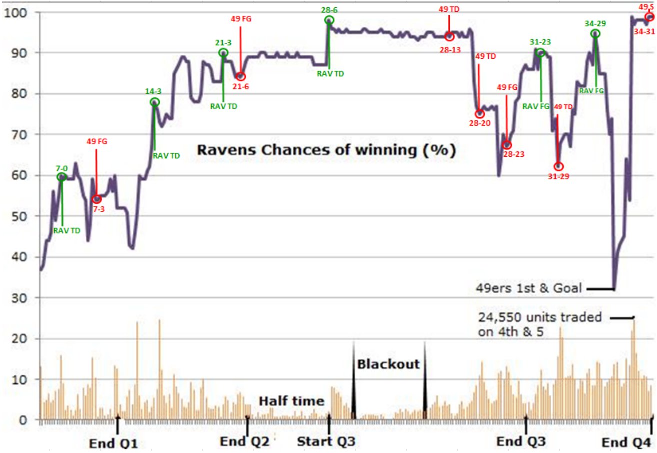


Figure 7: SuperBowl XLVII

Super Bowl XLVII: Ravens vs 49ers

- Super Bowl XLVII was held at the Superdome in New Orleans on February 3, 2013.
- We will track $X(t)$ which corresponds to the Raven's lead over the 49ers at each point in time. Table 3 provides the score at the end of each quarter.

t	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
Ravens	0	7	21	28	34
49*ers*	0	3	6	23	31
$X(t)$	0	4	15	5	3

SuperBowl XLVII by Quarter

Initial Market

Initial *point spread* Ravens being a four point underdog, $\mu = -4$.

$$\mu = \mathbb{E}\left(X(1)\right) = -4.$$

The Ravens upset the 49ers by $34 - 31$ and $X(1) = 34 - 31 = 3$ with the point spread being beaten by 7 points.

To determine the markets' assessment of the probability that the Ravens would win at the beginning of the game we use the *money-line* odds.

Initial Market

- San Francisco -175
- Baltimore Ravens $+155$.

This implies that a bettor would have to place \$175 to win \$100 on the 49ers and a bet of \$100 on the Ravens would lead to a win of \$155.

Convert both of these money-lines to *implied probabilities* of the each team winning

$$p_{SF} = \frac{175}{100+175} = 0.686 \ \text{and} \ p_{Bal} = \frac{100}{100+155} = 0.392$$

Probabilities of Winning

The probabilities do not sum to one. This "overound" probability is talso known as the bookmaker's edge.

$$p_{SF} + p_{Bal} = 0.686 + 0.392 = 1.078$$

providing a 7.8% edge for the bookmakers.

the *"market vig"* is the implied probability of the bookie making money on the bet.

```
[1] 0.07235622
```

We use the mid-point of the spread to determine p implying that

$$p = \frac{1}{2} p_{Bal} + \frac{1}{2}(1 - p_{SF}) = 0.353$$

From the Ravens perspective, we have $p = \mathbb{P}(X(1) > 0) = 0.353$.

Baltimore's win probability started trading at $p_0^{mkt} = 0.38$

Half Time Analysis

The Ravens took a commanding $21-6$ lead at half time. Market was trading at $p^{mkt}_{\frac{1}{2}} = 0.90$.

- During the 34 minute blackout 42760 contracts changed hands with Baltimore's win probability ticking down from 95 to 94.
- The win probability peak of 95% occurred after a third-quarter kickoff return for a touchdown.
- At the end of the four quarter, however, when the 49ers nearly went into the lead with a touchdown, Baltimore's win probability had dropped to 30%.

Implied Volatility

To calculate the implied volatility of the Superbowl we substitute the pair $(\mu, p) = (-4, .353)$ into our definition and solve for σ_{IV}.

$$\sigma = \frac{\mu}{\Phi^{-1}(p)} \, ,$$

- We obtain

$$\sigma_{IV} = \frac{\mu}{\Phi^{-1}(p)} = \frac{-4}{-0.377} = 10.60$$

where $\Phi^{-1}(p) = qnorm(0.353) = -0.377$. The 4 point advantage assessed for the 49ers is under a $\frac{1}{2}\sigma$ favorite.

- The outcome $X(1) = 3$ was within one standard deviation of the pregame model which had an expectation of $\mu = -4$ and volatility of $\sigma = 10.6$.

Half Time Probabilities

What's the probability of the Ravens winning given their lead at half time?

At half time Baltimore led by 15 points, 21 to 6.

The conditional mean for the final outcome is $15 + 0.5 * (-4) = 13$ and the conditional volatility is $10.6\sqrt{1 - t}$.

These imply a probability of .96 for Baltimore to win the game.

- A second estimate of the probability of winning given the half time lead can be obtained directly from the betting market.

From the online betting market we also have traded contracts on `TradeSports.com` that yield a half time probability of $p_{\frac{1}{2}} = 0.90$.

What's the implied volatility for the second half?

p_t^{mkt} reflects all available information

- For example, at half-time $t = \frac{1}{2}$ we would update

$$\sigma_{IV,t=\frac{1}{2}} = \frac{l + \mu(1-t)}{\Phi^{-1}(p_t)\sqrt{1-t}} = \frac{15-2}{\Phi^{-1}(0.9)/\sqrt{2}} = 14$$

where $qnorm(0.9) = 1.28$.

- As $14 > 10.6$, the market was expecting a more volatile second half–possibly anticipating a comeback from the 49ers.

How can we form a valid betting strategy?

Given the initial implied volatility $\sigma = 10.6$.

At half time with the Ravens having a $l + \mu(1-t) = 13$ points edge

- We would assess with $\sigma = 10.6$

$$p_{\frac{1}{2}} = \Phi\left(13/(10.6/\sqrt{2})\right) = 0.96$$

probability of winning versus the $p_{\frac{1}{2}}^{mkt} = 0.90$ rate.

- To determine our optimal bet size, ω_{bet}, on the Ravens we might appeal to the Kelly criterion (Kelly, 1956) which yields

$$\omega_{bet} = p_{\frac{1}{2}} - \frac{q_{\frac{1}{2}}}{O^{mkt}} = 0.96 - \frac{0.1}{1/9} = 0.60$$

Multivariate Normal

In the multivariate case, the normal-normal model is

$$\theta \sim N(\mu_0, \Sigma_0), \quad y \mid \theta \sim N(\theta, \Sigma).$$

The posterior distribution is

$$\theta \mid y \sim N(\mu_1, \Sigma_1),$$

where

$$\Sigma_1 = (\Sigma_0^{-1} + \Sigma^{-1})^{-1}, \quad \mu_1 = \Sigma_1(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}y).$$

The predictive distribution is

$$y_{new} \mid y \sim N(\mu_1, \Sigma_1 + \Sigma).$$

Normal With Unknown Variance

Consider, another example, when mean μ is fixed and variance is a random variable which follows some distribution $\sigma^2 \sim p(\sigma^2)$. Given an observed sample y, we can update the distribution over variance using the Bayes rule

$$p(\sigma^2 \mid y) = \frac{p(y \mid \sigma^2) p(\sigma^2)}{p(y)}.$$

Now, the total probability in the denominator can be calculated as

$$p(y) = \int p(y \mid \sigma^2) p(\sigma^2) d\sigma^2.$$

Normal With Unknown Variance

A conjugate prior that leads to analytically calculable integral for variance under the normal likelihood is the inverse Gamma. Thus, if

$$\sigma^2 \mid \alpha, \beta \sim IG(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{2(-\alpha-1)} \exp\left(-\frac{\beta}{\sigma^2}\right)$$

and

$$y \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$$

Then the posterior distribution is another inverse Gamma $IG(\alpha_{\text{posterior}}, \beta_{\text{posterior}})$, with

$$\alpha_{\text{posterior}} = \alpha + 1/2, \quad \beta_{\text{posterior}} = \beta + \frac{y-\mu}{2}.$$

Normal With Unknown Variance

Now, the predictive distribution over y can be calculated by

$$p(y_{new} \mid y) = \int p(y_{new}, \sigma^2 \mid y) p(\sigma^2 \mid y) d\sigma^2.$$

Which happens to be a t-distribution with $2\alpha_{\text{posterior}}$ degrees of freedom, mean μ and variance $\alpha_{\text{posterior}}/\beta_{\text{posterior}}$.

Black-Litterman

- Black and Litterman (1991, 1992) work for combining investor views with market equilibrium.
- In a multivariate returns setting the optimal allocation rule is

$$\omega^\star = \frac{1}{\gamma}\Sigma^{-1}\mu$$

  The question is how to specify (μ, Σ) pairs?

- For example, given $\hat{\Sigma}$, BL derive Bayesian inference for μ given market equilibrium model and *a priori* views on the returns of pre-specified portfolios which take the form

$$(\hat{\mu}|\mu) \sim \mathcal{N}\left(\mu, \tau\hat{\Sigma}\right) \text{ and } (Q|\mu) \sim \mathcal{N}\left(P\mu, \hat{\Omega}\right) \ .$$

Posterior Views

- Combining views, the implied posterior is

$$(\mu|\hat{\mu}, Q) \sim \mathcal{N}\left(Bb, B\right)$$

- The mean and variance are specified by

$$B = (\tau\hat{\Sigma})^{-1} + P'\hat{\Omega}^{-1}P \text{ and } b = (\tau\hat{\Sigma})^{-1}\hat{\mu} + P'\Omega^{-1}Q$$

These posterior moments then define the optimal allocation rule.

Satya Nadella: CEO of Microsoft

- In 2014, Satya Nadella became the CEO of Microsoft.
- The stock price of Microsoft has been on a steady rise since then.
- Suppose that you are a portfolio manager and you are interested in analyzing the returns of Microsoft stock compared to the market.
- Suppose you are managing a portfolio with two positions stock of Microsoft (MSFT) and an index fund that follows S&P500 index and tracks overall market performance.
- What is the mean returns of the positions in our portfolio?

Satya Nadella: CEO of Microsoft

- Assume the prior for the mean returns is a bivariate normal distribution, let $\mu_0 = (\mu_M, \mu_S)$ represent the prior mean returns for the stocks.
- The covariance matrix Σ_0 captures your beliefs about the variability and the relationship between these stocks' returns in the prior.

$$\Sigma_0 = \begin{bmatrix} \sigma_M^2 & \sigma_{MS} \\ \sigma_{MS} & \sigma_S^2 \end{bmatrix},$$

We will use the sample mean and covariance matrix of the historical returns as the prior mean and covariance matrix.

Satya Nadella: CEO of Microsoft

- The likelihood of observing the data, given the mean returns, is also a bivariate normal distribution.

  $$\Sigma = \begin{bmatrix} \sigma_M^2 & \sigma_{MS} \\ \sigma_{MS} & \sigma_S^2 \end{bmatrix},$$

  where σ_M^2 and σ_S^2 are the sample variances of the observed returns of MSFT and SPY, respectively, and σ_{MS} is the sample covariance of the observed returns of MSFT and SPY. The likelihood mean is given by

  $$\mu = \begin{bmatrix} \mu_M \\ \mu_S \end{bmatrix},$$

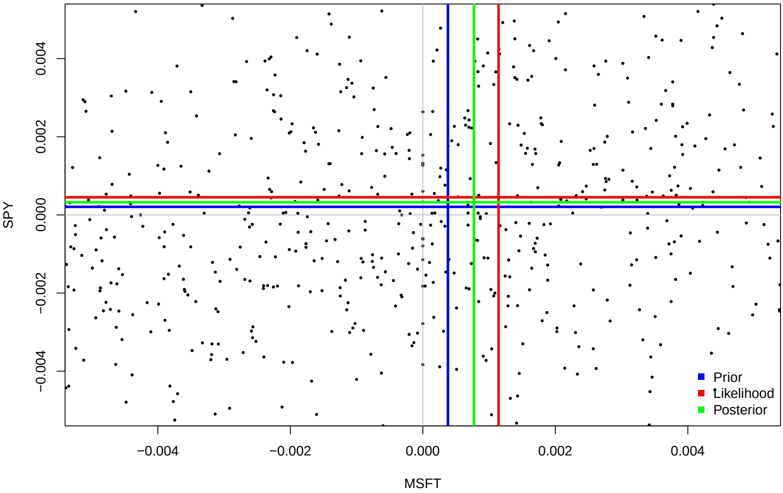
  where μ_M and μ_S are the sample means of the observed returns of MSFT and SPY, respectively.

Satya Nadella: CEO of Microsoft

- You update your beliefs (prior) about the mean returns using the observed data (likelihood).
- The posterior distribution, which combines your prior beliefs and the new information from the data, is also a bivariate normal distribution.
- The mean μ_{post} and covariance Σ_{post} of the posterior are calculated using Bayesian updating formulas, which involve μ_0, Σ_0, μ, and Σ.
- We use observed returns prior to Nadella's becoming CEO as our prior and analyze the returns post 2014.

Satya Nadella: CEO of Microsoft

```
[1] "MSFT" "SPY"
```



Mixtures of Conjugate Priors

- The mixture of conjugate priors is a powerful tool for modeling complex data.

  $$\theta \sim p(\theta) = \sum_{k=1}^{K} \pi_k p_k(\theta).$$

  Then the posterior is also a mixture of normal distributions, that is

  $$p(\theta \mid y) = p(y \mid \theta) \sum_{k=1}^{K} \pi_k p_k(\theta) / C.$$

Mixtures of Conjugate Priors

We introduce a normalizing constant for each component

$$C_k = \int p(y \mid \theta) p_k(\theta) d\theta.$$

then

$$p_k(\theta \mid y) = p_k(\theta) p(y \mid \theta) / C_k$$

is a proper distribution and our posterior is a mixture of these distributions

$$p(\theta \mid y) = \sum_{k=1}^{K} \pi_k C_k p_k(\theta \mid y) / C.$$

Meaning that we need to require

$$\frac{\sum_{k=1}^{K} \pi_k C_k}{C} = 1.$$

or

$$C = \sum_{k=1}^{K} \pi_k C_k$$

Mixture of two normal distributions

The prior distribution is a mixture of two normal distributions, that is

$$\mu \sim 0.5N(0,1) + 0.5N(5,1).$$

The likelihood is a normal distribution with mean μ and variance 1, that is

$$y \mid \mu \sim N(\mu, 1).$$

The posterior distribution is a mixture of two normal distributions, that is

$$p(\mu \mid y) \propto \phi(y \mid \mu, 1) \left(0.5\phi(\mu \mid 0, 1) + 0.5\phi(\mu \mid 5, 1)\right),$$

where $\phi(x \mid \mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.

Mixture of two normal distributions

We can calculate it using property of a normal distribution

$$\phi(x \mid \mu_1, \sigma_1^2)\phi(x \mid \mu_2, \sigma_2^2) = \phi(x \mid \mu_3, \sigma_3^2)\phi(\mu_1 - \mu_2 \mid 0, \sigma_1^2 + \sigma_2^2)$$

where

$$\mu_3 = \frac{\mu_1/\sigma_2^2 + \mu_2/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}, \quad \sigma_3^2 = \frac{1}{1/\sigma_1^2 + 1/\sigma_2^2}.$$

Mixture of two normal distributions

Given, we observed $y = 2$, we can calculate the posterior distribution for μ

```
[1] "Component parameters:\nMean = (1.0,3.5)\nVar = (0.5,0.
```

Exponential-Gamma Model

- Waiting times between events: consecutive arrivals of a Poisson process is exponentially distributed with mean $1/\lambda$.

$$f(x;\lambda) = \lambda e^{-\lambda x},\ x \geq 0$$

- λ is the rate parameter, which is the inverse of the mean
- special case of the Gamma distribution with shape 1 and scale $1/\lambda$.

Exponential Distribution	Parameters
Expected value	$\mu = EX = 1/\lambda$
Variance	$\sigma^2 = VarX = 1/\lambda^2$

Exponential Model: Examples

- Lifespan of Electronic Components: The exponential distribution can model the time until a component fails in systems where the failure rate is constant over time.
- Time Between Arrivals: In a process where events (like customers arriving at a store or calls arriving at a call center) occur continuously and independently, the time between these events can often be modeled with an exponential distribution.
- Radioactive Decay: The time until a radioactive atom decays is often modeled with an exponential distribution.

Exponential-Gamma Model

The *Exponential-Gamma* model assumes that the data follows an exponential distribution (likelihood). - The Gamma distribution is a flexible two-parameter family of distributions and can model a wide range of shapes.

$$\lambda \sim \text{Gamma}(\alpha, \beta)$$
$$x_i \sim \text{Exponential}(\lambda)$$

The posterior distribution of the rate parameter λ is given by:

$$p(\lambda \mid x_1, \ldots, x_n) \propto \lambda^{\alpha-1} e^{-\beta\lambda} \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^{\alpha+n-1} e^{-(\beta+\sum_{i=1}^{n} x_i)\lambda}$$

Exponential-Gamma Model

Posterior is a Gamma distribution with shape parameter $\alpha + n$ and rate parameter $\beta + \sum_{i=1}^{n} x_i$. The posterior mean and variance are given by:

$$\mathbb{E}[\lambda | x_1, \ldots, x_n] = \frac{\alpha + n}{\beta + \sum_{i=1}^{n} x_i}, \quad \mathrm{Var}[\lambda | x_1, \ldots, x_n] = \frac{\alpha + n}{(\beta + \sum_{i=1}^{n} x_i)^2}.$$

Notice, that $\sum x_i$ is the sufficient statistic for inference about parameter λ!

Exponential-Gamma Model

- Reliability Engineering: In situations where the failure rate of components or systems may not be constant and can vary, the Exponential-Gamma model can be used to estimate the time until failure, incorporating uncertainty in the failure rate.
- Medical Research: For modeling survival times of patients where the rate of mortality or disease progression is not constant and varies across a population. The variability in rates can be due to different factors like age, genetics, or environmental influences.
- Ecology: In studying phenomena like the time between rare environmental events (e.g., extreme weather events), where the frequency of occurrence can vary due to changing climate conditions or other factors.

Exploratory Data Analysis

Before deciding on a parametric model for a dataset. There are several tools that we use to choose the appropriate model. These include

1. Theoretical assumptions underlying the distribution (our prior knowledge about the data)
2. Exploratory data analysis
3. Formal goodness-of-fit tests

The two most common tools for exploratory data analysis are Q-Q plot, scatter plots and bar plots/histograms.

Q-Q plot

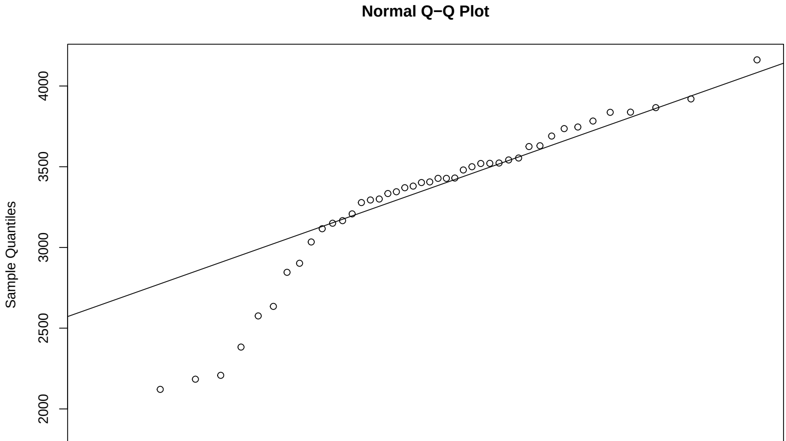
- Q-Q plot simply compares the quantiles of your data with the quantiles of a theoretical distribution (like normal, exponential, etc.).
- Quantile is the fraction (or percent) of points below the given value.
- That is, the i-th quantile is the point x for which i% of the data lies below x.
- On a Q-Q plot, if the two data sets come from a population with the same distribution, we should see the points forming a line that's roughly straight.

Q-Q plot

- If the two data sets x and y come from the same distribution, then the points $(x_{(i)}, y_{(i)})$ should lie roughly on the line $y = x$.
- If y comes from a distribution that's linear in x, then the points $(x_{(i)}, y_{(i)})$ should lie roughly on a line, but not necessarily on the line $y = x$.

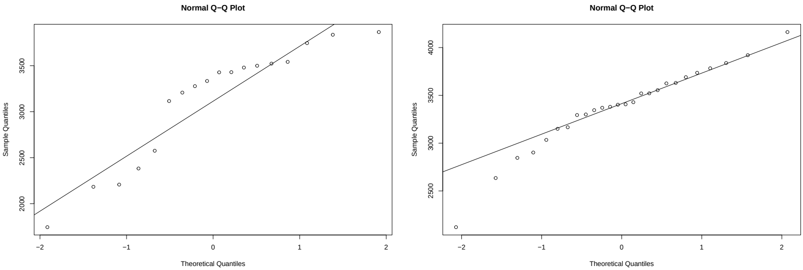
Noraml Q-Q plot

Q-Q plot for the Data on birth weights of babies born in a Brisbane hospital on December 18, 1997. The data set contains 44 records. A more detailed description of the data set can be found in `UsingR manual`.



Noraml Q-Q plot

The Q-Q plots look different if we split the data based on the gender



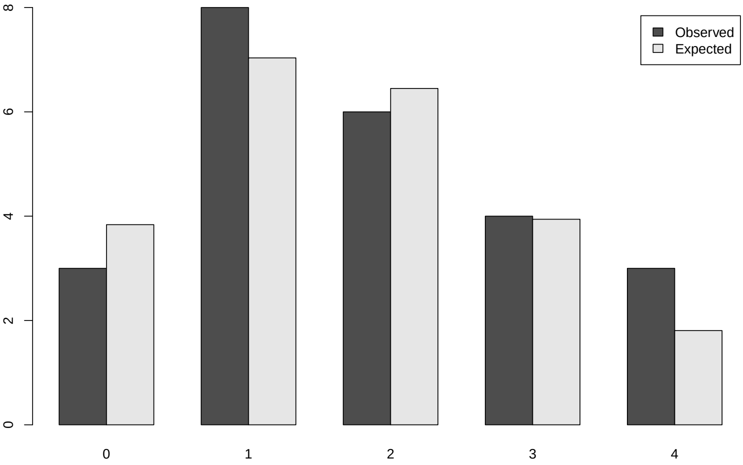
(a) Girls

(a) Boys

Histogram of baby weights by gender

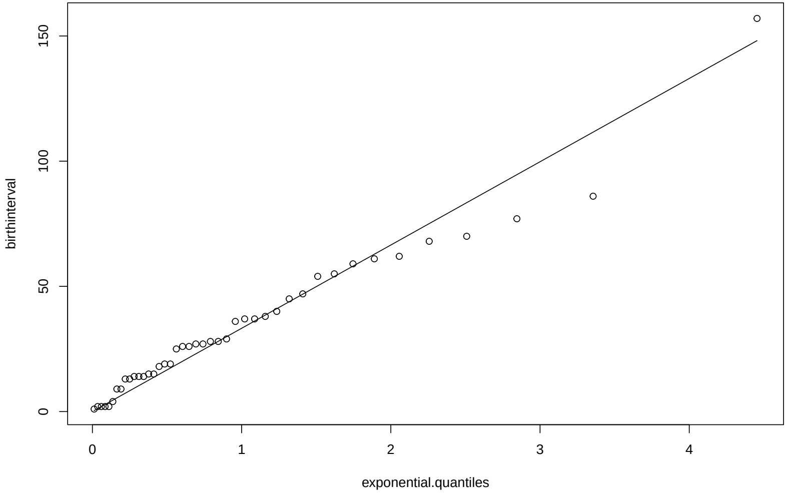
Noraml Q-Q plot

How about the times in hours between births of babies?



Exponential Q-Q plot

What about the Q-Q plot?



Brief List of Conjugate Models

Likelihood	Prior	Posterior
Binomial	Beta	Beta
Negative	Binomial	Beta
Poisson	Gamma	Gamma
Geometric	Beta	Beta
Exponential	Gamma	Gamma
Normal (mean unknown)	Normal	Normal
Normal (variance unknown)	Inverse Gamma	Inverse Gamm
Normal (mean and variance unknown)	Normal/Gamma	Normal/Gamm
Multinomial	Dirichlet	Dirichlet