Introduction to Bayesian Statistics

Bayesian vs frequentist debate

Pragmatic approach

- both are useful

Bayes theorem

$$P(x \mid y) = P(x \text{ and } y) \,/\, P(y)$$

$$= P(y \mid x)\, P(x) \,/\, P(y)$$

Bayes theorem

$P(x \mid y) = P(x \text{ and } y) / P(y)$

$= P(y \mid x) P(x) / P(y)$

Eg Draw a coin from a jar with 99% normal coins and 1% double headers. In 3 tosses observe 3 heads.

What is probability that this is a double headed coin?

Bayes theorem

	P(x or x’)	P(y | x or x’)	P(y|x)* P(x)
Fair coin	0.99	0.125	0.124
Double headeed	0.01	1.0	0.01
Total = P(y)			0.134

$$P(x \mid y) = P(y \mid x)\, P(x) / P(y)$$

$$= 1.0 * 0.01 / 0.135 = 0.075 = 0.01 / (0.124 + 0.01)$$

Definition of probability

Frequentists	Bayesians
Long run frequency	Subjective
Discriminate b/t random variable and parameter	Don’t discriminate
Use Bayes theorem only with random variable	Use Bayes theorem for both

Estimating a parameter

Frequentist approach

$y = u + s + e \qquad e \sim N(0,\sigma^2)$

$L(s) = P(y \mid s)$

ML estimate s-hat = value of s the maximizes L(s)

Estimating a parameter

Frequentist approach

s-hat ~ N(s, se^2)

P(s- 2*se < s-hat < s + 2*se) = 0.95

P(s-hat - 2*se < s < s-hat +2*se) < 0.95

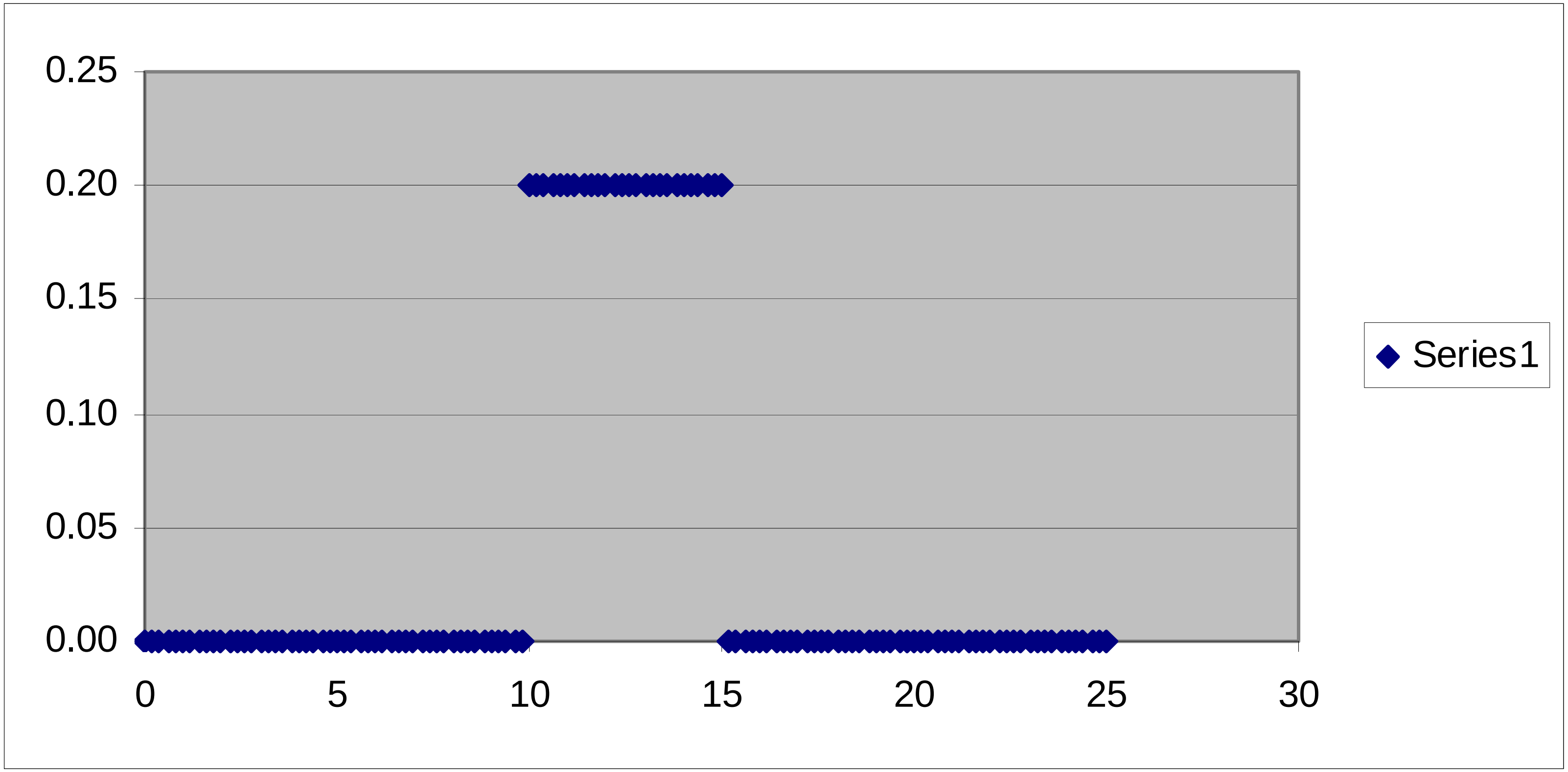
NOT P(20 < s < 30) = 0.95

Estimating a parameter

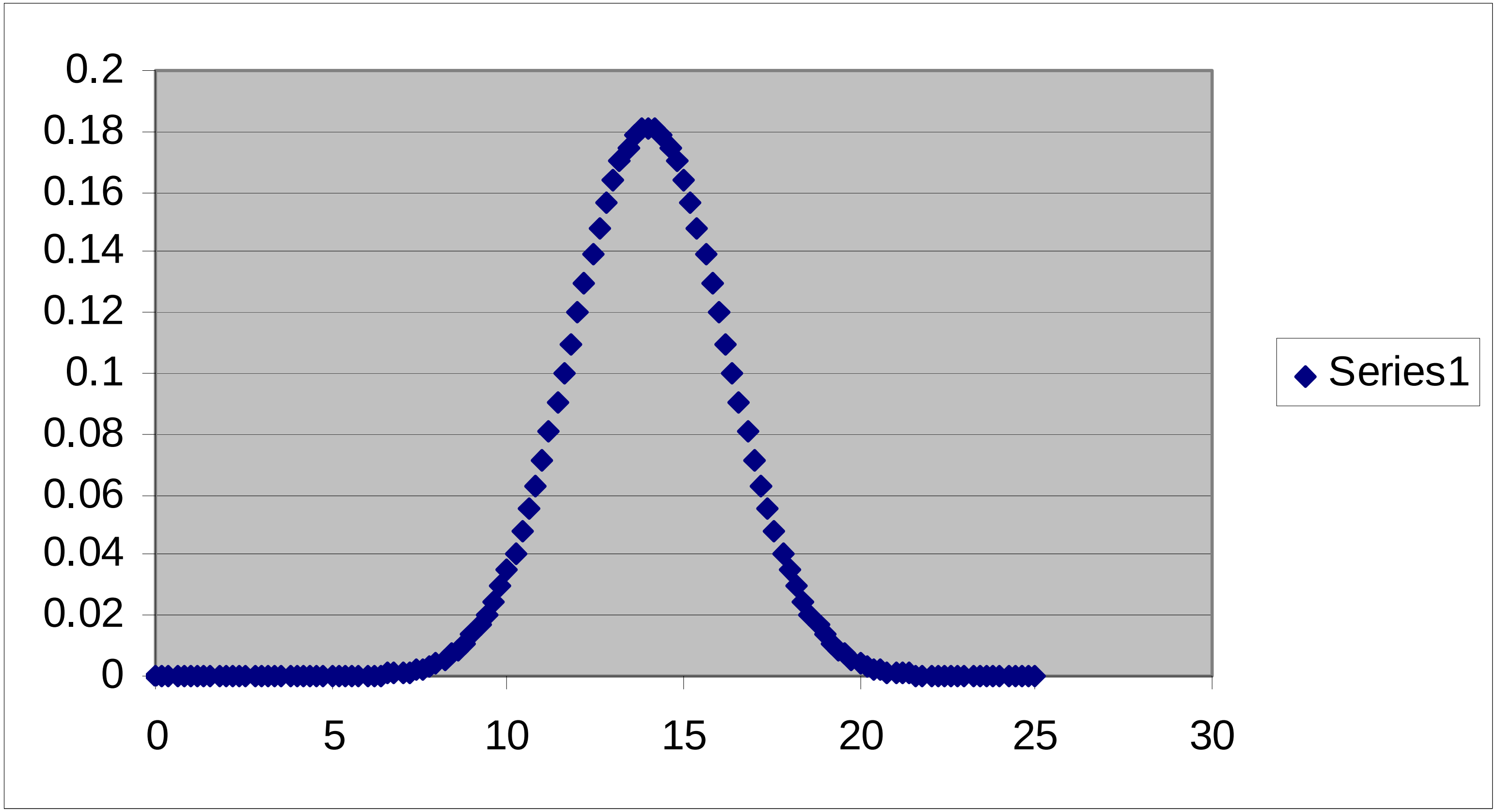
Bayesian approach

$P(s \mid y) \propto P(y \mid s) * P(s)$

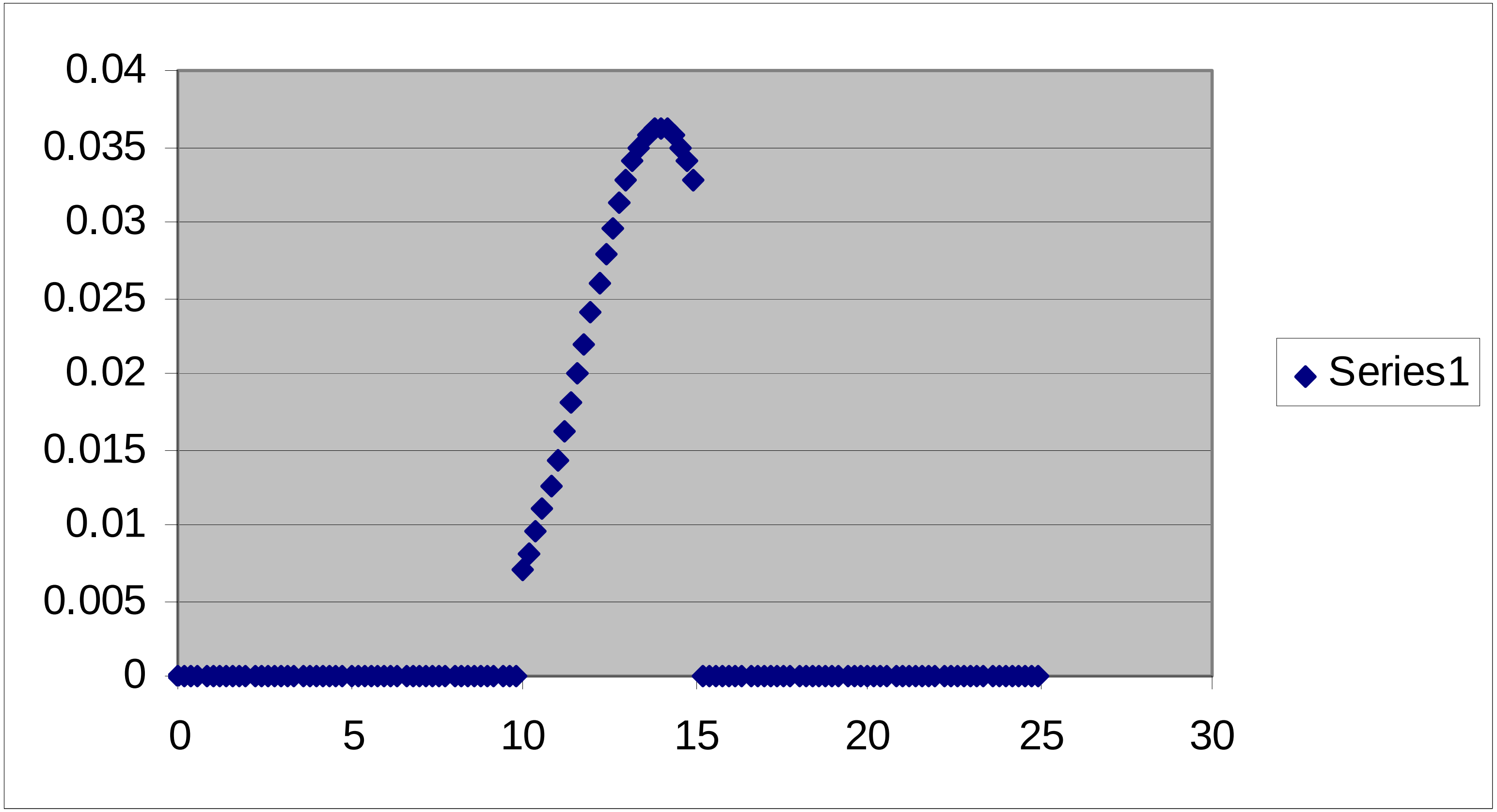
Prior distribution



Likelihood



Posterior



Fixed and random effects

Frequentist approach

$y = u + s + e$ $e \sim N(0,\sigma^2)$ $s \sim N(0,\sigma_s^2)$

$\text{s-hat} = \sum(y-u) / (n+\lambda)$ $\lambda = \sigma^2 / \sigma_s^2$

$\text{s-hat} = E(s \mid y)$

eg = 100 *100/ 116 = 86

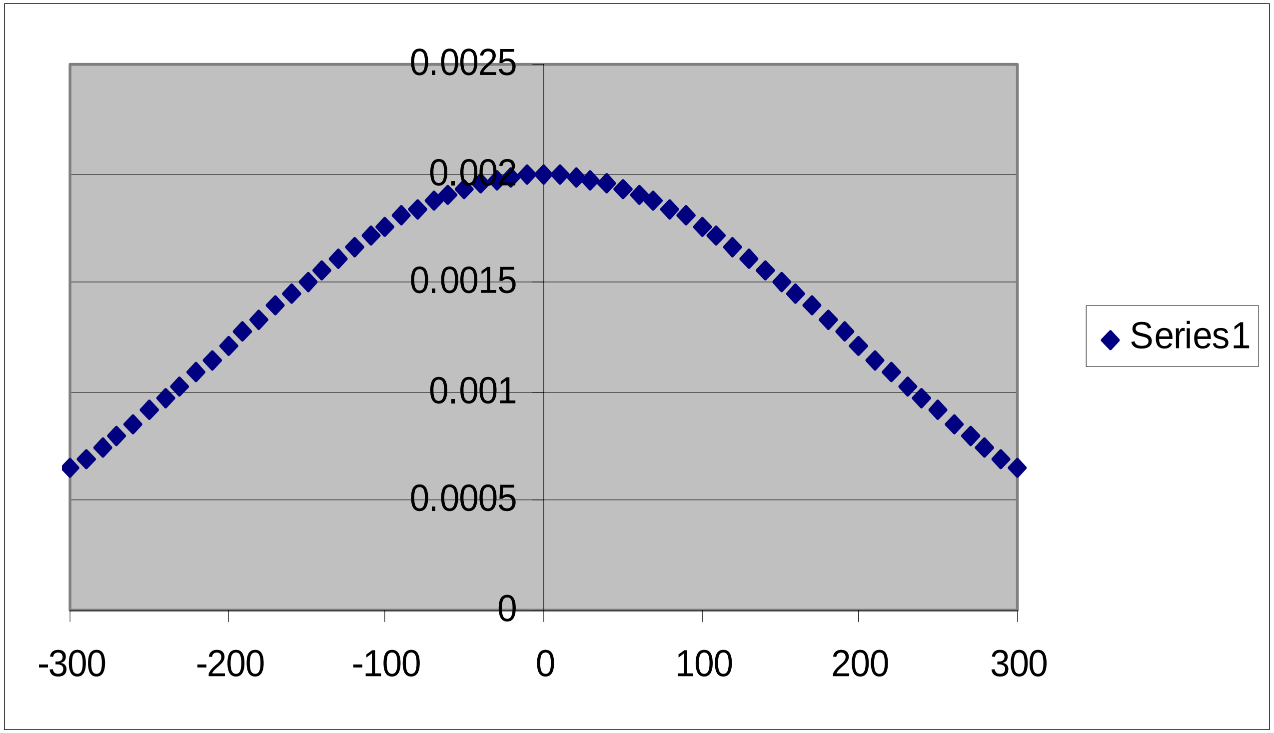
Fixed and random effects

Bayesian approach

$$P(s \mid y) = P(s) * P(y \mid s)$$

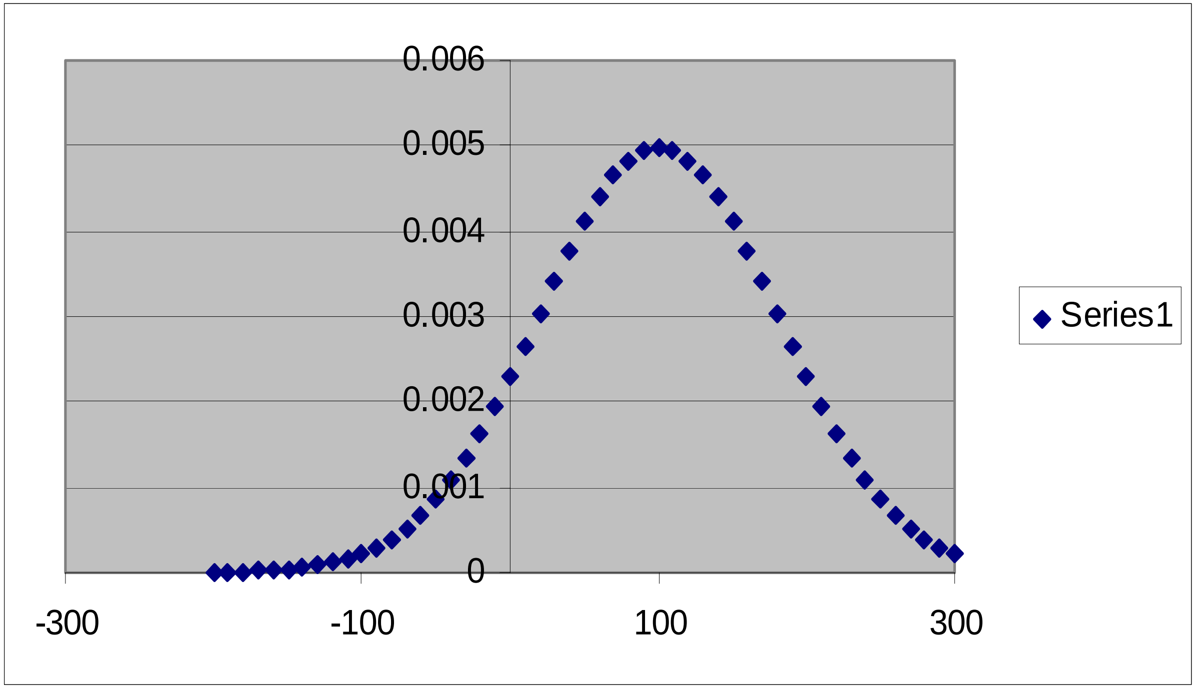
Fixed and random effects

Bayesian approach - Prior



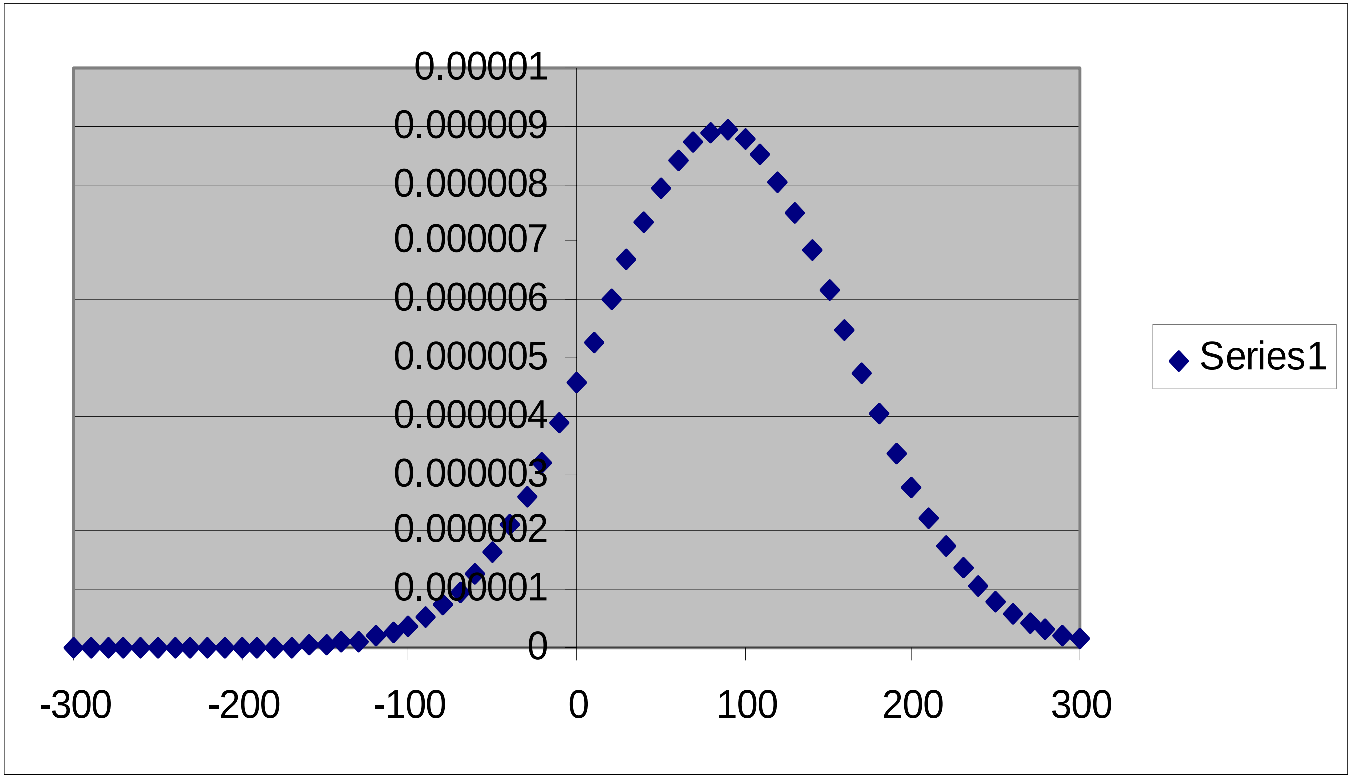
Fixed and random effects

Bayesian approach - Likelihood



Fixed and random effects

Bayesian approach - Posterior



Fixed and random effects

Frequentist approach

Fixed	Random
E(s-hat | s) = s	E(s | s-hat) = s-hat
unbiased	regressed back
exaggerated	best
summary of experiment	decision making

Fixed and random effects

Bayesian approach

Fixed Random

No difference

Treat all as ‘random’ with appropriate prior

Estimating multiple parameters

Frequentist approach

$L(\mu, \sigma) = P(y \mid \mu, \sigma)$

ML estimates μ-hat, σ-hat are the joint maximium of L

σ^2-hat = $\sum(y\text{-ybar})^2/n$ which is biased

Estimating multiple parameters

Bayesian approach

- $P(\mu,\sigma^2 \mid y) = P(y \mid \mu, \sigma^2) * P(\mu,\sigma^2)$.
- If the joint distribution is known, it is possible to calculate the marginal distribution by integrating over one of the variables. For instance, the marginal distribution of σ^2 is
- $P(\sigma^2 \mid y) = \int P(\mu, \sigma^2 \mid y)\, d\mu$

Nuisance parameters

Frequentist approach

1. Fit in the model (eg hys effects)
2. Integrate them out (eg sire effects)
3. Restricted ML (eg REML of variances)

Nuisance parameters

Bayesian approach

Integrate them out to get marginal posterior

Statistical inference

Frequentist approach

H0 vs H1

$P(\text{lower} < \text{test statistic} < \text{upper} \mid H0) = 0.95$

Statistical inference

Bayesian approach

Posterior

$P(\text{lower} < \text{parameter} < \text{upper}) = 0.95$

Conclusions

Advantages of Bayesian approach

Removes fixed vs random effect distinction

Better estimates for decision making provided prior reasonable

Conclusions

Advantages of Frequentist approach

Summary of experiment, unpolluted

Don’t need priors

Simpler hypothesis testing

Gibbs sampling

MCMC

MCMC = Markov Chain
Monte Carlo

Used to draw samples from posterior distribution

Numerical solutions for complex model by solving small simple steps

Heavy on computing time

Eg Gibbs, Metropolis-Hastings

Gibbs sampling

Sample one parameter at a time assuming the current values of all the other parameters are correct

Eg Variance components

σ_g^2 is easy to estimate if you know g's

A very simple example of Gibbs sampling

Sire S mated to Dam D produces offspring O who carries a recessive lethal gene whose allele frequency in the population is 0.1.

What is the probability that S carries the lethal?

P(S | O)?

Need conditional distributions

We will sample from $P(S \mid D, O)$ and $P(D \mid S,O)$

Obtain using Bayes theorem

$$P(S \mid D,O) = P(O \mid S,D) * P(S \mid D) / P(O \mid D)$$
$$= P(O \mid S,D) * P(S) / P(O \mid D)$$

D= ++

S	P(S)	P(O=+m | S, D)	P(S)*P(O=+m | S,D)	P(S | O, D)
++	0.81	0	0	0
+m	0.18	0.5	0.09	1
total = P(O = +m | D = ++)			0.09	

D = +*m*

++	0.81	0.5	0.405	0.82
+m	0.18	0.5	0.09	0.18
total = P(O = +m | D = +m)			0.495	

Cycle	Gd	Gs
1	++	+m
2	++	+m
3	++	+m
4	++	+m
5	++	+m
6	+m	++
7	+m	++
8	+m	+m
9	+m	++
10	+m	+m
11	++	+m
12	++	+m
13	++	+m
14	+m	++
15	+m	++
16	+m	++
17	+m	++
18	+m	+m
19	++	+m
20	++	+m

In these 20 cycles, we sampled Gs = ++ 7times and Gs=+m 13 times. Therefore, using these samples we would estimate that the P(Gs = +m | Go=+m) is 13/20 = 0.65.

Gibbs practicalities

- Burn-in
- Autocorrelation
- reducibility
- joint sampling
- length of chain(s)

Normally distributed data

We sample 10 observations from a population.

What is the mean and variance of the population?

Normally distributed data

$y = \mu + e$, $e \sim N(0,\sigma^2)$

What we want are the marginal posterior distributions

$P(\mu \mid y)$ and $P(\sigma^2 \mid y)$

For gibbs sampling we need the conditional distributions

$P(\mu \mid y, \sigma^2)$ and $P(\sigma^2 \mid y, \mu)$

Normally distributed data

For gibbs sampling we need the conditional distributions

$P(\mu \mid y, \sigma^2) \propto P(y \mid \mu , \sigma^2)* P(\mu)$

$P(\sigma^2 \mid y, \mu) \propto P(y \mid \mu , \sigma^2)* P(\sigma^2)$

$P(y \mid \mu , \sigma^2) \propto (\sigma^2)^{-n/2} \exp\{-\sum(y- \mu)^2/(2 \sigma^2)\}$

Normally distributed data

$P(y \mid \mu , \sigma^2) \propto (\sigma^2)^{-n/2} \exp\{-\sum(y- \mu)^2/(2 \sigma^2)\}$

As a funcion in μ this is a normal distribution

$\mu \sim N(\sum y/n, \sigma^2/n)$

As a function of σ^2 it is a scaled inverse chi-square distribution with n-2 degrees of freedom and scaled by $\sum(y- \mu)^2$

Normally distributed data

As a funcion in μ this is a normal distribution

$\mu \sim N(\sum y/n, \sigma^2/n)$

As a function of σ^2 it is a scaled inverse chi-square distribution with n-2 degrees of freedom and scaled by $\sum(y- \mu)^2$

So sample μ from a normal and sample σ^2 by sampling a chi-square, inverting it and multiplying by $\sum(y- \mu)^2$

Gibbs in a linear model

$Y = Xb + e$, $e \sim N(0,V)$

$b \sim N(0, W)$

$\sigma_i^2 \sim$ scaled inverted chi-square

$(X'V^{-1} X + W^{-1})b = X'V^{-1} y$

$C\,b = z$

$b \sim N(C^{-1} z, C^{-1})$

$\sigma_a^2 \sim$ scaled inverted chi-square with scale

$(a'A^{-1}a + S_a)$ and with $(n_a + v_a)$ df

Conclusions

- Gibbs sampling is easy to do because conditional distributions are usually easy
- Especially if you introduce variables for ‘missing data’
- Similar to EM algorithm
- Numerical problems
- All variables (parameters and random variables) are treated alike