Statistics 225
Bayesian Statistical Analysis

Hal S. Stern
Department of Statistics
University of California, Irvine

Prerequisites

- Probability (distns, transformations)
- Statistical Inference (standard procedures)
- Ideally two semesters at graduate level

Broad Outline

- Univariate/multivariate models
- Hierarchical models and model checking
- Computation
- Other models (glm’s, missing data, etc.)

Computing

- Splus - mostly covered in class
- BUGS - completely covered in class
- Other - at your own risk
Stat 225
History

- Bayes & Laplace (late 1700s) - inverse probability
  - probability - statements about observables given assumptions about unknown parameters
  - inverse probability - statements about unknown parameters given observed data values

- Ex: given \( y \) successes in \( n \) iid trials with probability of success \( \theta \), find \( \Pr(a < \theta < b) \)

- Little after that except for isolated individuals (e.g., Jeffreys)

- Interest resumes in mid 1900s (the term Bayesian statistics is born)

- Computational advances in late 20th century have led to increase in interest
Stat 225
Bayes/Frequentist Controversy

- Bayes
  - parameters as random variables
  - subjective probability (for some people)

- Frequentist
  - parameters as fixed but unknown quantities
  - probability as long-run frequency

- Some controversy in the past (and present)

- Message in this course is NOT adversarial
Stat 225
Some Things Not Discussed (Much)

- The following terms are sometimes associated with Bayesian statistics but will not receive much attention here:
  - decision theory
  - nonparametric Bayesian methods
  - subjective probability
  - objective Bayesian methods
  - maximum entropy
Stat 225
Motivating Example: Cancer Maps

- Kidney cancer mortality rates
  (Manton et al. - JASA, 1989)
  - Analyses of age-standardized death rates for cancer of kidney/ureter by U.S. county
  - Two maps of estimated rates
    * Direct calculation: use observed rates in county/age-group cells to form estimates
    * Empirical Bayes: modeling to stabilize estimated rates
Stat 225
Motivating Example: SAT coaching

- SAT coaching study
  - Randomized experiments in 8 schools
  - Separate analyses
  - Outcome is SAT-Verbal score
  - Effect of treatment (coaching) estimated using analysis of covariance

<table>
<thead>
<tr>
<th>School</th>
<th>Estimated treatment effect</th>
<th>Standard error of effect estimate</th>
<th>Treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28</td>
<td>15</td>
<td>?</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>10</td>
<td>?</td>
</tr>
<tr>
<td>C</td>
<td>-3</td>
<td>16</td>
<td>?</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
<td>11</td>
<td>?</td>
</tr>
<tr>
<td>E</td>
<td>-1</td>
<td>9</td>
<td>?</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>11</td>
<td>?</td>
</tr>
<tr>
<td>G</td>
<td>18</td>
<td>10</td>
<td>?</td>
</tr>
<tr>
<td>H</td>
<td>12</td>
<td>18</td>
<td>?</td>
</tr>
</tbody>
</table>
Stat 225
Bayesian inference: Two key ideas

• Explicit use of probability for quantifying uncertainty
  – probability models for data given parameters
  – probability distributions for parameters

• Inference for unknowns conditional on observed data
  – inverse probability
  – Bayes’ theorem (hence the modern name)
  – formal decision-making
Introduction to Bayesian Methods
Notation/Terminology

- $\theta =$ unobservable quantities (parameters)
- $y =$ observed data (outcomes, responses, random variable)
- $x =$ explanatory variables (covariates, often treated as fixed)
- Don’t usually distinguish between upper and lower case roman letters since everything is a random variable
- $\hat{y} =$ unknown but potentially observable quantities (predictions, response to a different treatment)
- NOTE: don’t usually distinguish between univariate, multivariate quantities
Introduction to Bayesian Methods
Notation/Terminology

• \( p(\cdot) \) or \( p(\cdot|\cdot) \) denote distributions (generic)

• It would take too many letters if each distn received its own letter

• We write \( Y|\mu, \sigma^2 \sim N(\mu, \sigma^2) \) to denote that \( Y \) has a normal density

• We write \( p(y|\mu, \sigma^2) = N(y|\mu, \sigma^2) \) to refer to the normal density with argument \( y \)

• Same for other distributions: Beta\((a, b)\), Unif\((a, b)\), Exp\((\theta)\), Pois\((\lambda)\), etc.
Introduction to Bayesian Methods
The Bayesian approach

• Focus here is on three step process
  – specify a full probability model
  – posterior inference via Bayes’ rule
  – model checking/sensitivity analysis

• Usually an iterative process - specify model, fit and check, then respecify model
Introduction to Bayesian Methods
Specifying a full probability model

- Data distribution $p(y|\theta) = p(\text{data} \mid \text{parameters})$
  - also known as sampling distribution
  - $p(y|\theta)$ when viewed as a function of $\theta$ is also known as the likelihood function $L(\theta|y)$

- Prior distribution $p(\theta)$
  - may contain subjective prior information
  - often chosen vague/uninformative
  - mathematical convenience

- Marginal model
  - above can be combined to determine implied marginal model for $y$ .... $p(y) = \int p(y|\theta)p(\theta)d\theta$
  - useful for model checking
  - Bayesian way of thinking leads to new distns that can be useful even for frequentists
Introduction to Bayesian Methods
Posterior inference/Model checking

• Posterior inference
  – Bayes’ thm to derive posterior distribution

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

  – probability statements about unknowns

  – formal decision-making is based on posterior distn

  – sometimes write \( p(\theta|y) \propto p(\theta)p(y|\theta) \) because the denominator is a constant in terms of \( \theta \)

• Model checking/sensitivity analysis
  – does the model fit

  – are conclusions sensitive to choice of prior distn/likelihood
Introduction to Bayesian Methods
Likelihood, Odds, Posteriors

• Recall that $p(\theta|y) \propto p(\theta)p(y|\theta)$

  – posterior $\propto$ prior $\times$ likelihood

  – consider two possible values of $\theta$, say $\theta_1$ and $\theta_2$

  \[
  \frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)}{p(\theta_2)} \times \frac{p(y|\theta_1)}{p(y|\theta_2)}
  \]

  – posterior odds $=\$ prior odds $\times$ likelihood ratio

  – note likelihood ratio is still important
Introduction to Bayesian Methods
Likelihood principle

- Likelihood principle - if two likelihood functions agree, then the same inferences about \( \theta \) should be drawn

- Traditional frequentist methods violate this

- Example: given a sequence of coin tosses with constant probability of success \( \theta \) we wish to test \( H_0 : \theta = 0.5 \)
  - observe 9 heads, 3 tails in 12 coin tosses
  - if binomial sampling (\( n = 12 \) fixed), then
    \[
    L(\theta|y) = p(y|\theta) = \binom{12}{9} \theta^9 (1 - \theta)^3
    \]
    and \( p \)-value is .073
  - if negative binomial sampling (sample until 3 tails), then
    \[
    L(\theta|y) = p(y|\theta) = \binom{11}{9} \theta^9 (1 - \theta)^3
    \]
    and \( p \)-value is .033
  - but data (and likelihood function) is the same ... 9 successes, 3 failures ... and should carry the same information about \( \theta \)
Introduction to Bayesian Methods

Independence

- A common statement in statistics: assume $Y_1, \ldots, Y_n$ are iid r.v.’s

- In Bayesian class, we need to think hard about independence

- Why?
  - Consider two "indep" Bernoulli trials with probability of success $\theta$
  - It is true that
    \[ p(y_1, y_2 | \theta) = \theta^{y_1 + y_2} (1-\theta)^{2-y_1-y_2} = p(y_1 | \theta)p(y_2 | \theta) \]
    so that $y_1$ and $y_2$ are independent given $\theta$
  - But ... $p(y_1, y_2) = \int p(y_1, y_2 | \theta)p(\theta)d\theta$ may not factor
  - If $p(\theta) = \text{Unif}(\theta|0,1) = 1$ for $0 < \theta < 1$, then
    \[ p(y_1, y_2) = \Gamma(y_1 + y_2 + 1) \Gamma(3 - y_1 - y_2) / \Gamma(4) \]
    so $y_1$ and $y_2$ are not independent in their marginal distribution
**Introduction to Bayesian Methods**

**Exchangeability**

- If independence is no longer the key, then what is?

- Exchangeability
  - Informal defn: subscripts don’t matter
  - Formally: given events $A_1, \ldots, A_n$, we say they are exchangeable if
    \[ P(A_1 A_2 \ldots A_k) = P(A_{i_1} A_{i_2} \ldots A_{i_k}) \]
    for every $k$ where $i_1, i_2, \ldots, i_n$ are a permutation of the indices
  - Similarly, given random variable $Y_1, \ldots, Y_n$, we say they are exchangeable if
    \[ P(Y_1 \leq y_1, \ldots, Y_k \leq y_k) = P(Y_{i_1} \leq y_1, \ldots, Y_{i_k} \leq y_k) \]
    for every $k$
Introduction to Bayesian Methods
Exchangeability and independence

- Relationship between exchangeability and independence
  - r.v.’s that are iid given $\theta$ are exchangeable
  - an infinite sequence of exchangeable r.v.’s can always be thought of as iid given some parameter (Definetti)
  - note previous point requires an infinite sequence

- What is not exchangeable?
  - time series, spatial data
  - may become exchangeable if we explicitly include time in the analysis
  - i.e., $y_1, y_2, \ldots, y_t, \ldots$ are not exchangeable but $(t_1, y_1), (t_2, y_2), \ldots$ may be
Introduction to Bayesian Methods
A simple example

- Hemophilia - blood clotting disease
  - sex-linked genetic disease on X chromosome
  - males (XY) - affected or not
  - females (XX) - may have 0 copies of disease gene (not affected), 1 copy (carrier), 2 copies (usually fatal)

- Consider a woman – brother is a hemophiliac, father is not
  - we ignore the possibility of a mutation introducing the disease
  - woman’s mother must be a carrier
  - woman inherits one X from mother
    -- > 50/50 chance of being a carrier

- Let $\theta = 1$ if woman is carrier, 0 if not
  - a priori we have $\Pr(\theta = 1) = \Pr(\theta = 0) = 0.5$

- Let $y_i =$ status of woman’s $i$th male child
  (1 if affected, 0 if not)
Introduction to Bayesian Methods

A simple example (cont’d)

• Given two unaffected sons (not twins), what inference can be drawn about $\theta$?

• Assume two sons are iid given $\theta$

• $\Pr(y_1 = y_2 = 0 | \theta = 1) = 0.5 \times 0.5 = .25$
  $\Pr(y_1 = y_2 = 0 | \theta = 0) = 1 \times 1 = 1.00$

• Posterior distn by Bayes’ theorem

\[
\Pr(\theta = 1 | y) = \frac{\Pr(y | \theta = 1) \Pr(\theta = 1)}{\Pr(y)} \]

\[
= \frac{\Pr(y | \theta = 1) \Pr(\theta = 1)}{\Pr(y | \theta = 1) \Pr(\theta = 1) + \Pr(y | \theta = 0) \Pr(\theta = 0)}
\]

\[
= \frac{.25 \times .5}{.25 \times .5 + 1 \times .5} = .2
\]
Introduction to Bayesian Methods
A simple example (cont’d)

• Odds version of Bayes’ rule
  – prior odds \( \Pr(\theta = 1)/\Pr(\theta = 0) = 1 \)
  – likelihood ratio \( \Pr(y|\theta = 1)/\Pr(y|\theta = 0) = 1/4 \)
  – posterior odds = 1/4
    (posterior prob = .25/(1 + .25) = .20)

• Updating for new information
  – suppose that a 3rd son is born (unaffected)
  – note: if we observe an affected child, then we know \( \theta = 1 \) since that outcome is assumed impossible when \( \theta = 0 \)
  – two approaches to updating analysis
    * redo entire analysis \( (y_1, y_2, y_3 \text{ as data}) \)
    * update using only new data \( (y_3) \)
Introduction to Bayesian Methods
A simple example (cont’d)

- Updating for new information - redo analysis
  - as before but now $y = (0, 0, 0)$
  - $\Pr(y|\theta = 1) = .5 \times .5 \times .5 = .125$,
    \[ \Pr(y|\theta = 0) = 1 \]
  - $\Pr(\theta = 1|y) = \frac{.125 \times .5}{(.125 \times .5 + 1 \times .5)} = .111$

- Updating for new information - updating
  - take previous posterior distn as new prior distn
    ($\Pr(\theta = 1) = .2$ and $\Pr(\theta = 0) = .8$)
  - take data as consisting only of $y_3$
  - $\Pr(\theta = 1|y_3) = \frac{.5 \times .2}{(.5 \times .2 + 1 \times .8)} = .111$
  - same answer!
Introduction to Bayesian Methods
Probability review

- Probability (mathematical definition):
  A set function that is
  - nonnegative
  - additive over disjoint sets
  - sums to one over entire sample space

- For Bayesian methods probability is a fundamental measure of uncertainty
  - Pr(1 < y < 3|θ = 0) or Pr(1 < y < 3) is interesting before data has been collected
  - Pr(1 < θ < 3|y) is interesting after data has been collected

- Where do probabilities come from?
  - frequency argument (e.g., 10,000 coin tosses)
  - physical argument (e.g., symmetry in coin toss)
  - subjective (e.g., if would be willing to bet on NY Giants given 1:1 odds, then must believe the probability Giants win is greater than .5)
Introduction to Bayesian Methods
Probability review

• Some terms/defns you should know
  – joint distn \( p(u, v) \)
  – marginal distn \( p(u) = \int p(u, v)dv \)
  – conditional distn \( p(u|v) = \frac{p(u, v)}{p(v)} \)
  – moments: \( E(u) = \int up(u)du = \int \int up(u, v)dvdv \)
    \[ \text{Var}(u) = \int (u - E(u))^2p(u)du \]
    \[ E(u|v) = \int up(u|v)du \text{ (a fn of } v) \]
  – conditional distns play a large role in Bayesian inference so the following rules are useful
    * \( E(u) = E(E(u|v)) \)
    * \( \text{Var}(u) = E(\text{Var}(u|v)) + \text{Var}(E(u|v)) \)

• transformations (one-to-one)
  * denote distn of \( u \) by \( p_u(u) \)
  * take \( v = f(u) \)
  * distribution of \( v \) is
    \( p_v(v) = p_u(f^{-1}(v)) \text{ in discrete case} \)
    \[ p_v(v) = p_u(f^{-1}(v))|J| \text{ in continuous case} \]
  where Jacobian \( J \) is \( \left| \frac{\partial u_i}{\partial v_j} \right| = \left| \frac{\partial f^{-1}(v)}{\partial v_j} \right| \)
• Simulation plays a big role in modern Bayesian inference and one particular transformation is important in this context

• Probability integral transform
  – suppose $X$ is a continuous r.v. with cdf $F_X(x)$
  – then $Y = F_X(X)$ has uniform distn on 0 to 1

• Application in simulations
  – if $U$ is uniform on $(0,1)$ and $F(\cdot)$ is cdf of a continuous r.v.
  – then $Z = F^{-1}(U)$ is a r.v. with cdf $F$
  – example:
    * let $F(x) = 1 - e^{-x/\lambda}$ = exponential cdf
    * then $F^{-1}(u) = -\lambda \log(1 - u)$
    * if we have a source of uniform random numbers then we can easily transform to construct samples from an exponential distn
Single Parameter Models

Introduction

- We introduce important concepts/computations in the one-parameter case

- There is little advantage to the Bayesian approach in these cases

- The benefits of the Bayesian approach are in hierarchical (often random effects) models

- Main approach is to teach via example

- First example is binomial data (Bernoulli trials)
  - easy
  - historical interest (Bayes, Laplace)
  - representative of a large class of distns (exponential families)
Single Parameter Models
Binomial Model

- Consider \( n \) exchangeable trials
- Data can be summarized by total \# of successes
- Natural model: define \( \theta \) as probability of success and take \( Y \sim \text{Bin}(n, \theta) \)
  \[ p(y|\theta) = \text{Bin}(y|n, \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \]
- Question - do we have to be explicit about conditioning on \( n \)? (usually are not)
- Prior distn: \( p(\theta) = \text{Unif}(\theta|0,1) \)
- Posterior distn:
  \[
p(\theta|y) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} / \int \binom{n}{y} \theta^y (1 - \theta)^{n-y} d\theta
  = (n + 1) \binom{n}{y} \theta^y (1 - \theta)^{n-y} = \frac{(n + 1)!}{y!(n - y)!} \theta^y (1 - \theta)^{n-y}
  = \frac{\Gamma(n + 2)}{\Gamma(y + 1) \Gamma(n - y + 1)} \theta^{y+1-1} (1 - \theta)^{n-y+1-1}
  = \text{Beta}(y + 1, n - y + 1)
\]
- Note: could have noticed \( p(\theta|y) \propto \theta^y (1 - \theta)^{n-y} \) and inferred it is a \( \text{Beta}(y + 1, n - y + 1) \) distn (formal calculation confirms this)
Single Parameter Models
Binomial Model

- Inference
  - draw inferences from posterior distn
  - point estimation
    * posterior mean = \( \frac{(y + 1)}{(n + 2)} \)
      (compromise between sample proportion \( \frac{y}{n} \)
       and prior mean 1/2)
    * posterior mode = \( \frac{y}{n} \)
    * best point estimate depends on loss function
    * posterior variance = \( \left( \frac{y+1}{n+2} \right) \left( \frac{n-y+1}{n+2} \right) \left( \frac{1}{n+3} \right) \)
  - interval estimation
    * 95\% central posterior interval - find a,b s.t.
      \[ \int_0^a \text{Beta}(\theta|y+1, n-y+1) d\theta = .025 \text{ and } \int_0^b \text{Beta}(\theta|y+1, n-y+1) d\theta = .975 \]
    * alternative is highest posterior density region
    * note this interval has the interpretation we want to give to traditional CIs
  - hypothesis test – don’t say anything now
Single Parameter Models
Binomial Model

- Inference by simulation
  - all of the inferences mentioned (point estimation, interval estimation) can be done via simulation
  - simulate 1000 draws from the posterior distribution

* available in standard packages
* MCMC for harder problems later

- point estimates easy to compute
  (now include Monte Carlo error)

- interval estimates easy – find percentiles of the simulated values
Single Parameter Models
Prior distributions

- Where do prior distributions come from?
  - a priori knowledge about $\theta$ (“deep thoughts”)
  - population interpretation (a population of possible $\theta$ values)
  - mathematical convenience

- Frequently rely on asymptotic results (to come) which guarantee that likelihood will dominate the prior distn in large samples
Single Parameter Models
Conjugate prior distributions

- Consider Beta(α, β) prior distn for binomial model
  - think of α, β as fixed now (but these could also be random and given their own prior distn)
  - $p(\theta|y) \propto \theta^{y+\alpha-1}(1 - \theta)^{n-y+\beta-1}$
  - recognize as kernel of Beta$(y + \alpha, n - y + \beta)$
  - example of conjugate distn - posterior distn is in the same parametric family as the prior distn
  - convenient mathematically
  - convenient interpretation - prior in this case is like observing α successes in α + β “prior” trials
Single Parameter Models
Conjugate prior distributions - general

- **Definition:**
  Let \( F \) be a class of sampling distn \( (p(y|\theta)) \).
  Let \( P \) be a class of prior distns \( (p(\theta)) \).
  \( P \) is **conjugate** for \( F \) if \( p(\theta) \in P \) and \( p(y|\theta) \in F \)
  implies that \( p(\theta|y) \in P \)

- Not a great definition ... trivially satisfied by \( P = \{ \text{ all distns} \} \) but this is not an interesting case

- **Exponential families** (most common distns):
  the only distns that are finitely parametrizable and have conjugate prior families
  - density of exponential families is
    \[
    p(y|\theta) = f(y)g(\theta)e^{\phi(\theta)^t u(y)}
    \]
    with \( \phi(\theta) \) denoting the natural parameter
  - \( p(\theta) \propto g(\theta)^{\eta}e^{\phi(\theta)^t \nu} \) will be conjugate family
  - **binomial:** \( \phi(\theta) = \log(\theta/(1-\theta)) \) and \( g(\theta) = 1-\theta \)
    conjugate prior distn is \( \theta^\nu(1-\theta)^{\eta-\nu} \)
Single Parameter Models
Conjugate prior distributions - general

• Advantages
  – mathematically convenient
  – easy to interpret
  – can provide good approx to many prior opinions (especially if we allow mixtures of distns from the conjugate family)

• Disadvantages
  – may not be realistic
Single Parameter Models
Nonconjugate prior distributions

- No real difference conceptually in how analysis proceeds

- Harder computationally

- Grid-based simulation
  - specify prior distn on a grid \( \Pr(\theta = \theta_i) = \pi_i \)
  - compute likelihood on same grid \( l_i = p(y|\theta_i) \)
  - posterior distn lives on the grid with \( \Pr(\theta = \theta_i|y) = \pi_i^* = \pi_i l_i / (\sum_j \pi_j l_j) \)
  - can sample from this posterior distn easily in Splus
  - can do better with a trapezoidal approx to the prior distn

- There are serious problems with grid-based simulation

- We will see better computational approaches
Single Parameter Models
Noninformative prior distributions

- Often there is a desire to have the prior distn play a minimal role the posterior distn (why?)

- Example: consider \( y_1, \ldots, y_n | \theta \sim \text{iid} \mathcal{N}(\theta, \sigma^2) \) and
\[
p(\theta | \mu, \tau^2) = \mathcal{N}(\theta | \mu, \tau^2)
\]
where \( \sigma^2, \mu, \tau^2 \) are known
  - a conjugate family
  - \( p(\theta | y) = \mathcal{N}(\theta | \hat{\mu}, V) \) with
\[
\hat{\mu} = \frac{n \bar{y} + \frac{1}{\tau^2} \mu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}
\]
  - note: if \( n \to \infty \) then posterior distn resembles \( p(\theta | y) = \mathcal{N}(\theta | \bar{y}, \sigma^2 / n) \); like classical sampling distn result (data dominates prior distn)
  - if \( \tau^2 \to \infty \), then \( p(\theta | y) \approx \mathcal{N}(\theta | \bar{y}, \sigma^2 / n) \)
  (this yields the same estimates and intervals as classical methods; can be thought of as noninformative)

  - same result would be obtained by taking \( p(\theta) \propto 1 \) BUT that is not a proper prior distn
  - we can use improper prior distn but must check that the posterior distn is a proper distn
Single Parameter Models
Noninformative prior distributions

• How do we find noninformative prior distributions?

• Flat or uniform distributions
  – did the job in the binomial and normal cases
  – makes each value of $\theta$ equally likely
  – but on what scale (should every value of $\log \theta$
    be equally likely or every value of $\theta$)

• Jeffrey’s prior
  – invariance principle – a rule for creating
    noninformative prior distns should be invariant
to transformation
  – if $p_\theta$ is prior distn for $\theta$ and we consider
    $\phi = h(\theta)$, so that $p_\phi(\phi) = p_\theta(h^{-1}(\phi)) |d\theta/d\phi|$
  – Jeffrey’s suggestion $p(\theta) \propto I(\theta)^{1/2}$ where
    $I(\theta)$ is the Fisher information
  – gives flat prior for $\theta$ in normal case
  – does this work for multiparameter problems?
Single Parameter Models
Noninformative prior distributions

• How do we find noninformative prior distributions? (cont’d)

• Pivotal quantities
  – location family has $p(y-\theta|\theta) = f(y-\theta)$ so should expect $p(y-\theta|y) = f(y-\theta)$ as well …… this suggests $p(\theta) \propto 1$
  – similarly for scale family we find $p(\theta) \propto 1/\theta$ (where $\theta$ is a scale parameter like normal s.d.)

• Vague, diffuse distributions
  – use conjugate or other prior distn with large variance
Single Parameter Models
Noninformative prior distributions - example

- Binomial case
  - Uniform on $\theta$ is Beta$(1, 1)$
  - Jeffreys’ prior distn is Beta$(1/2, 1/2)$
  - Uniform on natural parameter $\log(\theta/(1 - \theta))$ is Beta$(0, 0)$ (an improper prior distn)

- Summary on noninformative distn
  - very difficult to make this idea rigorous since it requires a definition of “information”
  - informally — this is a useful but dangerous idea
  - useful as a first approximation or first attempt
  - dangerous if applied automatically without thought
  - improper distributions can cause serious problems (improper posterior distns) that are hard to detect
  - some prefer vague or diffuse proper distributions as a way of expressing ignorance
Multiparameter Models

Introduction

- Now write \( \theta = (\theta_1, \theta_2) \) (at least two parameters)

- \( \theta_1 \) and \( \theta_2 \) may be vectors as well

- Key point here is how Bayesian approach handles “nuisance” parameters

- Posterior distn \( p(\theta_1, \theta_2 | y) \propto p(y | \theta_1, \theta_2)p(\theta_1, \theta_2) \)

- Suppose \( \theta_1 \) is of primary interest, i.e., want \( p(\theta_1 | y) \)
  - \( p(\theta_1 | y) = \int p(\theta_1, \theta_2 | y) d\theta_2 \) analytically or by numerical integration
  - \( p(\theta_1 | y) = \int p(\theta_1 | \theta_2, y)p(\theta_2 | y) d\theta_2 \) (often a convenient way to calculate)
  - \( p(\theta_1 | y) = \int p(\theta_1, \theta_2 | y) d\theta_2 \) by simulation (generate simulations of both and toss out the \( \theta_2 \)'s)

- Note: Bayesian results still usually match those of traditional methods. We don’t see differences until hierarchical models
Multiparameters Models
Normal example

- $y_1, y_2, \ldots, y_n | \mu, \sigma^2$ are iid $N(\mu, \sigma^2)$

- Prior distn: $p(\mu, \sigma^2) \propto 1/\sigma^2$
  - indep non-informative prior distns for $\mu$ and $\sigma^2$
  - equivalent to $p(\mu, \log \sigma) \propto 1$
  - not a proper distn

- Posterior distn:
  
  $p(\mu, \sigma^2 | y) \propto \left(\frac{1}{\sigma^2}\right)^{\bar{y} + 1} \exp\left[-\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2\right]$

  $\propto \left(\frac{1}{\sigma^2}\right)^{\bar{y} + 1} \exp\left[-\frac{1}{2\sigma^2} \left(\sum_i (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\right)\right]$

  - note that $\mu, \sigma^2$ are not indep in their posterior distn

  - posterior distn depends on data only through the sufficient statistics
Multiparameters Models
Normal example (cont’d)

- Further examination of joint posterior distribution

\[ p(\mu, \sigma^2 | y) \propto \left( \frac{1}{\sigma^2} \right)^{\frac{n+1}{2}} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_i (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \right) \right] \]

- conditional posterior distn \( p(\mu | \sigma^2, y) \)
  * examine joint posterior distn but now think of \( \sigma^2 \) as known
  * focus only on \( \mu \) terms
  * \( p(\mu | \sigma^2, y) \propto \exp[-\frac{1}{2\sigma^2}n(\bar{y} - \mu)^2] \)
  * just like known variance case
  * recognize \( \mu | \sigma^2, y \sim N(\bar{y}, \sigma^2/n) \)

- marginal posterior distn of \( \sigma^2 \), i.e., \( p(\sigma^2 | y) \)
  * \( p(\sigma^2 | y) = \int p(\mu, \sigma^2 | y) d\mu \)
  * alternative: note \( p(\sigma^2 | y) = p(\mu, \sigma^2 | y)/p(\mu | \sigma^2, y) \)
    (LHS doesn’t have \( \mu \), RHS does .... must be true for any \( \mu \))
  * \( p(\sigma^2 | y) \propto (\sigma^2)^{-(n+1)/2} \exp[-\frac{1}{2\sigma^2} \sum_i (y_i - \bar{y})^2] \)
  * known as scaled-inverse-\( \chi^2(n-1, s^2) \) distn with \( s^2 = \sum_i (y_i - \bar{y})^2/(n - 1) \)
Multiparameters Models
Normal example (cont’d)

• Further examination of joint posterior distribution

\[ p(\mu, \sigma^2 | y) \propto \left( \frac{1}{\sigma^2} \right)^{\frac{n}{2} + 1} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_i (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \right) \right] \]

– so far, \( p(\mu, \sigma^2 | y) = p(\sigma^2 | y)p(\mu | \sigma^2, y) \)

– this factorization can be used to simulate from joint posterior distn
  * generate \( \sigma^2 \) from \( \text{Inv-}\chi^2(n - 1, s^2) \) distn
  * then generate \( \mu \) from \( N(\bar{y}, \sigma^2/n) \) distn

– often most interested in \( p(\mu | y) \)
  * \( p(\mu | y) = \int_0^\infty p(\mu, \sigma^2 | y) d\sigma^2 \propto \left[ 1 + \frac{n(\mu - \bar{y})}{(n-1)s^2} \right]^{-n/2} \)

  * \( \mu | y \sim t_{n-1}(\bar{y}, s^2/n) \) (a t-distn)
  * recall traditional result \( \frac{\bar{y} - \mu}{s/\sqrt{n}} | \mu, \sigma^2 \sim t_{n-1} \)
    (note result doesn’t depend at all on \( \sigma^2 \) )
Multiparameters Models
Normal example (cont’d)

- Further examination of joint posterior distribution

\[
p(\mu, \sigma^2|y) \propto \left(\frac{1}{\sigma^2}\right)^{\tilde{y}+1} \exp\left[-\frac{1}{2\sigma^2} \left(\sum_i (y_i - \tilde{y})^2 + n(\tilde{y} - \mu)^2\right)\right]
\]
- consider $\tilde{y}$ a future draw from the same population
- what is the predictive distn of $\tilde{y}$, i.e., $p(\tilde{y}|y)$
- $p(\tilde{y}|y) = \int \int p(\tilde{y}|\mu, \sigma^2, y)p(\mu, \sigma^2|y)d\mu \ d\sigma^2$
- note first term in integral doesn’t depend on $y$ .... given params we know distn of $\tilde{y}$ is $N(\mu, \sigma^2)$
- predictive distn by simulation
  (simulate $\sigma^2 \sim \text{Inv-}\chi^2(n-1, s^2)$,
  then $\mu \sim N(\tilde{y}, \sigma^2/n)$, then $\tilde{y} \sim N(\mu, \sigma^2)$)
- predictive distn analytically (can proceed as for $\mu$ by first conditioning on $\sigma^2$)
  $\tilde{y}|y \sim t_{n-1}(\tilde{y}, (1 + \frac{1}{n})s^2)$
Multiparameters Models
Normal example - other prior distns (cont’d)

- Semi-conjugate analysis
  - for conjugate distn, the prior distn for \( \mu \)
    depends on scale parameter \( \sigma \) (unknown)
  - may want to allow info about \( \mu \) that does not
    depend on \( \sigma \)
  - consider independent prior distributions
    \( \sigma^2 \sim \text{Inv-}\chi^2(\nu_o, \sigma_o^2) \) and \( \mu \sim N(\mu_o, \tau_o^2) \)
  - may call this semi-conjugate
  - note that given \( \sigma^2 \), analysis for \( \mu \) is conjugate
    normal-normal case so that \( \mu | \sigma^2, y \sim N(\mu_n, \tau_n^2) \)
    with
    \[
    \mu_n = \frac{1}{\tau_o^2} \mu_o + \frac{n}{\sigma^2} \bar{y} \quad \text{and} \quad \tau_n^2 = \frac{1}{\tau_o^2 + \frac{n}{\sigma^2}}
    \]
Multiparameters Models
Normal example - other prior distns (cont’d)

• Semi-conjugate analysis (cont’d)
  
  – $p(\sigma^2|y)$ is not recognizable distn
    * calculate as
      $$p(\sigma^2|y) = \int \prod_{i=1}^{n} N(y_i|\mu, \sigma^2)N(\mu|\mu_o, \tau_o^2)\text{Inv-}\chi^2(\sigma^2|\nu_o, \sigma_o^2)d\mu$$
    * or calc $p(\sigma^2|y) = p(\mu, \sigma^2|y)/p(\mu|\sigma^2, y)$
      (RHS evaluated at convenient choice of $\mu$)
    * use a 1-dimensional grid approximation or some other simulation technique

• Multivariate normal case
  – no details here (see book)
  – discussion is almost identical to that for univariate normal distn with Inv-Wishart distn in place of the Inv-\chi^2
Multiparameters Models
Multinomial data

- Data distribution

\[ p(y|\theta) = \prod_{j=1}^{k} \theta_j^{y_j} \]

where \( \theta \) = vector of probabilities with \( \sum_{j=1}^{k} \theta_j = 1 \) and \( y \) = vector of counts with \( \sum_{j=1}^{k} y_j = n \)

- Conjugate prior distn is the Dirichlet(\( \alpha \)) distn (multivariate generalization of the beta distn)

\[ p(\theta) = \prod_{j=1}^{k} \theta_j^{\alpha_j-1} \]

for vectors \( \theta \) such that \( \sum_{j=1}^{k} \theta_j = 1 \) and \( \alpha > 0 \)

- \( \alpha = 1 \) yields uniform prior distn on \( \theta \) vectors such that \( \sum_j \theta_j = 1 \) (noninformative? ... favors uniform distn)

- \( \alpha = 0 \) uniform on log \( \theta \) (noninformative but improper)

- Posterior distn is Dirchlet(\( \alpha + y \))
Multiparameters Models
A non-standard example: logistic regression

• A toxicology study (Racine et al, 1986, Applied Statistics)

• $x_i = \log(\text{dose}), \ i = 1, \ldots, k$ ($k$ dose levels)

• $n_i = \text{animals given } i\text{th dose level}$

• $y_i = \text{number of deaths}$

• Goals:
  – traditional inference for parameters $\alpha, \beta$
  – special interest in inference for LD50 (dose at which expect 50% would die)
Multiparameters Models
Logistic regression (cont’d)

• Data model specification
  – within group (dose): exchangeable animals
    so model \( y_i | \theta_i \sim \text{Bin}(n_i, \theta_i) \)
  – between groups: non-exchangeable (higher dose
    means more deaths); many possible models
    including
    \[
    \text{logit}(\theta_i) = \log \left( \frac{\theta_i}{1 - \theta_i} \right) = \alpha + \beta x_i
    \]
  – resulting data model
    \[
    p(y | \alpha, \beta) = \prod_{i=1}^{k} \left( \frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}} \right)^{y_i} \left( \frac{1}{1 + e^{\alpha + \beta x_i}} \right)^{n_i - y_i}
    \]

• Prior distn
  – noninformative: \( p(\alpha, \beta) \propto 1 \) … is posterior distn
    proper?
  – answer is yes but it is not-trivial to show
  – should we restrict \( \beta > 0 \) ??
Multiparameters Models
Logistic regression example (cont’d)

• Posterior distn: \( p(\alpha, \beta | y) \propto p(y | \alpha, \beta) p(\alpha, \beta) \)

\[
p(\alpha, \beta | y) = \prod_{i=1}^{k} \left( \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right)^{y_i} \left( \frac{1}{1 + e^{\alpha+\beta x_i}} \right)^{n_i - y_i}
\]

• Grid approximation
  – obtain crude estimates of \( \alpha, \beta \)
    (perhaps by standard logistic regression)
  – define grid centered on crude estimates
  – evaluate posterior density on 2-dimensional grid
  – sample from discrete approximation
  – refine grid and repeat if necessary

• Grid approximations are risky (may miss important parts of distn)

• More sophisticated approaches will be developed later (MCMC)
Multiparameters Models
Logistic regression example (cont’d)

• Inference for LD50
  – want \( x_i \) such that \( \theta_i = 0.5 \)
  – turns out \( x_i = -\alpha/\beta \)
  – with simulation it is trivial to get posterior distn of \( -\alpha/\beta \)
  – note that using MLEs it would be easy to get estimate but hard to get standard error
  – doesn’t make sense to talk about LD50 if \( \beta < 0 \)
  .... could do inference in two steps
  * \( \Pr(\beta > 0) \)
  * distn of LD50 given \( \beta > 0 \)

• Real-data example (handout)