**Frequentist vs Bayes**: In frequentist, unknown model parameters are fixed and only estimable by replications of data from an experiment. In Bayesian, unknown parameters are random and have distributions like data, while data are fixed. Make inference on distribution of parameters conditional on observed data. Bayesian allows incorporating a guess even when we have no data available. **Bayesian Inference Process**: Write down a prior guess (π(θ)), add data, obtain a prior (p(θ|x)), perform inference by summarizing posterior. Posterior info >= prior info >= 0 (= 0 when prior is flat). **Why Bayesian**: Incorporate prior info. Expands class of models we fit that are awkward or infeasible from classical view. Ease of interpretation. Bayesian methods obtain posterior estimates by weighting the data and prior info and allow the data to dominate as *n* grows. Also allows for early stopπng – Bayesian Adaptive Design. **Confidence interval vs credible interval**: Frequentist 95%CI: If we generate 100 samples of size *n* and compute δ(x) for each sample, 95 of all obtained intervals will capture true value. Under this paradigm the interval is random and depends on (θ, σ^2). Uncertainty is quantified by repeated data sampling and reliant on future, uncollected data. In Bayesian, conditional on observed data, the probability of the interval covering the true value of θ is 0.95. To Bayesian, data sets which might have been, but were not, observed are irrelevant to making inferences about unknown parameters. **Hypothesis testing**: Frequentist testing uses P-values, where hypotheses μst be nested, p-value can only offer evidence against the null, the p-value is not the probability that H\_0 is true. Frequentist hypothesis testing could violate likelihood principle. In making inferences or decisions about θ after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. Two likelihood functions contain the same information about θ if they are proportional to each other as functions of θ. Bayesians always condition on the data observed and the long-run performance of a procedure is of (at most) a secondary interest. Using a prior, a Bayesian analysis returns the same posteriors for the binomial/neg binomial example. **Credible intervals:** Posterior quantiles: credible interval extends from alpha/2 posterior quantile to 1 – alpha/2 quantile. Or highest posterior density: region with posterior probability (1 – alpha), such that the density within the region is never lower than outside, which will be the narrowest (1 – alpha) credible interval. These intervals are not always the same, equal tail intervals do not work as well when the distribution is more than unimodal as there may be areas within the interval that have lower density than areas outside the interval. **Advantages to Bayesian Inference**: Provides an intuitive approach for specifying complex models. Ability to formally incorporate prior info. The reason for stopπng experimentation does not affect the inference. Intuitive interpretation of results. Inferences are conditional on actual data. Does not rely on asymptotics. **Disadvantages of Bayesian inference**: Can be dependent on prior (how do we control for inference of prior?). No direct connection with Type-1 error, which regulators care about (although many Bayesian procedures have good frequentist properties). Can be computational extensive. **Bayes Theorem**: or more generally . As applied to a statistical model: = . The marginal distribution of y is a function of y alone, so it’s called a normalizing constant and generally left out, with . **Single parameter model with known variance**: Consider a single observation y ~ N(y|θ, σ^2) with known σ. Prior on θ ~ N(θ | μ, τ^2) where μ and τ are hyperparameters. The posterior of θ is then p(θ|y) = N(θ | σ^2 / (σ^2 + τ^2) \* μ + τ^2 / (σ^2 + τ^2) \* y, σ^2 \* τ^2/(σ^2 + τ^2). We can write B = σ^2 / (σ^2 + τ^2) and note that 0 < B < 1. Then the E(θ|y) = B μ + (1 – B) \* y, a weighted average of prior mean and observed data value with weights determined by variances. **Bayes and sufficiency**: A statistic T(y) is sufficient for θ if the likelihood can be factored as f(y|θ) = h(y)g(T(y)|θ), so the posterior is proportional to g(T(y)|θ)π(θ). **What is a conjugate prior:** A conjugate prior is a prior such that the posterior will be the same form as the prior. These are computationally convenient but rarely possible in complex settings.  **Conjugates:** ***Beta-Binomial***: prior: beta(a, b), likelihood: binomial(N, θ), posterior: beta(a + x, b + N – x) [θ^(a + x – 1) (1 – θ)^(b + N – x – 1)]. With *a* ***Bernoulli*** likelihood (Bernoulli(θ)), posterior become beta(a +1, b) or beta(a, b + 1)[θ^a(1 – θ)^(b – 1) if x = 1, switch minus 1 on exponent if x = 0] . ***Geometric likelihood***: Geometric distribution pmf: θ^x(1 – θ). With beta prior, posterior becomes beta(a + x, b + 1)[ (θ^(a + x + 1)(1 – θ)^b]. ***Normal-Normal with variance known***: prior: N(μprior, σ^2prior), likelihood: N(θ, σ^2), posterior ~ N(μposterior, σ^2posterior). [exp(-(θ – μposterior)^2 / (2 \* σ^2posterior))]. Μposterior = (σ^-2prior \* μprior + σ^-2 \* x) / (σ^-2prior + σ^-2), σ^2posterior = 1 / (σ^-2prior + σ^-2). ***Normal-Normal with variance known and more than one obs****.*: μposterior / σ^2posterior = (μprior / σ^2prior) + n\*x\_bar / σ^2 and 1/σ^2posterior = 1/σ^prior + n/σ^2. ***Poisson-gamma***: Gamma(y + alpha, (2/beta)-1) [thetay + alpha – 1e –theta(2/beta), where the 2 may be replaced by T + 1 if the variable has some Poisson(lambda \* T). **Jeffrey’s Prior**: Non-informative prior, invariant under 1-1 transformations. Π(θ) I(θ)1/2, where I(θ) = Eθ[()2] = -E\_θ[()2]. May be improper for many models. **Point estimations**: Mode, mean, median. Mean tends to chase heavy tails, median is awkward to compute as it is . A Bayes point estimate is a weighted average of a common frequentist estimate and a parameter estimate obtained only from the prior distribution. Weight on the frequentist estimate tends to 1 as n increases to infinity. **Bayes Factor:** Is the posterior odds ratio to the prior odds ratio: . Large Bayes factor favor HA. It has limitations: not well-defined when the prior is improper and may be sensitive to choices of priors. Can use modified Bayes factor (partial BF, fractional BF) or use conditional predictive distribution f(yi|y(i)). Or use information criteria like AIC, BIC, or DIC. **Bayesian prediction**: The left hand side of that equation is the posterior predictive distribution. **How to specify a prior**: Objective and informative: historical data, πlot data, “today’s posterior is tomorrow’s prior”, noninformative: priors meant to express ignorance about the unknown parameters, conjugate: posterior and prior belong to the same distribution family. **Effective sample size**: The approximate number of subjects’ worth of data found in the prior. **Noninformative priors:** Meant to express ignorance about the unknown parameter or have minimal impact on the posterior distribution of θ. May be improper and this may lead to proper or improper posteriors. **Μulti-parameter models:** Want to estimate joint posterior, but really interested in marginal posteriors usually. 2 ways to specify: independent (π(θ1, θ2) = π1(θ1)π2(θ2)) or hierarchical (pi­1(theta1|theta2)pi2(theta2). If theta1 is of interest and theta2 is a nuisance parameter then the objective is or if the marginal posterior is not of a closed form: and then use a sampling method for approximation. At each iteration 1) draw from the marginal distribution of theta2 given y, and then draw from the conditional distribution of theta1 given y and theta2. Can then also draw a new y value on the conditional of the other 2 parameters. **Hierarchical models**: Intuitive and can be used to deal with complicated cases. Allow us to borrow strength across subgroups, resulting in more efficient estimation. The joint posterior of a hierarchical model p(theta1, theta2 | y) is proportional to f(y|theta1, theta2)pi(theta­­1|theta2)pi(theta2). The conditional posterior of theta1 p(theta1|theta2, y) = p(theta1, theta2 | y)/p(theta­­2|y) = f(y|theta1, theta2) pi(theta1|theta2)pi(theta2) / p(theta­2 |y). **Non-iterative Monte Carlo methods**: **Asymptotpic normal approximation**: With iid X and general likelihood conditional on theta and a prior pi(theta), as n -> infinity, then p(theta|x) ~ N(theta\_hatp, [IP(X)]-1) where theta\_hatp is the posterior mode obtained by solving with p\* as the unnormalized posterior. [IP(x)]-1 is the negative inverse Hessian of log p\*(theta|x) evaluated at the mode. **Importance Sampling**: Approximate E(f(theta)|y) = with a proposal density g(theta). Draw a sample theta from g(theta). E(f(theta|y)) is estimated by . A good match of p(theta|y) and g(theta) will produce roughly equal weights and a good approximation. Need a density equation (the mixture-Normal model from homework). Sample from a candidate density (a wide N distribution in homework with sd = sqrt(20)). Calculate weights by plugging the sample from the candidate density into the target density equation and divide that by the density for the samples from their own distribution. Use weighted mean of your samples, weighted by the weights we just calculated to get mean. **Rejection sampling**: Objective is to sample from our posterior. We use an envelope function g(theta) and constant M such that p\*(theta|y) < M \* g(theta) (our envelope function should have a higher density than our target distribution). Draw a sample from g(theta). Draw a sample from a uniform distribution (0, 1). Accept your draw if U is < p\*(thetaj|y)/Mg(thetaj). Resulting sample will be approximately p(theta|y). **Gibbs Sampling**: Find the full conditionals for all parameters (pi = thetai | theta1, … thetai – 1, y). Draw thetat1 from the full conditional for theta­1 with initial values. Then draw thetat2 from the full conditional of theta2 with initial values + the just drawn sample for theta1. Repeat for all theta. **Metropolis Hastings algorithm**: Draw theta\* from proposal (symmetric) density q(.|thetat -1). Compute alpha = p\*(theta\*|y)/p\*(thetat-1|y) (MH: p\*(theta\*|y) \* q(thetat-1|theta\*) / p\*(thetat-1|y) \* q(theta\*|thetat=1). Accept theta\* and set thetat = theta\* with probability min(alpha, 1). Reject theta\* and set thetat = thetat – 1 otherwise. From the homework, the unnormalized posterior of theta was the density function for the mixture model. This would be divided by M \* our candidate function which was a normal distribution with large SD.

**Example WinBugs code**

# WiNBUGS code for Normal-Normal Model #

###################################

model {

# Likelihood: y[i] ~ N(theta,sigma^2)

for(i in 1:n){

y[i] ~ dnorm(theta,prec.y)

}

prec.y <- 1/sigma2

# Prior: theta ~ N(mu,tau2)

theta ~ dnorm(mu,prec.theta)

prec.theta <- 1/tau2

}

# Data

list(

mu=2,sigma2=1,tau2=1,n=10,

y=c(7.2,5.0,6.7,4.6,6.4,5.5,3.6,6.5,7.9,6.6)

)

# Initial values

list(theta=0)

list(theta=6)

**Example JAGS code – hierarchical poisson model from homework**

#devtools::install.packages("jagsplot")

library(jagsplot)

poisson\_model <- function() {

# Priors

for (i in (1:2)) {

invb[i] ~ dgamma(1, 1)

b[i] <- 1 / invb[i]

theta[i] ~ dgamma(0.5, b[i])

}

# Likelihood

for (i in 1:n) {

y[i] ~ dpois(theta[group[i]])

}

R <- theta[1] / theta[2]

}

jags.data <- list(y = y,

n = n,

k = k,

group = group)

params <- c("theta", "R")

jags.sample <- jags.parallel(model = poisson\_model,

parameters.to.save = params,

data = jags.data,

n.chains = 3,

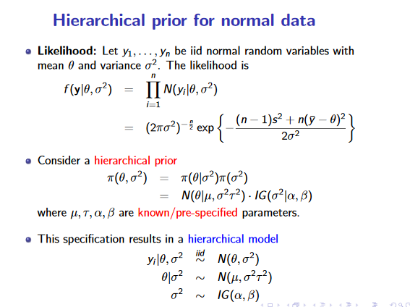
n.burnin = 1000,

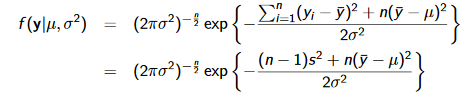
n.iter = 10000)

**Deriving joint posterior when mu and sigma are unknown**

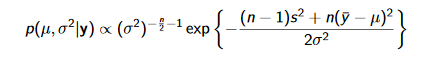
Consider improper priors

****

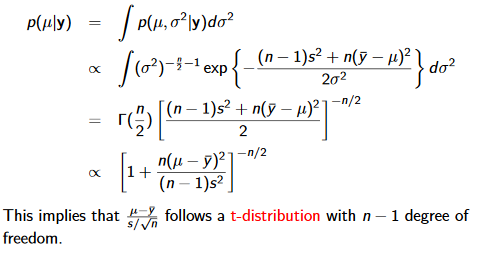
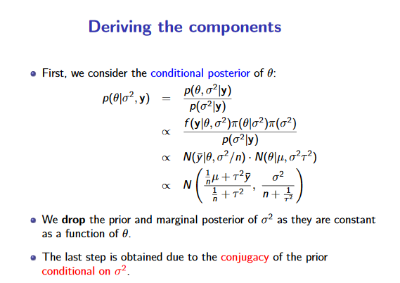
Likelihood can be rewritten using sufficient statistics

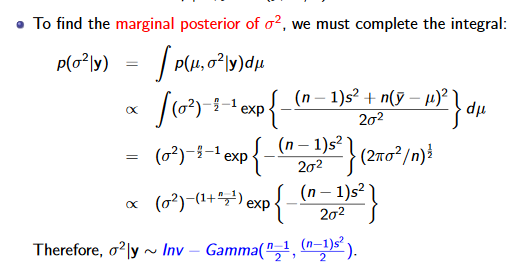
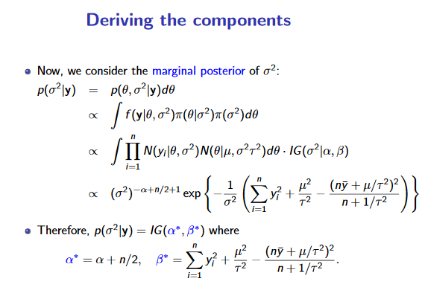


Joint posterior is therefore

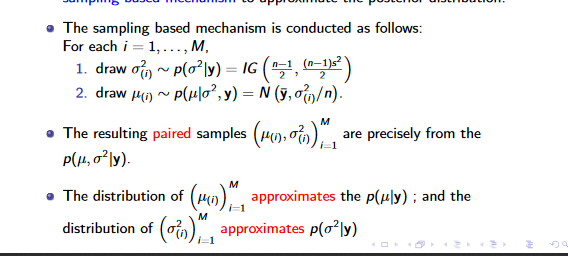
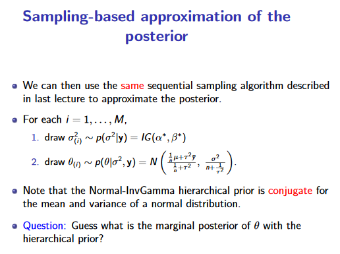


Marginal posterior:

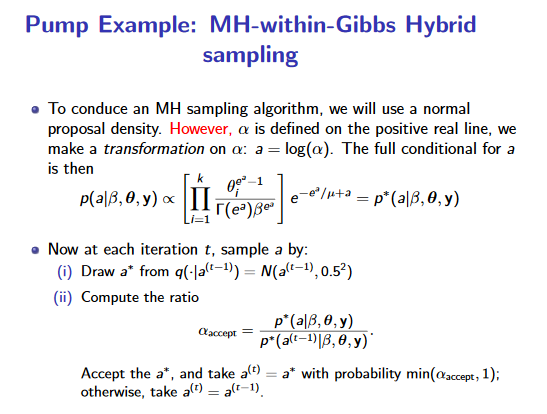
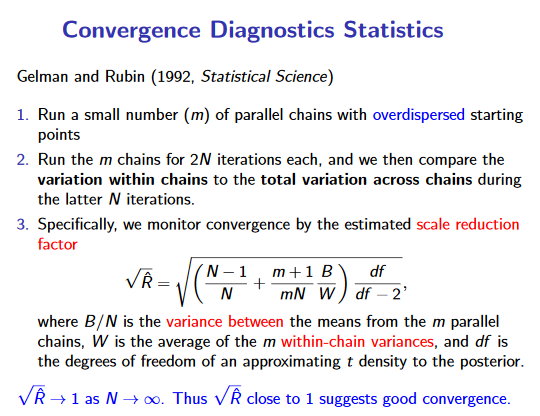
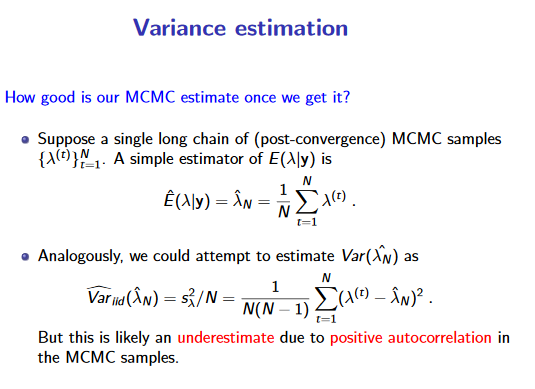
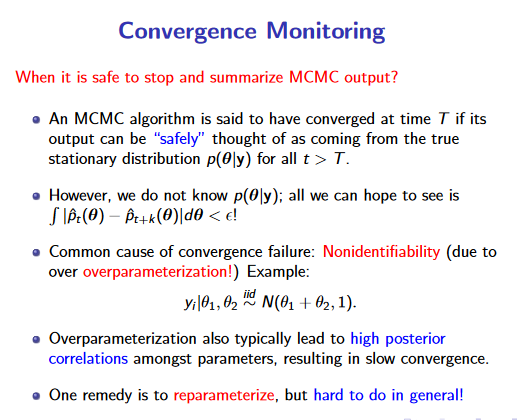
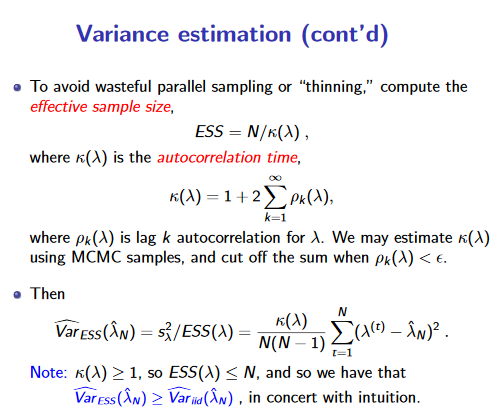
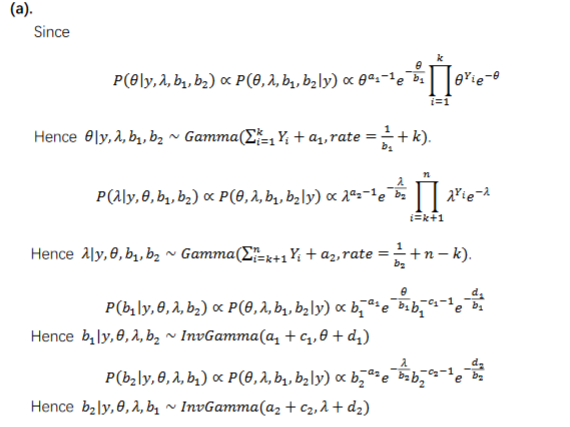


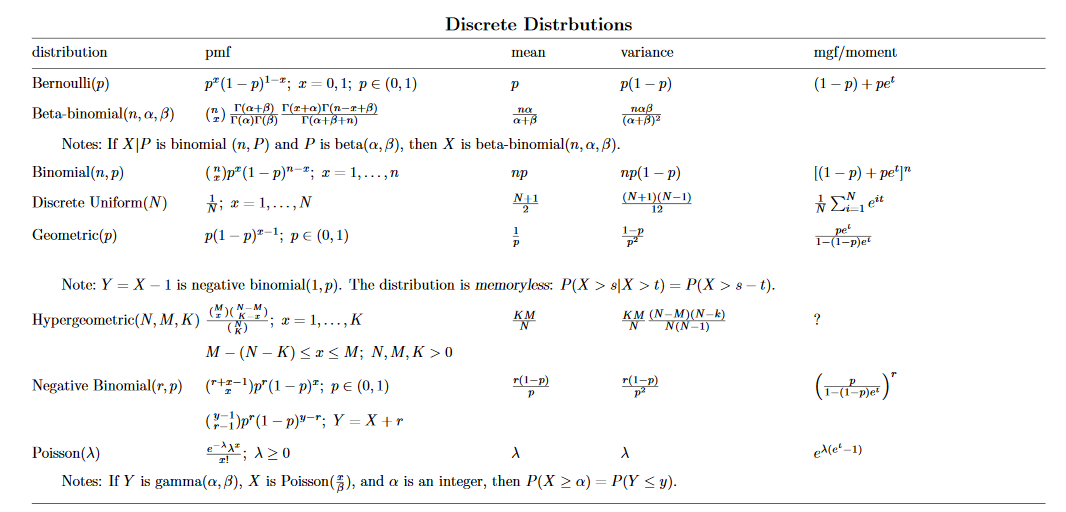


Sampling based approximation



Hierarchical model for normal example ->

Figure 1Answer to hw3question c

