Bayesian Hierarchical Modeling for the Social Sciences

Basic Mechanics and Normal Models

JEFF GILL
Department of Government
Department of Mathematics and Statistics
Center for Behavioral Neurosciences
American University

The Denominator From Bayes Law

- The "integrated likelihood" is the denominator of Bayes law calculated here by:

$$p(\mathbf{x}) = \int \underbrace{L(\theta|\mathbf{x})p(\theta)}_{\text{likelihood}\times\text{prior}} \; d\theta$$

- This is also called the "marginal likelihood," the "marginal probability of the data," or the "predictive probability of the data".
- Why do we treat this as a constant?
- This quantity is often ignored since it can be recovered later, but it is important in Bayesian model comparison.

Standard Bayesian Conventions

- Uncertainty always described with probability.
- The use of *precisions* rather than *variances*.
- Posterior description with quantiles.
- Required statement of all statistical assumptions.
- Much less emphasis on asymptotics.

Reporting Posterior Results

- Consider a single parameter θ and some generic (unspecified structure) data $\mathbf{D}$.
- Bayesians generally report $p(\theta|\mathbf{D})$ to readers via distributional summaries such as means, modes, quantiles, probabilities over regions, traditional-level probability intervals, and graphical displays.
- Once the posterior distribution has been calculated, everything about is known and it is entirely up to the researcher to highlight features of interest.
- Often it is convenient to report the posterior mean and variance in papers and reports since this is what non-Bayesians do by default.
- The posterior mean:

$$E[\theta|\mathbf{D}] = \int_{-\infty}^{\infty} \theta\pi(\theta|\mathbf{D})d\theta$$

Reporting Posterior Results

- The posterior variance:

$$
\begin{aligned}
Var[\theta|\mathbf{D}] &= E\left[(\theta - E[\theta|\mathbf{D}])^2|\mathbf{D}\right] \\
&= \int_{-\infty}^{\infty} (\theta - E[\theta|\mathbf{D}])^2 \pi(\theta|\mathbf{D}) d\theta \\
&= \int_{-\infty}^{\infty} \left(\theta^2 - 2\theta E[\theta|\mathbf{D}]) + E[\theta|\mathbf{D}]^2\right) \pi(\theta|\mathbf{D}) d\theta \\
&= \int_{-\infty}^{\infty} \theta^2 \pi(\theta|\mathbf{D}) d\theta - 2E[\theta|\mathbf{D}] \int_{-\infty}^{\infty} \theta \pi(\theta|\mathbf{D}) d\theta + \left(\int_{-\infty}^{\infty} \theta \pi(\theta|\mathbf{D}) d\theta\right)^2 \\
&= E[\theta^2|\mathbf{D}] - E[\theta|\mathbf{D}]^2
\end{aligned}
$$

Summarizing a Posterior

- Suppose we had data, $\mathbf{D}$, distributed $p(\mathbf{D}|\theta) = \theta e^{-\theta \mathbf{D}}$, which can be either a single scalar or a vector for our purposes.
- Thus $\mathbf{D}$ is exponentially distributed with the support $(0{:}\infty)$.
- The prior distribution for θ is $p(\theta) = 1$, where $\theta \in (0{:}\infty)$.
- The resulting posterior distribution is:

$$\pi(\theta|\mathbf{D}) \propto p(\theta)p(\mathbf{D}|\theta) = (1)\theta e^{-\theta \mathbf{D}} = \theta e^{-\theta \mathbf{D}}.$$

- This posterior distribution has mean:

$$E[\theta|\mathbf{D}] = \int_0^\infty (\theta) \left(\theta e^{-\theta \mathbf{D}}\right) d\theta = \frac{2}{\mathbf{D}^3},$$

  which is found easily with two iterations of integration-by-parts.

Summarizing a Posterior

- The expectation of $\theta^2|\mathbf{D}$ is:

$$E[\theta^2|\mathbf{D}] = \int_0^\infty (\theta^2)\left(\theta e^{-\theta\mathbf{D}}\right) d\theta = \frac{6}{\mathbf{D}^4},$$

  which is found with three iterations of integration-by-parts now.

- So the posterior variance is:

$$Var[\theta|\mathbf{D}] = E[\theta^2|\mathbf{D}] - E[\theta|\mathbf{D}]^2 = 6\mathbf{D}^{-4} - 4\mathbf{D}^{-6}.$$

Credible Intervals and Sets

- The Bayesian analogue to the confidence interval is the credible interval and more generally the credible set, which does not have to be contiguous.
- Most of the time in practice, it is calculated in *exactly the same way* as the confidence interval.
- For instance calculating a 95% credible interval under the Gaussian normal assumption means marching out 1.96 standard errors from the mean in either direction, just like the analogous confidence interval is created. (The difference lies in the interpretation.)
- A $100(1-\alpha)\%$ credible interval gives the region of the parameter space where the probability of covering $\boldsymbol{\theta}$ is at least $1-\alpha$.
- In contrast, applying this new definition to the confidence interval means that the probability of coverage is either zero or one, since it either covers the true $\boldsymbol{\theta}$ or it doesn't.

Credible Intervals and Sets

- Define C as a *contiguous* subset of the parameter space, $\boldsymbol{\Theta}$, such that a $100(1-\alpha)$ credible interval meets the condition:

$$1 - \alpha = \int_C \pi(\boldsymbol{\theta}|\mathbf{X}) d\theta$$

  for some chosen α level.

- Conventions: centered at mean or mode, equal tails.

- So credible intervals are *not* unique!

Credible Intervals and Sets, Example

- Suppose we have duration data, $\mathbf{X}$, exponentially distributed $p(X|\theta) = \theta e^{-\theta X}$ defined over $(0, \infty)$, where interest is in the posterior distribution of the unknown parameter θ.
- Specify the prior distribution of $p(\theta) = 1/\theta$, for $\theta \in (0{:}\infty)$.
- The posterior is:

$$\pi(\theta|\mathbf{X}) \propto p(\theta)L(\theta|\mathbf{X}) = \frac{1}{\theta}\theta^n \exp\left[-\theta \sum_{i=1}^{n} x_i\right] = \theta^{n-1} \exp\left[-\theta \sum_{i=1}^{n} x_i\right].$$

- This means that $\theta|\mathbf{X} \sim \mathcal{G}(\theta|n, \sum x_i)$, where putting the constants back in front to recover the full form of this gamma posterior distribution produces:

$$\pi(\theta|\mathbf{X}) = \frac{(\sum x_i)^n}{\Gamma(n)}\theta^{n-1} \exp\left[-\theta \sum x_i\right].$$

- Since we know everything about this posterior distribution, we are free to choose any desired credible interval.

Credible Intervals and Sets, Example

- Browne, Freidreis, and Gleiber (1986) tabulate complete cabinet duration for eleven Western European countries from 1945 to 1980:

Table 1: European Cabinet Duration Annualized, 1945-1980

Country	N	Average Duration
Austria	15	2.114
Belgium	27	1.234
Denmark	20	1.671
Finland	28	1.070
Iceland	15	2.080
Ireland	14	2.629
Italy	38	0.833
Netherlands	12	2.637
Norway	17	2.065
Sweden	15	2.274

Credible Intervals and Sets, Example

- Country averages from the third column of the table are weighted by N in the second column to reflect the number of such events: $\mathbf{X}_i N_i$.

- For a chosen α the end-points of an equal-tail credible interval can be calculated with:

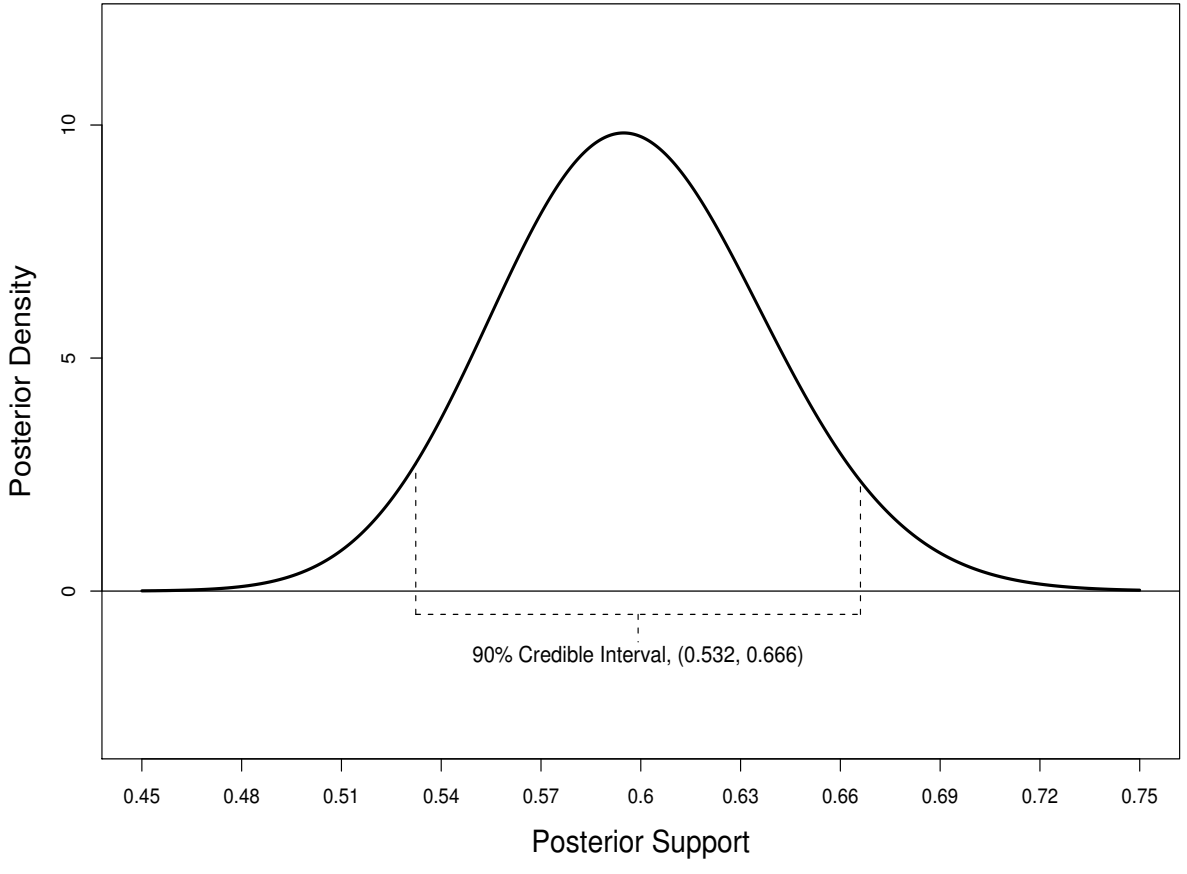
$$\frac{\alpha}{2} = \int_0^L \pi(\boldsymbol{\theta}|\mathbf{X})d\theta \qquad \frac{\alpha}{2} = \int_H^\infty \pi(\boldsymbol{\theta}|\mathbf{X})d\theta$$

  or we could simply use the following `R` commands for a 95% credible interval:

```
dur <- c(2.114,1.234,1.671,1.070,2.168,2.080, 2.629,0.833,2.637,2.065,2.274)
N <- c(15,27,20,28,15,15,14,38,12,17,15)
qgamma(0.025,shape=sum(N),rate=sum(N*dur))
[1] 0.52056
qgamma(0.975,shape=sum(N),rate=sum(N*dur))
[1] 0.67988
```

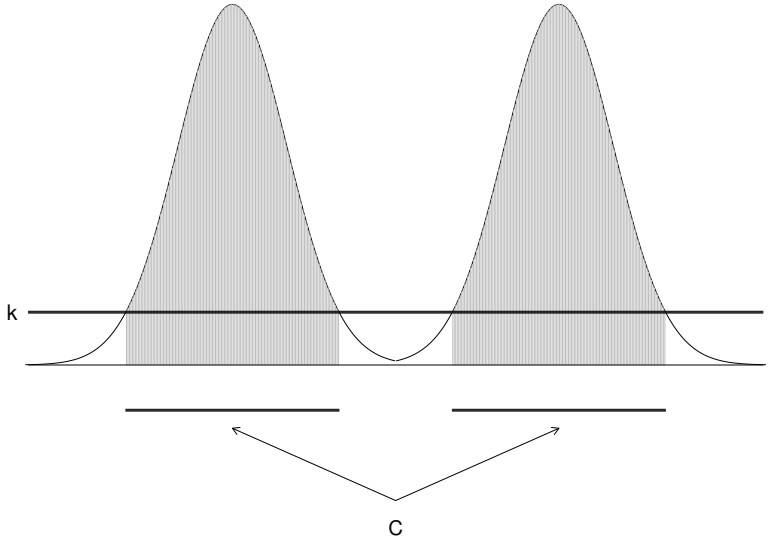
Credible Intervals and Sets, Example

Equal Tail Credible Interval for Cabinet Duration



Highest Posterior Density Intervals and Sets

- When looking at posterior distributions, we really care where the highest density exists on the support of the posterior density, regardless of whether it is contiguous or not.
- HPD created such that no region outside of the interval will have higher posterior density than any region inside the HPD.
- Therefore HPDs are not necessarily contiguous.



Highest Posterior Density Intervals and Sets

- A $100(1-\alpha)\%$ highest posterior density (HPD) is the subset of the support of the posterior distribution for some parameter, θ, that meets the criteria:

$$C = \{\theta : \pi(\theta|\mathbf{x}) \geq k\},$$

  where k is the largest number such that:

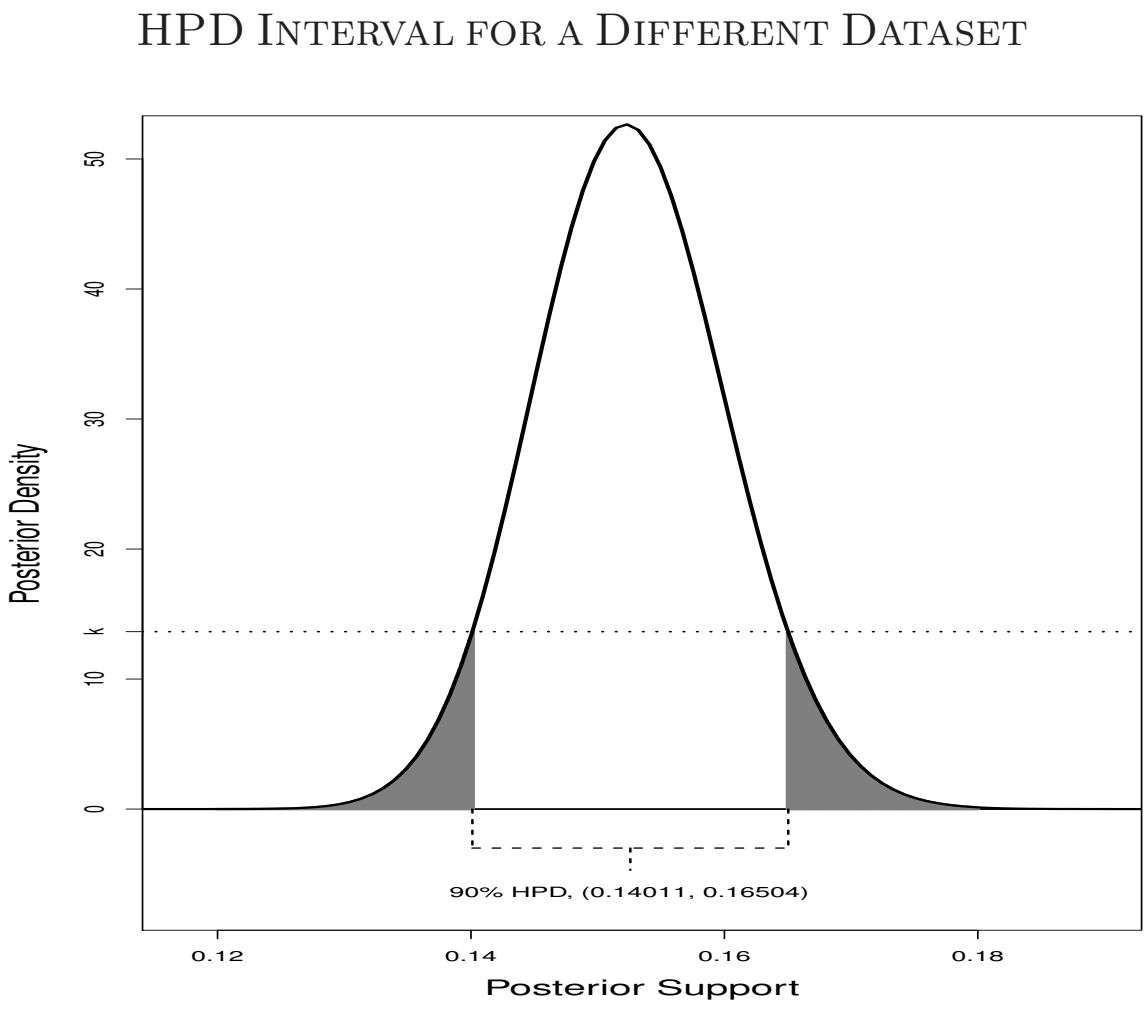
$$1 - \alpha = \int_{\theta:\pi(\theta|\mathbf{x})>k} \pi(\theta|\mathbf{x})d\theta$$

- The important difference is $\theta \ : \ \pi(\theta|\mathbf{x}) \ > \ k$ instead of a single contiguous interval as with the credible interval.

- Sometimes this can be done analytically.

- The `R` code for this example is in Chapter 2.

Highest Posterior Density Intervals and Sets, Example



Data Exercise 1: HPD Calculations

- The HPD region is constructed with the following in way by starting at the posterior mode, then incrementing a horizontal line down vertically until the separation between the higher density and lower density regions reflects the desired coverage.
- So for each value of k, the level on the y-axis, we separately sum the area inside and outside the coverage area, regardless of contiguity.
- This process was quite easy to implement in `R` since we know the exact form of the gamma distribution from the model.
- Later when estimating marginal posterior forms with Bayesian stochastic simulation (MCMC), we will see that there are similarly easy ways to make this calculation even when we do not have an exact parametric description of the posterior distribution.
- There is also the `HPDinterval` function in the `CODA` library for this, but it does not illustrate the underlying theory as directly as the exposition here.

Data Exercise 1: HPD Calculations

http://jeffgill.org/files/jeffgill/files/hpd.gamma_.r.txt}

```
hpd.gamma <- function(data.df,g.shape,g.rate,target=0.90,steps=300,tol=0.01) {
   if (steps %% 2 == 1) steps <- steps + 1
   g.mode  <- sum(data.df$N)/sum(data.df$N*data.df$dur)
   g.range <- seq(qgamma(0.001,g.shape,g.rate), qgamma(0.999,g.shape,
              g.rate),length=steps)
   g.range <- c(g.range[1:(steps/2)],g.mode,g.range[(steps/2+1):steps])
   g.dens  <- dgamma(g.range,g.shape,g.rate)
   g.probs <- pgamma(g.range,g.shape,g.rate)
   for (i in 1:(steps/2)) {
      k.dir <- which(c(g.dens[(steps/2-i)],g.dens[(steps/2+i)]) ==
                max(g.dens[(steps/2-i)],g.dens[(steps/2+i)]))
      k <- c(g.dens[(steps/2-i)],g.dens[(steps/2+i)])[k.dir]
      k.loc <- c((steps/2-i),(steps/2+i))[k.dir]
      if (k.dir == 2) k2.range <- c(1:(steps/2))
      else k2.range <- c((steps/2 + 1):steps)
      k2.min <- which(abs(k-g.dens[k2.range])==min(abs(k-g.dens[k2.range])))
      if (k.dir == 1) k2.min <- k2.min + steps/2
      if (g.probs[k.loc] + (1-g.probs[k2.min]) < 1-target)  break
      bounds <- c(g.range[k.loc],g.range[k2.min])
   }
   return(list("cdf.vals"=c(g.probs[k.loc],g.probs[k2.min]),
           "bounds"=bounds,"k"=k))
}
```

Data Exercise 1: HPD Calculations

```
state.df <-
  read.table("http://jeffgill.org/files/jeffgill/files/stat.short_.data_.txt",
  header=TRUE)
state.df <- state.df[-35,] # NO DATA FOR OHIO
state.hpd <- hpd.gamma(state.df,g.shape=sum(state.df$N),
          g.rate=sum(state.df$N*state.df$dur))
```

- Assignment: rerun with the code a with a different alpha level.
- Plot the HPD interval (many ways to do this).
- Optional: modify for the normal PDF, find some normal data, run.

Bayesian Updating: Overview

- Start with the prior distribution: $p(\theta)$, on an unknown variable θ.
- Observe a first set of iid data, $\mathbf{x}_1$, and calculate the posterior: $\boxed{\pi_1(\theta|\mathbf{x}_1) \propto p(\theta)L(\mathbf{x}_1|\theta).}$
- Now observe a second set of iid data, $\mathbf{x}_2$ from the same data-generating process and *update* the posterior and therefore improve our state of knowledge by treating the previous posterior as a prior:

$$\pi_2(\theta|\mathbf{x}_1, \mathbf{x}_2) \propto \pi_1(\theta|\mathbf{x}_1)L(\mathbf{x}_2|\theta) = p(\theta)L(\mathbf{x}_1|\theta)L(\mathbf{x}_2|\theta) = p(\theta)L(\mathbf{x}_1, \mathbf{x}_2|\theta).$$

- This is exactly the same result we would have obtained if all the data had arrived at once: $\pi(\theta|\mathbf{x})$.
- This process can continue *ad infinitum* and the model will continue to update the posterior conclusions as new information continues to roll in.

Bayesian Normal Models

- Why Be Normal?
  - A great deal of standard theory is based on normal assumptions.
  - Nature loves the normal: CLT.
  - Even non-normal data can often be modeled with normals.
  - Mixtures of normals are extremely flexible.
- Bayesian Normal Models
  - Easy.
  - Have good frequentist properties.
  - Lead directly to the Bayesian linear regression model (Lindley & Smith 1972).
  - Today: conjugacy mania.

Bayesian Normal Models, Mean Unknown, Variance Known

- Assume that the data are iid with unknown mean μ and known variance σ_0^2:

$$X|\mu,\sigma_0^2 \sim \mathcal{N}(\mu,\sigma_0^2) = (2\pi\sigma_0^2)^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma_0^2}(X-\mu)^2\right]$$
$$-\infty < \mu < \infty, \sigma_0^2 \text{ known}$$

- and specify a normal prior distribution for μ:

$$\mu|m,s^2 \sim \mathcal{N}(m,s^2)$$
$$= (2\pi s^2)^{-\frac{1}{2}}\exp\left[-\frac{1}{2s^2}(\mu-m)^2\right]$$
$$m, s \text{ given.}$$

Bayesian Normal Models, Mean Unknown, Variance Known

- Posterior Calculation:

$$\pi(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu)$$

$$\propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma_0^2}(x_i-\mu)^2\right] \exp\left[-\frac{1}{2s^2}(\mu-m)^2\right]$$

$$= \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{s^2}(\mu-m)^2\right)\right].$$

- Now expand the two squares.

$$\pi(\mu|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i^2-2x_i\mu+\mu^2) + \frac{1}{s^2}(\mu^2-2\mu m+m^2)\right)\right]$$

Bayesian Normal Models, Mean Unknown, Variance Known

- Continue with the expansion...

$$\pi(\mu|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_0^2}\frac{s^2}{s^2}\sum_{i=1}^{n}(x_i^2 - 2x_i\mu + \mu^2) + \frac{1}{s^2}\frac{\sigma_0^2}{\sigma_0^2}(\mu^2 - 2\mu m + m^2)\right)\right]$$

$$= \exp\left[-\frac{1}{2}\frac{1}{\sigma_0^2 s^2}\left(s^2\sum_{i=1}^{n}x_i^2 - 2s^2\mu n\bar{x} + n\mu^2 s^2 + \sigma_0^2\mu^2 - 2\sigma_0^2\mu m + \sigma_0^2 m^2\right)\right]$$

and gather by order of μ...

$$= \exp\left[-\frac{1}{2}\frac{1}{\sigma_0^2 s^2}\left(\mu^2(\sigma_0^2 + ns^2) - 2\mu(m\sigma_0^2 + s^2 n\bar{x}) + \underbrace{(m^2\sigma_0^2 + s^2\sum_{i=1}^{n}x_i^2)}_{k}\right)\right].$$

The last term in the expansion can be treated as part of the normalizing constant.

Bayesian Normal Models, Mean Unknown, Variance Known

- Rearrange into a Normal Form:

$$\pi(\mu|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left(\mu^2\left(\frac{1}{s^2}+\frac{n}{\sigma_0^2}\right)-2\mu\left(\frac{m}{s^2}+\frac{n\bar{x}}{\sigma_0^2}\right)+k\right)\right]$$

$$= \exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{n}{\sigma_0^2}\right)\left(\mu^2\frac{\left(\frac{1}{s^2}+\frac{n}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{n}{\sigma_0^2}\right)}-2\mu\frac{\left(\frac{m}{s^2}+\frac{n\bar{x}}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{n}{\sigma_0^2}\right)}+k\right)\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{n}{\sigma_0^2}\right)\left(\mu-\frac{\left(\frac{m}{s^2}+\frac{n\bar{x}}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{n}{\sigma_0^2}\right)}\right)^2\right].$$

Bayesian Normal Models, Mean Unknown, Variance Known, Results

- Therefore the point estimate of the mean is:

$$\hat{\mu} = \left(\frac{m}{s^2} + \frac{n\bar{x}}{\sigma_0^2}\right) \Bigg/ \left(\frac{1}{s^2} + \frac{n}{\sigma_0^2}\right),$$

- and the variance is:

$$\hat{\sigma}_\mu^2 = \left(\frac{1}{s^2} + \frac{n}{\sigma_0^2}\right)^{-1}.$$

- Notice that the posterior mean depends on the data only through $\bar{x}$ (the *sufficient statistic*).

- Proportionality and later normalizing with k made things much easier.

Bayesian Normal Models, Mean Unknown, Variance Known, Precisions

- $\frac{1}{s^2}$ is the *prior precision*
- $\frac{n}{\sigma_0^2}$ is the *data precision*
- and the *posterior precision* is the sum of these:

$$\frac{1}{\hat{\sigma}_\mu^2} = \frac{1}{s^2} + \frac{n}{\sigma_0^2}$$

- Note what happens as the data size increases for fixed σ_0^2 (this is why precisions are convenient for Bayesians).

Bayesian Normal Models, Mean Unknown, Variance Known, Asymptotics

- The posterior mean estimate:

$$\lim_{n\to\infty}\hat{\mu} = \lim_{n\to\infty}\frac{\frac{m}{s^2}+\frac{n\bar{x}}{\sigma_0^2}}{\frac{1}{s^2}+\frac{n}{\sigma_0^2}} = \lim_{n\to\infty}\frac{\frac{m\sigma_0^2}{ns^2}+\bar{x}}{\frac{\sigma_0^2}{ns^2}+1} = \bar{x},$$

- The posterior variance of the mean estimate (not the variance of the data):

$$\lim_{n\to\infty}\hat{\sigma_\mu^2} = \lim_{n\to\infty}\frac{1}{\frac{1}{s^2}+\frac{n}{\sigma_0^2}} = \lim_{n\to\infty}\frac{\sigma_0^2}{\frac{\sigma_0^2}{s^2}+n} = \frac{\sigma_0^2}{n}.$$

- Keep in mind that $\hat{\sigma_\mu^2}$ is the variance of the posterior of μ not the posterior of σ^2.

Bayesian Normal Models, Mean Known, Variance Unknown

- Now assume:

$$p(X|\mu_0, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(X-\mu_0)^2\right].$$

- The corresponding likelihood function is:

$$L(\sigma^2|\mathbf{x}) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{n}{2\sigma^2}\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2\right)}_{\text{sufficient statistic}}\right].$$

- Relabel the sufficient statistic for σ^2 as a convenience:

$$\tilde{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2$$

giving the simplified form:

$$L(\sigma^2|\mathbf{x}) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{n}{2\sigma^2}\tilde{x}\right].$$

Bayesian Normal Models, Mean Known, Variance Unknown

- Assign an inverse gamma prior for σ^2:

$$\mathcal{IG}(\sigma^2|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)}\exp[-\beta/\sigma^2]$$

$$\text{where:} \quad \sigma^2 > 0, \alpha > 0, \beta > 0.$$

- This has some moment limitations as well:

$$E[\sigma^2] = \frac{\beta}{\alpha-1}, \qquad \alpha > 1$$

$$Var[\sigma^2] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \qquad \alpha > 2.$$

Bayesian Normal Models, Mean Known, Variance Unknown

- Posterior calculation:

$$\pi(\sigma^2|\mathbf{x}) \propto L(\sigma^2|\mathbf{x})p(\sigma^2|\alpha,\beta)$$

$$= (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{n}{2\sigma^2}\tilde{x}\right] \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)} \exp[-\beta/\sigma^2]$$

$$\propto (\sigma^2)^{-((\alpha+\frac{n}{2})+1)} \exp\left[-\left(\beta+\frac{n}{2}\tilde{x}\right)/\sigma^2\right].$$

- Which actually gives the kernel of a different inverse gamma PDF:

$$\sigma^2|\mathbf{x} \sim \mathcal{IG}(\alpha+\frac{n}{2}, \beta+\frac{n}{2}\tilde{x}).$$

Multivariate Normal Model, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ Both Unknown

- The most realistic in this family and therefore worthy of considerable attention here.
- The conjugate prior specification for the mean has the same added complexity as before: it must be specified with a dependency the variance: $p(\mu|\sigma^2)$.
- If this is unrealistic, then a nonconjugate prior should be specified.
- For the multivariate case assume:
  - each of the n $\mathbf{X}$ rows is a k-dimensional vector representing a single case,
  - so now $\boldsymbol{\mu}$ is a vector and $\boldsymbol{\Sigma}$ is a matrix, both to be estimated.
  - From the PDF of the multivariate normal, the likelihood function can be expressed and manipulated as follows. . .

Both Unknown, Looking at the Likelihood Function

- Since:

$$\left(\sum_{i=1}^{n}(\mathbf{x}_i'\mathbf{x}_i) - n\bar{\mathbf{x}}'\bar{\mathbf{x}}\right) = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}}) \equiv S^2,$$

  then $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X})$ is a function of the data only through the two-component sufficient statistic: $[\bar{\mathbf{x}}, S^2]$, simplifying the likelihood:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = |\boldsymbol{\Sigma}|^{-n/2}\exp\left[-\frac{1}{2}\left(\text{tr}(\boldsymbol{\Sigma}^{-1})S^2 + n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})\right)\right].$$

- The conjugate priors for this setup are:

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathcal{N}_k\left(\mathbf{m}, \frac{\boldsymbol{\Sigma}}{n_0}\right), \qquad \boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\alpha, \boldsymbol{\beta}),$$

  where $\mathcal{W}()$ denotes the Wishart distribution, which is a multivariate generalization of the gamma PDF (an obvious choice for modeling multivariate variances).

Both Unknown (cont.)

- Wishart Form:

$$\mathcal{W}(\boldsymbol{\Sigma}^{-1}|\alpha,\boldsymbol{\beta}) = \frac{|\boldsymbol{\Sigma}^{-1}|^{(\alpha-(k+1))/2}}{\Gamma_k(\alpha)|\boldsymbol{\beta}|^{\alpha/2}} \exp[-\text{tr}(\boldsymbol{\beta}^{-1}\boldsymbol{\Sigma}^{-1})/2]$$

$$\text{where: } \Gamma_k(\alpha) = 2^{\alpha k/2}\pi^{k(k-1)/4}\prod_{i=1}^{k}\Gamma\left(\frac{\alpha+1-i}{2}\right), \quad 2\alpha > k-1, \quad \text{and } \boldsymbol{\beta} \text{ nonsingular.}$$

  where the term $\Gamma_k(\alpha)$ is the k-dimensional generalized gamma function, and is ignorable except for normalizing considerations.

- The parameter n_0 here is not a prior sample size; it is intended to be a reflection of prior precision relative to the sample size that is tunable by the researcher to reflect prior confidence in representability.

- The smaller the ratio n_0/n, the less weight on the prior, and therefore the closer the results will be closer to classical results.

- For additional mathematical details, see:
  `http://www.tc.umn.edu/~nydic001/docs/unpubs/Wishart_Distribution.pdf`.

Both Unknown (cont.)

- The resulting marginal posteriors are produced by taking integrals (reasonable agony involved):

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathcal{N}_k\left(\frac{n_0\mathbf{m} + n\bar{\mathbf{x}}}{n_0 + n}, \frac{\boldsymbol{\Sigma}}{n_0 + n}\right)$$

$$\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}_k\left(\alpha + n, \boldsymbol{\beta}^{-1} + S^2 + \frac{n_0 n}{n_0 + n}(\bar{x} - \mathbf{m})(\bar{x} - \mathbf{m})'\right).$$

- Note that the dependency exists here in the multivariate case as well.

Example: Variance Estimation with Public Health Data

- Consider data from the 2000 U.S. census and North Carolina public records (North Carolina Division of Public Health, Women's and Children's Health Section in Conjunction with State Center for Health Statistics).
- Each case is one of 100 North Carolina counties, and we will use only the following subset of the variables.
- `Substantiated.Abuse`: within family documented abuse for the county.
- `Percent.Poverty`: percent within the county living in poverty, U.S. definition (`http://www.census.gov/hhes/www/poverty/threshld/thresh98.html`).
- `Total.Population`: county population/1000.
- Each $\mathbf{X}$ row is a k-dimensional (3 here) vector representing a single case, distributed $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Example: Variance Estimation with Public Health Data (cont.)

- Relatively uninformed, $\alpha = 3$, $\mathbf{m} = (250, 16, 88)$, $n_0 = 0.01$, $\boldsymbol{\beta}$ a diagonal matrix w/100:

$\boldsymbol{\mu}$ Quantile	Abuse	%Poverty	Population
0.01	195.8976	14.2399	77.9827
0.25	199.6618	14.3123	79.7873
0.50	201.2110	14.3409	80.5230
0.75	202.7294	14.3698	81.2590
0.99	206.4080	14.4400	83.0124

$$\bar{\boldsymbol{\Sigma}} = \begin{bmatrix} 531.553969 & -3.2723672 & 200.207935 \\ -3.272367 & 0.1870651 & -1.672702 \\ 200.207935 & -1.6727021 & 117.901661 \end{bmatrix}$$

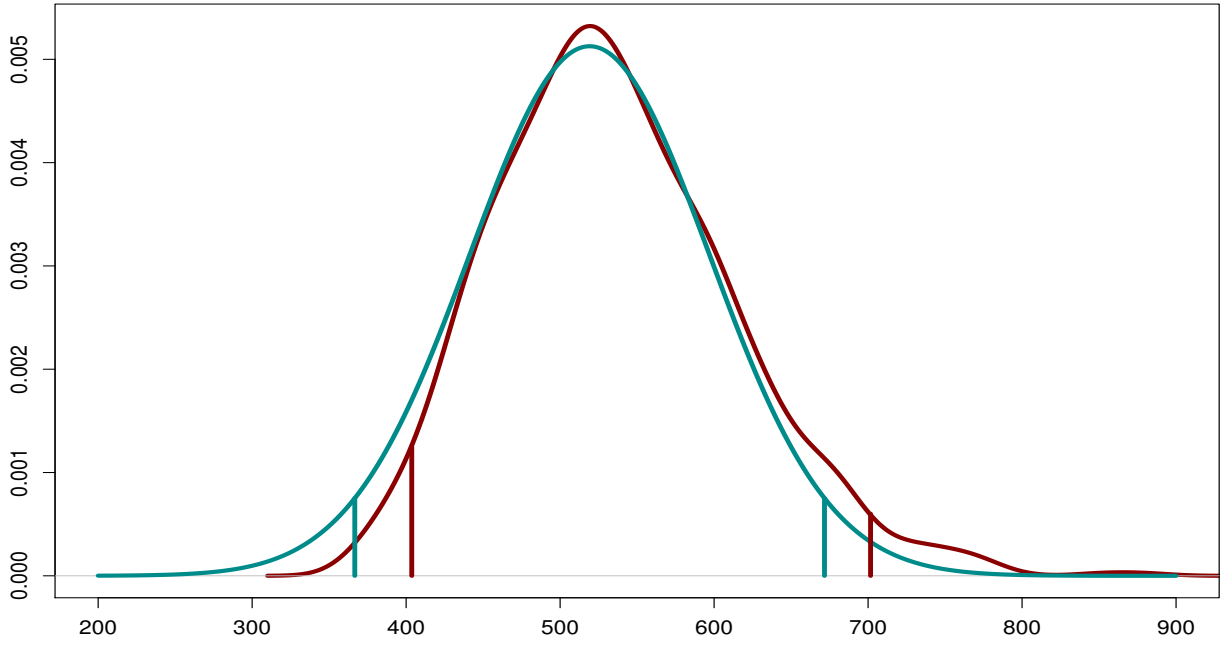
Example: Variance Estimation with Public Health Data (cont.)

- Now add strong priors, $\alpha = 3$, $\mathbf{m} = (100, 6, 88)$, $n_0 = 99$, $\boldsymbol{\beta}$ a diagonal matrix w/10:

$\boldsymbol{\mu}$ Quantile	Abuse	%Poverty	Population
0.01	138.4181	9.190786	82.33058
0.25	147.1816	9.902820	83.64427
0.50	150.7495	10.187523	84.19891
0.75	154.3365	10.478351	84.79181
0.99	163.0994	11.159200	86.25384

$$\bar{\boldsymbol{\Sigma}} = \begin{bmatrix} 5678.6595 & 421.23489 & -181.05113 \\ 421.2349 & 35.20976 & -33.15966 \\ -181.0511 & -33.15966 & 146.30970 \end{bmatrix}$$

Comparing the Posterior Distributions for the $\boldsymbol{\Sigma}[1, 1]$ Parameter



Note: green line for the likelihood, and red line for the posterior with uninformed prior parameter values.

R Code for the Example

```
nc.sub.df <- read.table("http://jeffgill.org/files/jeffgill/files/nc.sub_.dat_.txt",
  header=TRUE)
library(bayesm); library(BaM)   # FOR THE rwishart AND rmultinorm FUNCTION

Alpha <- 3 + nrow(nc.sub.df)
Beta.inv <- solve(diag(3)*100)
m <- c(250,16,88)
n0 <- 0.01
x.bar <- apply(nc.sub.df,2,mean)
S.sq <- var(nc.sub.df)

k <- (n0 * nrow(nc.sub.df))/(n0 + nrow(nc.sub.df))
p.Beta <- solve( Beta.inv + S.sq + k * round((x.bar-m) %*% t(x.bar-m),2) )
Sigma <- array(NA,dim=c(3,3,1))
for (i in 1:10000) Sigma <- array(c(Sigma,rwishart(Alpha,p.Beta)$IW),dim=c(3,3,(i+1)))
Sigma <- Sigma[,,-1]
```

R Code for the Example

```
Sigma.Mean <- apply(Sigma,c(1,2),mean)
     [,1]    [,2]    [,3]
[1,] 531.553969 -3.2723672 200.207935
[2,] -3.272367 0.1870651 -1.672702
[3,] 200.207935 -1.6727021 117.901661

# ANALYTICAL MEAN OF THE INVERSE WISHART:
( (Alpha-ncol(nc.sub.df)-1)^(-1) )*solve(p.Beta)
         Substantiated.Abuse Percent.Poverty
Substantiated.Abuse     531.736988   -3.2745226
Percent.Poverty        -3.274523    0.1872254
Total.Population       200.408410   -1.6760040
         Total.Population
Substantiated.Abuse    200.408410
Percent.Poverty      -1.676004
Total.Population     118.023667
```

R Code for the Example

```
Sigma.SD <- apply(Sigma,c(1,2),sd)
     [,1]     [,2]     [,3]
[1,] 75.510375 1.04926933 32.2891887
[2,]  1.049269 0.02689763  0.5020452
[3,] 32.289189 0.50204522 16.8975802

# VECTOR MEAN BY SIMULATION
Mu <- rmultinorm(5000,(n0*m + nrow(nc.sub.df)*x.bar)/(n0 + nrow(nc.sub.df)),
  Sigma.Mean/(n0+nrow(nc.sub.df)))
apply(Mu,2,quantile, probs = c(0.01,0.25,0.50,0.75,0.99))
    [,1]   [,2]   [,3]
1%  195.8976 14.23990 77.98269
25% 199.6618 14.31230 79.78725
50% 201.2110 14.34090 80.52302
75% 202.7294 14.36977 81.25900
99% 206.4080 14.44002 83.01237
```

Data Exercise 2: Normal Model Code

- Rerun the Variance Estimation code using different priors that you select.
- Are you able to noticeably change the posterior results with different priors?

General Comments on Uninformative Priors (more later)

- Somewhat antithetical to the Bayesian principle.
- Uninformative priors are never really totally "uninformative" since every specified prior has information.
- Usually mathematically more difficult, but an easier "sell."
- Current trends: mildly informed priors, nonparametric priors.
- Warning #1: it is possible to specify a uninformative prior such that posterior credible regions end up with pathological properties such as $P(C|\mathbf{X})$ being dissimilar than $P(C|\theta)$ for all θ (Bernardo and Smith 1994).
- Warning #2: it is possible to specify an improper prior such that the posterior distribution is also improper (Hobert and Casella 1998).

Bayesian Normal Models, Uninformative Priors

- The posterior distribution of the mean parameter is:

$$\pi(\mu|\mathbf{x}) \propto \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{1}{((\pi(n-1))^{\frac{1}{2}}} \left(\frac{n}{s^2}\right)^{\frac{1}{2}} \left(1 + \frac{1}{n-1}\left(\frac{\mu-\bar{x}}{s/\sqrt{n}}\right)^2\right)^{-\frac{1}{2}n}$$

- Therefore the marginal posterior of $\frac{\mu-\bar{x}}{s/\sqrt{n}}$ is student's-t with $\theta = n-1$ degrees of freedom, so the marginal posterior of μ is also student's-t with non-centrality parameter $\bar{x}$.

- Now obtain the marginal posterior of σ by dividing the joint posterior by the conditional distribution of μ assuming that σ is known.

$$\pi(\sigma|\mathbf{x}) = \frac{\left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \frac{\left(\frac{(n-1)s^2}{2}\right)^{\frac{n-1}{2}}}{\frac{1}{2}\Gamma\left(\frac{n-1}{2}\right)} \sigma^{-(n+1)} \exp\left[-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\mu-\bar{x})^2\right)\right]}{\sqrt{n}(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right]}$$

$$\propto \sigma^{-((n-1)+1)} \exp\left[-\frac{1}{2}(n-1)s^2/\sigma^2\right].$$

- So the marginal posterior of σ^2 is distributed $\mathcal{IG}((n-2)/2, (n-1)s^2/2)$.

Bayesian Normal Models, IQ Example

- IQ tests are purported to be biased towards Western Europeans and North Americans given their wording and structure.
- Question: is there evidence of economic and cultural biases in national level aggregation of IQ scores.
- The test is designed to have a mean response of 100 with a standard deviation of 15 (the Stanford-Binet version has a standard deviation of 16).

Bayesian Normal Models, IQ Example (cont.)

- Consider recently collected IQ data (Lynn & Vanhanen 2001) for 81 countries.

Argentina	96	Australia	98	Austria	102	Barbados	78
Belgium	100	Brazil	87	Bulgaria	93	Canada	97
China	100	Congo (Br.)	73	Congo (Zr.)	65	Croatia	90
Cuba	85	Czech Repub.	97	Denmark	98	Ecuador	80
Egypt	83	Eq. Guinea	59	Ethiopia	63	Fiji	84
Finland	97	France	98	Germany	102	Ghana	71
Greece	92	Guatemala	79	Guinea	66	Hong Kong	107
Hungary	99	India	81	Indonesia	89	Iran	84
Iraq	87	Ireland	93	Israel	94	Italy	102
Jamaica	72	Japan	105	Kenya	72	Korea (S.)	106
Lebanon	86	Malaysia	92	Marshall I.	84	Mexico	87
Morocco	85	Nepal	78	Netherlands	102	New Zealand	100
Nigeria	67	Norway	98	Peru	90	Phillipines	86
Poland	99	Portugal	95	Puerto Rico	84	Qatar	78
Romania	94	Russia	96	Samoa	87	Sierra Leone	64
Singapore	103	Slovakia	96	Slovenia	95	South.Africa	72
Spain	97	Sudan	72	Surname	89	Sweden	101
Switzerland	101	Taiwan	104	Tanzania	72	Thailand	91
Tonga	87	Turkey	90	Uganda	73	U.K.	100
U.S.	98	Uruguay	96	Zambia	77	Zimbabwe	66

Bayesian Normal Models, IQ Example (cont.)

- Using the priors: $p(\mu) \propto c$, $p(\sigma) \propto \sigma^{-1}$, we get the posterior summary:

Quantile:	0.01	0.10	0.25	0.50	0.75	0.90	0.99
μ	85.05	86.48	87.30	88.21	89.11	89.93	91.38
σ^2	56.74	63.42	67.71	72.97	78.81	84.61	96.12

- Note that the distribution of μ is centered at 88 rather than 100, and the mode of the posterior variance implies a standard error of roughly 8.5.

Data Exercise 3: Normal Data Summary

- Data from:
  Katherine Steffen, Allan Doctor, Julie Hoerr, Jeff Gill, Chris Markham, Sarah M. Brown, Daniel Cohen, Rose Hansen, Emily Kryzer, Jessica Richards, Sara Small, Stacey Valentine, Jennifer L. York, Enola K. Proctor, Philip C. Spinella. "Controlling Phlebotomy Volume Diminishes PICU Transfusion: Implementation Processes and Impact." Pediatrics, Forthcoming 2017.

- This study indicates that excessive phlebotomy in critically ill children is common, however use of simple patient blood management strategies reduced blood overdraw volumes. Implementation science provides robust models to facilitate acceptance and adoption of these patient blood management strategies.

- Rerun the normal summary model with `PIM2` (Revised Paediatric Index of Mortality) from these data to get a summary of both the mean and variance.

Data Exercise 3: Normal Data Summary

- Here is the code for the mean, you need to write the code for the variance.

```
pim2 <- scan("http://jeffgill.org/files/jeffgill/files/pbm.pim2_.dat_.txt")
n <- length(pim2)
t.pim2 <- (pim2-mean(pim2))/(sd(pim2)/sqrt(n))
r.t <- (rt(100000, n-1)*(sd(pim2)/sqrt(n))) + mean(pim2)
quantile(r.t,c(0.01,0.10,0.25,0.5,0.75,0.90,0.99))
```