# Harvard Department of Government 2017
# Chapter 3, Normal and Students-$t$ Models

JEFF GILL
*Visiting Professor, Fall 2024*

# 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 $\mu$ and known variance $\sigma_0^2$:

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

- And specify a normal prior distribution for $\mu$:

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

# 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 $\mu$...

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

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:

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

- 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 $\sigma_0^2$ (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 $\mu$ not the posterior of $\sigma^2$.

# 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:

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

- Relabel the sufficient statistic for $\sigma^2$ as a convenience:

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

giving the simplified form:

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

# Bayesian Normal Models, Mean Known, Variance Unknown

- Assign an inverse gamma prior for $\sigma^2$:

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

where: $\sigma^2 > 0, \alpha > 0, \beta > 0.$

- This has some moment limitations as well:

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

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

# 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:

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

# Normal Model with Both Mean and Variance Unknown

- We can now develop the model with both $\mu$ and $\sigma^2$ unknown, with conjugate priors for both.
- The new wrinkle is that the conjugate prior for the mean is conditional on the variance.
- We use the following inverse gamma and normal conjugate priors:

$$p(\sigma^2|\alpha,\beta) \propto (\sigma^2)^{-(\alpha+1)} \exp\left[-\beta/\sigma^2\right]$$

$$p(\mu|\sigma^2/s_0, m) \propto (\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2/s_0}(\mu - m)^2\right],$$

- The parameter $s_0$ is a so-called "confidence parameter" that measures the researcher's strength of belief in the prior dependence of $\sigma^2$ on $\mu$.

# Normal Model with Both Mean and Variance Unknown, Joint Posterior

- Now rearrange in a similar fashion as before:

$$\pi(\mu,\sigma^2|\mathbf{x}) \propto (\sigma^2)^{-\alpha-\frac{n}{2}-\frac{3}{2}} \exp\left[-\frac{1}{\sigma^2}\beta - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{x}_i-\mu)^2 - \frac{1}{2\sigma^2/s_0}(\mu-m)^2\right]$$

$$= (\sigma^2)^{-\alpha-\frac{n}{2}-\frac{3}{2}} \exp\left[-\frac{1}{\sigma^2}\beta - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\mathbf{x}_i^2 - 2n\bar{\mathbf{x}}\mu + n\mu^2\right) - \frac{1}{2\sigma^2/s_0}(\mu^2 - 2\mu m + m^2)\right]$$

$$= (\sigma^2)^{-\alpha-\frac{n}{2}-\frac{1}{2}} \exp\left[-\frac{1}{\sigma^2}\beta - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\mathbf{x}_i^2 - n\mathbf{x}^2\right)\right] \times (\sigma^2)^{-1} \exp\left[-\frac{1}{2\sigma^2}\left((n+s_0)\mu^2 - 2(n\bar{\mathbf{x}} + ms_0)\mu + (n\bar{\mathbf{x}}^2 + s_0m^2)\right)\right]$$

# Normal Model with Both Mean and Variance Unknown, Marginal Posterior for $\sigma^2$

- Because of the separation in the last form (and proportionality) the marginalization is easy:

$$\pi(\sigma^2|\mathbf{x}) = \int_0^\infty \pi(\mu, \sigma^2|\mathbf{x}) d\mu$$
$$\propto (\sigma^2)^{-\alpha-\frac{n}{2}-\frac{1}{2}} \exp\left[-\frac{1}{\sigma^2}\left(\beta + \frac{1}{2}\sum_{i=1}^n \mathbf{x}_i^2 - \frac{1}{2}n\bar{\mathbf{x}}^2\right)\right].$$

- Therefore the posterior distribution of $\sigma^2$ is another inverted gamma according to:

$$\sigma^2|\mathbf{x} \sim \mathcal{IG}\left(\alpha + \frac{n}{2} - \frac{1}{2}, \beta + \frac{1}{2}\sum_{i=1}^n \mathbf{x}_i^2 - \frac{1}{2}n\bar{\mathbf{x}}^2\right).$$

# Normal Model with Both Mean and Variance Unknown, Marginal Posterior for $\mu$

- Since the prior (and posterior) distribution of $\mu$ is conditional on $\sigma^2$, the integral is not trivial, so so use a simple trick:

$$\begin{aligned}\pi(\mu|\mathbf{x}) &= \frac{\pi(\mu,\sigma^2|\mathbf{x})}{\pi(\sigma^2|\mathbf{x})} \\ &\propto \sigma^{-2}\exp\left[-\frac{1}{2\sigma^2}\left((n+s_0)\mu^2 - 2(n\bar{\mathbf{x}}+ms_0)\mu + (n\bar{\mathbf{x}}^2+s_0m^2)\right)\right] \\ &= \sigma^{-2}\exp\left[-\frac{1}{2\sigma^2/(n+s_0)}\left(\mu^2 - 2\frac{n\bar{\mathbf{x}}+ms_0}{n+s_0}\mu + \frac{n\bar{\mathbf{x}}^2+s_0m^2}{n+s_0}\right)\right]. \qquad (1)\end{aligned}$$

- We can see now that the posterior distribution of $\mu$ is the normal:

$$\mu|\sigma^2,\mathbf{x} \sim \mathcal{N}\left[\frac{n\bar{\mathbf{x}}+ms_0}{n+s_0}, \frac{\sigma^2}{n+s_0}\right].$$

- Note that the prior dependence on $\sigma^2$ flows through to the posterior for $\mu$.

# 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 Likehood Function

- Start with rearranging the likelihood function:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = \prod_{i=1}^{n} \left( (2\pi)^{-k/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \right)$$

$$\propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left[ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right]$$

$$= |\boldsymbol{\Sigma}|^{-n/2} \exp\left[ -\frac{1}{2} \sum_{i=1}^{n} \left( \mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \mathbf{x}_i - 2\mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right) \right]$$

$$= |\boldsymbol{\Sigma}|^{-n/2} \exp\left[ -\frac{1}{2} \left( \sum_{i=1}^{n} \mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \mathbf{x}_i - n\bar{\mathbf{x}}' \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}} + n\bar{\mathbf{x}}' \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}} - 2n\bar{\mathbf{x}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + n\boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right) \right]$$

$$= |\boldsymbol{\Sigma}|^{-n/2} \exp\left[ -\frac{1}{2} \left( \operatorname{tr}(\boldsymbol{\Sigma}^{-1}) \left( \sum_{i=1}^{n} (\mathbf{x}_i' \mathbf{x}_i) - n\bar{\mathbf{x}}' \bar{\mathbf{x}} \right) + n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \right) \right].$$

## Both Unknown, Looking at the Likehood 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:

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

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

  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.

# Both Unknown (cont.)

- This setup leads to an articulation of the joint posterior:

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

$$\times \left|\frac{\boldsymbol{\Sigma}}{n_0}\right|^{-n/2} \exp\left[-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{m})\left(\frac{\boldsymbol{\Sigma}}{n_0}\right)^{-1}(\boldsymbol{\mu} - \mathbf{m})\right]$$

$$\times |\boldsymbol{\beta}|^{-\alpha/2}|\boldsymbol{\Sigma}^{-1}|^{(\alpha-(k+1))/2}\exp[-\text{tr}(\boldsymbol{\beta}^{-1}\boldsymbol{\Sigma}^{-1})/2]$$

# 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 |

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

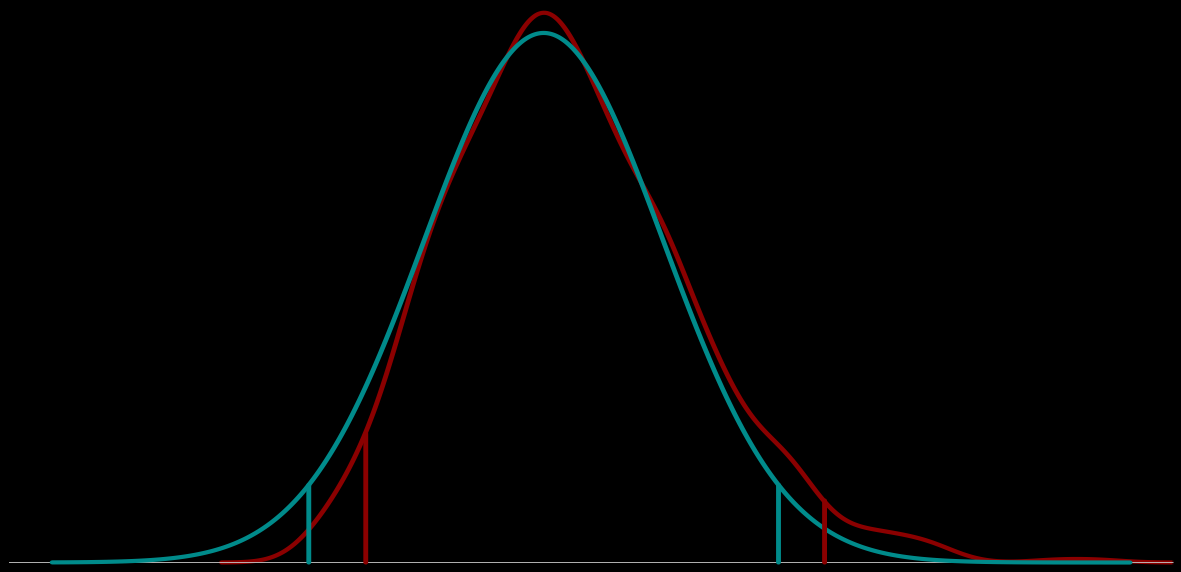
# Example: Variance Estimation with Public Health Data (cont.)

- 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("data/nc.sub.dat")
library(bayesm)     # FOR THE rwishart 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
```

## 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 $\theta$ (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

- Assume again iid normal data, rearrange the likelihood function:

$$L(\mu,\sigma|\mathbf{x}) = (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

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

$$= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\left(\sum(x_i-\bar{x})^2-2\sum(x_i\mu-x_i\bar{x}-\bar{x}\mu+\bar{x}^2)+n(\bar{x}-\mu)^2\right)\right]$$

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

# Bayesian Normal Models, Uninformative Priors

- This gives the likelihood expressed in terms of two sufficient statistics:

$$\bar{x} = \frac{1}{n}\sum(x_i)$$

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

- which are distributed $\mathcal{N}(\mu, \sigma^2/n)$ and $(\sigma^2/(n-1))\chi^2_{n-1}$, respectively, under classical results.

## Bayesian Normal Models, Uninformative Priors

- Common uninformative priors are given by:

$$p(\mu) \propto c, \quad -\infty < \mu < \infty$$
$$p(\sigma) \propto \sigma^{-1}, \quad 0 < \sigma < \infty,$$

- Note that both of these forms are not only uninformative (diffuse), but they do not integrate to a finite constant: *improper*.
- The purpose is to insert as little prior information into the posterior as possible.

# Bayesian Normal Models, Uninformative Priors

- The joint posterior distribution is calculated by:

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

  where we included the normalizing constant for a change.

- We can get the marginal posterior for $\mu$ by integrating out $\sigma^2$:

$$\pi(\mu|\mathbf{x}) \propto \int_0^\infty \sigma^{-(n+1)} \exp\left[-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\mu - \bar{x})^2\right)\right] d\sigma^2$$

- but this is too much trouble.

# Bayesian Normal Models, Uninformative Priors

- So lets make use of the handy integral (note the similarity to $\mathcal{IG}$):

$$\int_0^\infty x^{-b-1}\exp[-a/x^2]dx = \frac{1}{2}a^{-\frac{b}{2}}\Gamma\left(\frac{b}{2}\right),$$

- Set $\sigma = x$, $n = b$, and $\frac{1}{2}\left((n-1)s^2 + n(\mu - \bar{x})^2\right) = a$ so that our target integral is of this form.
- Specifically, if $k = c\left(\frac{n}{2\pi}\right)^{\frac{1}{2}}$, then:

$$\begin{aligned}\pi(\mu, \sigma|\mathbf{x}) &= k\sigma^{-(n+1)}\exp\left[-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\mu-\bar{x})^2\right)\right] \\ &= kx^{-b-1}\exp\left[-\frac{a}{x^2}\right].\end{aligned}$$

# Bayesian Normal Models, Uninformative Priors

- This of course means that:

$$\pi(\mu|\mathbf{x}) = \int_0^\infty \pi(\mu, \sigma|\mathbf{x}) d\sigma = \frac{1}{2} a^{-\frac{b}{2}} \Gamma\left(\frac{b}{2}\right),$$

  with $\sigma = x$, $n = b$, $\frac{1}{2}\left((n-1)s^2 + n(\mu - \bar{x})^2\right) = a$, $k = c\left(\frac{n}{2\pi}\right)^{\frac{1}{2}}$.

- So:

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

- And with a little clean up/rearranging plus judicious insertion of constants:

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

# Bayesian Normal Models, Uninformative Priors

- Big deal, right? Its easier to see if we make the transformation

$$t = \frac{\mu - \bar{x}}{s/\sqrt{n}}$$

with the Jacobian:

$$J = \left|\frac{\partial}{\partial t}\mu\right| = \left|\frac{\partial}{\partial t}\left(\frac{s}{\sqrt{n}}t + \bar{x}\right)\right| = \frac{s}{\sqrt{n}}$$

and reparameterize $n = \theta + 1$.

- Now the result is expressed as:

$$\begin{aligned}\pi(t|\mathbf{x}) &= \frac{\Gamma\left(\frac{\theta+1}{2}\right)}{\Gamma\left(\frac{\theta}{2}\right)}\frac{1}{(\theta\pi)^{\frac{1}{2}}}\left(\frac{\theta+1}{s^2}\right)^{\frac{1}{2}}\left(1+\frac{1}{\theta}(t)^2\right)^{-\frac{1}{2}(\theta+1)}\frac{s}{\sqrt{\theta+1}} \\ &= \frac{\Gamma\left(\frac{\theta+1}{2}\right)}{\Gamma\left(\frac{\theta}{2}\right)}\frac{1}{(\theta\pi)^{\frac{1}{2}}(1+t^2/\theta)^{(\theta+1)/2}}.\end{aligned}$$

- 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 $\mu$ is also student's-t with non-centrality parameter $\bar{x}$.

# Bayesian Normal Models, Uninformative Priors

- We will make use of the property $\pi(\sigma|\mathbf{x}) = \frac{\pi(\mu,\sigma|\mathbf{x})}{\pi(\mu|\sigma,\mathbf{x})}$ even though it artificially assumes $\sigma$ were known in the denominator.
- Now obtain the marginal posterior of $\sigma$ by dividing the joint posterior by the conditional distribution of $\mu$ assuming that $\sigma$ 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 $\sigma^2$ 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 |
|---|---|---|---|---|---|---|---|
| $\mu$ | 85.05 | 86.48 | 87.30 | 88.21 | 89.11 | 89.93 | 91.38 |
| $\sigma^2$ | 56.74 | 63.42 | 67.71 | 72.97 | 78.81 | 84.61 | 96.12 |

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

# An Example with Count Data

- As an example of the model-building process consider a dataset consisting of counts, for which we specify a Poisson likelihood function:

$$f(\mathbf{y}|\mu) = \left(\prod_{i=1}^{n} y_i!\right)^{-1} \exp\left[\log(\mu)\sum_{i=1}^{n} y_i\right] \exp[-n\mu].$$

- A convenient form of the prior for $\mu$ (the intensity parameter) is a gamma distribution with prior parameters $\alpha$ and $\beta$ which the researchers specifies:

$$f(\mu|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}\mu^{\alpha-1}e^{-\beta\mu}, \qquad \mu,\alpha,\beta > 0.$$

# An Example with Count Data

- This is handy because the gamma distribution is conjugate in the way previously described to the Poisson likelihood, meaning that the posterior is also a gamma form:

$$\begin{aligned}\pi(\mu|\mathbf{y}) \propto\ & p(\mu|\alpha,\beta)L(\mathbf{y}|\mu) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}\mu^{\alpha-1}\exp(-\beta\mu)\left(\prod_{i=1}^{n} y_i!\right)^{-1}\exp\left[\log(\mu)\sum_{i=1}^{n} y_i\right]\exp[-n\mu] \\ \propto\ & \mu^{\alpha-1}\exp(-\beta\mu)\exp\left[\log(\mu)\sum_{i=1}^{n} y_i\right]\exp[-n\mu] \\ \propto\ & \mu^{(\alpha+n\bar{y})-1}\exp\left[-(\beta+n)\mu\right].\end{aligned}$$

- Therefore the posterior distribution for $\mu$ is $\mathcal{G}(\alpha+n\bar{y}, \beta+n)$, and thus has mean $(\alpha+n\bar{y})/(\beta+n)$ and variance $(\alpha+n\bar{y})/(\beta+n)^2$.
- The neat part about this is we now have everything we need to know about the posterior and can either rely upon known properties of the gamma distribution or simply simulate values according to this parametrization and summarize empirically.

# An Example with Count Data

- As an illustration, consider counts of U.S. testing of thermonuclear devices up until the 1992 Test Ban Treaty, which the U.S. signed.
- The first genuinely thermonuclear device, "Mike," was detonated November 1, 1952 on the former Islet of Elugelab (it now no longer exists).
- See DeGroot (2004, p175-8) for interesting details on the history and policy of testing in the United States.
- So the dataset for hydrogen bombs is set to start in calendar year 1953 here.
- We will terminate the data at 1988, with a change of presidents leading up to the 1992 treaty.
- The data (source: U.S. Department of Energy) are given in Table **??**.

# Counts of U.S. Thermonuclear Tests By Year

| *Year* | Count | *Year* | Count | *Year* | Count | *Year* | Count | *Year* | Count |
|---|---|---|---|---|---|---|---|---|---|
| 1953 | 11 | 1954 | 6 | 1955 | 18 | 1956 | 18 | 1957 | 32 |
| 1958 | 77 | 1959 | 0 | 1960 | 0 | 1961 | 10 | 1962 | 98 |
| 1963 | 47 | 1964 | 47 | 1965 | 39 | 1966 | 48 | 1967 | 42 |
| 1968 | 56 | 1969 | 46 | 1970 | 39 | 1971 | 24 | 1972 | 27 |
| 1973 | 24 | 1974 | 23 | 1975 | 22 | 1976 | 21 | 1977 | 20 |
| 1978 | 21 | 1979 | 16 | 1980 | 17 | 1981 | 17 | 1982 | 19 |
| 1983 | 19 | 1984 | 20 | 1985 | 18 | 1986 | 15 | 1987 | 15 |
| 1988 | 15 | | | | | | | | |

# An Example with Count Data

- The data have mean $\bar{y} = 25.55$ over this period.
- Suppose we set $\alpha = 50$, and $\beta = 2$ to be somewhat sympathetic with the data.
- Let us also set $\alpha = 5$, and $\beta = 1$ as a cynical prior to serve as a comparison specification and thus a test of how influential the first prior turns out to be.
- The calculations can be done analytically, but it is perhaps easier and no less accurate to do them by simulation using `R`.
- The `R` object **`nukes`** is a data frame where the first column contains years and the second column contains counts.

# An Example with Count Data

```
nukes <- read.table("nukes.vec.data", header=FALSE)
alpha <- 50; beta=2; n <- nrow(nukes)
prior.sims <- rgamma(10000,shape=alpha,rate=beta)
post.sims  <- rgamma(10000,shape=alpha+n*mean(nukes[,2]),rate=beta+n)
c(mean(post.sims),var(post.sims))

alpha <- 5; beta=1; n <- nrow(nukes)
prior.sims <- rgamma(10000,shape=alpha,rate=beta)
post.sims  <- rgamma(10000,shape=alpha+n*mean(nukes[,2]),rate=beta+n)
c(mean(post.sims),var(post.sims))
```

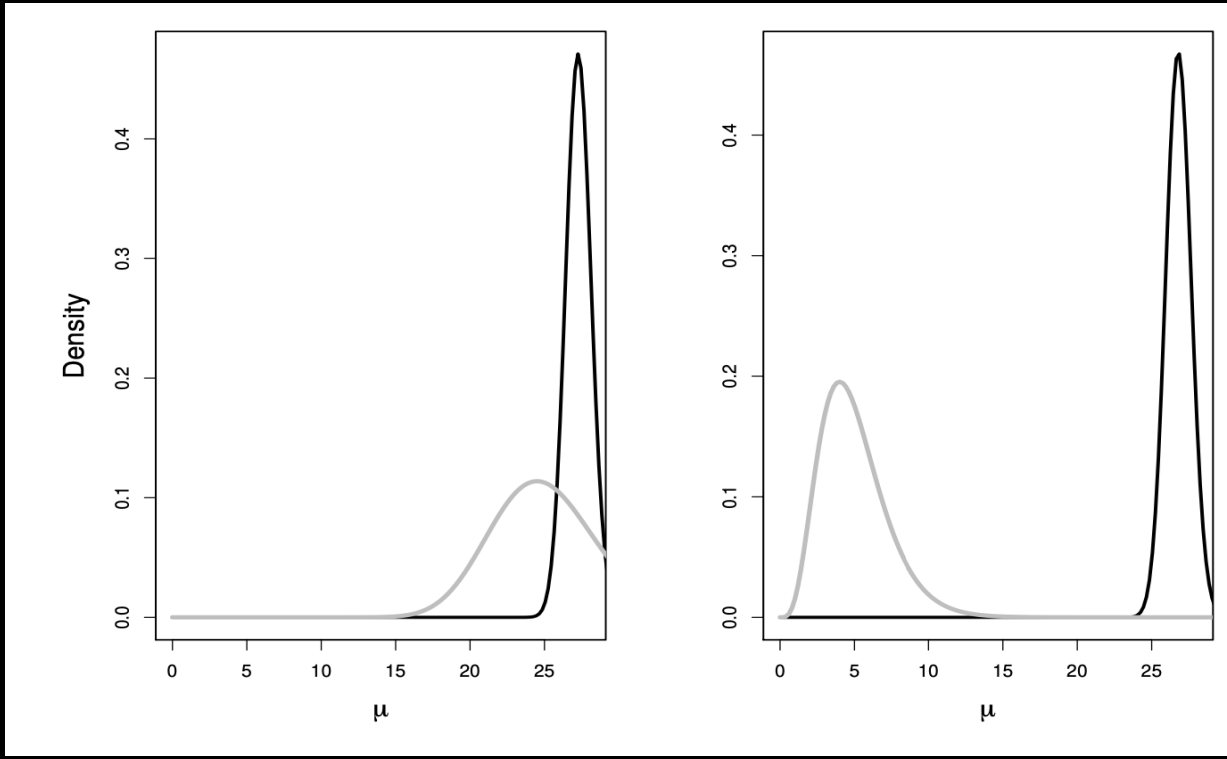
# Gamma-Poisson Model for Thermonuclear Tests

| | | *Resulting Moments for Distributions* | | | |
|---|---|---|---|---|---|
| | | Prior | | Posterior | |
| *Gamma Prior Parameters* | | Mean | Variance | Mean | Variance |
| $\alpha = 50,$ | $\beta = 2$ | 25.0 | 12.5 | 27.29 | 0.74 |
| $\alpha = 5,$ | $\beta = 1$ | 5.0 | 5.0 | 26.80 | 0.72 |

# An Example with Count Data

- What this shows is that the posterior distribution is relatively insensitive to the form of the prior that we select.
- This robustness property is reassuring in that it demonstrates that we are letting the data speak clearly here.
- To further illustrate these results consider the graph: we see here how dramatically "wrong" the second prior specification is, yet how it changes relatively little in the final analysis.
- The black curve is the posterior distribution and the grey curve is the associated prior distribution.

# An Example with Count Data



# An Example with Count Data

- So how would a non-Bayesian analysis of these data proceed?
- We know that the maximum likelihood estimate of the intensity parameter from a Poisson specification is the data mean, $\bar{y} = 25.55$, which differs little from the analysis above except that it buries the assumption of a uniform prior over the support of $\mu$.
- More specifically, every non-Bayesian result is equivalent to a Bayesian result where the prior is an appropriately bounded uniform distribution.
- So non-Bayesians assume a priori that all possible outcomes are equally likely.
- This does not seem like a terrible assumption except that we *know* something about anticipated thermonuclear tests here.
- For instance, five is much more likely than 500 and so on.
- Many people can still live with such uniform assumptions, but the usual procedure is not overtly stating this as a model component and therefore not having to defend it.

# An Example with Count Data

- How would we get the appropriate uniform prior for this example?
- It requires a form of the gamma prior used here such that the Poisson likelihood is identical to the posterior we derived.
- Working backwards, inserting $\alpha = 1$ and $\beta = 0$ gives the desired equality, except that now the prior is undefined due to the treatment of zero.
- So instead we say that the "appropriately bounded uniform distribution" is a $G(\alpha, \beta)$ PDF with $\alpha = 1$ and $\beta$ some infinitesimally small positive real number but not zero: $\log(\beta) \ll 0$.
- Obviously we cannot plug in an infinitesimally small $\beta$ value into a PDF in the measures sense, but we can graph this distribution for a very small value of $\beta$ to get a sense of what such a distribution looks like.

# An Example with Count Data

- The figure shows this form ($\alpha = 1$, very small $\beta$) and it easy to see the uniform structure.
- Actually, this uniformness extends to positive infinity and thus defines an *improper prior* since it does not integrate to a finite number.