STAT3016/4116/7016: Introduction to Bayesian Data Analysis

RSFAS, College of Business and Economics, ANU

Multiparameter models - the normal model and multinomial model for categorical data

Normal distribution - Review

What are the parameters of the normal distribution? Write down the probability density function of the normal distribution.

What are some key features of the normal distribution?

Why is the normal probability model the most useful?

Inference for the mean, conditional on the variance

Suppse $\{Y_1, ..., Y_n | \theta, \sigma^2\} \overset{\text{iid}}{\sim} Normal(\theta, \sigma^2)$.
Let's work out the posterior distribution of θ when σ^2 is known.

$$
\begin{aligned}
p(\theta | y_1, ..., y_n, \sigma^2) &\propto p(\theta|\sigma^2) p(y_1, ..., y_n | \theta, \sigma^2) \\
&\propto p(\theta|\sigma^2) \times e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}
\end{aligned}
$$

What class of prior distributions would be conjugate to the normal sampling model when σ^2 is known?

Gamma prior : $p(\theta) \sim \theta^{\alpha-1} \exp(-b\theta)$

(inappropriate prior) $p(\theta|y) \sim \theta^{\alpha-1} \exp(-b\theta)\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}$

NORMAL PRIOR: product of normal $\rightarrow$ normal.

Inference for the mean, conditional on the variance

Show that if $\theta \sim \mathrm{Norm}(\mu_0, \tau_0^2)$ and $y_1, ...y_n$ are i.i.d normal (θ, σ^2) then $p(\theta | y_1, ..., y_n, \sigma^2)$ is also a normal density with mean parameter

$$\boxed{\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}} \ ,$$

looks like the posterior mean is the weighted prior mean & weighted sample mean.

and variance parameter

prior precision

$$\boxed{\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}} \ .$$

$\frac{1}{\tau_n^2}$ posterior precision

$\frac{1}{\sigma^2}$ sampling precision

Inference for the mean, conditional on the variance

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \quad . \quad \tilde{\tau}_n^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2$$

1. Interpret the formula for the posterior variance τ_n^2
2. Interpret the formula for the posterior mean μ_n.
3. Consider predicting a new observation $\tilde{Y}$ from the normal population after having observed $(Y_1 = y_1, ..., Y_n = y_n)$. Find $E[\tilde{Y}|\sigma^2, y_1, ..., y_n]$ and $Var[\tilde{Y}|\sigma^2, y_1, ..., y_n]$.
   Interpret the result.

$$\mu_n = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\bar{y}$$

$$\{\tilde{Y} \mid \theta, \sigma^2\} \sim N(\theta, \sigma^2) \iff \tilde{Y} = \theta + \tilde{\varepsilon}, \ \{\tilde{\varepsilon} \mid \theta, \sigma^2\} \sim N(0, \sigma^2)$$

$$E[\tilde{Y} \mid y_1, \cdots, y_n, \sigma^2] = E[\theta + \tilde{\varepsilon} \mid y_1, \cdots, y_n, \sigma^2]$$

$$= E[\theta \mid y_1, \cdots, y_n, \sigma^2] + E[\tilde{\varepsilon} \mid y_1, \cdots, y_n, \sigma^2]$$

$$= \mu_n + 0 = \mu_n$$

$$Var[\tilde{Y} \mid y_1, \cdots, y_n, \sigma^2] = Var[\theta + \tilde{\varepsilon} \mid y_1, \cdots, y_n, \sigma^2]$$

$$= Var[\theta \mid y_1, \cdots, y_n, \sigma^2] + Var[\tilde{\varepsilon} \mid y_1, \cdots, y_n, \sigma^2]$$

$$= \tau_n^2 + \sigma^2$$

$$\tilde{Y} \mid \sigma^2, y_1, \cdots, y_n \sim N(\mu_n, \tau_n^2 + \sigma^2)$$

Example: Car speeds

Suppose you drive on a particular highway and typically drive at a constant speed of 110 km/hr (the speed limit). Suppose that speeds are normally distributed with unknown mean θ and known standard deviation σ. We have average speed data from 10 cars. We wish to make inference about the population mean θ.

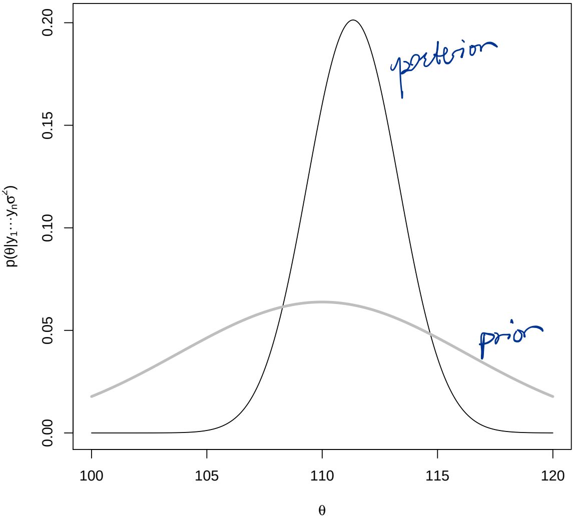
The data are (98, 100, 107, 110, 112, 117, 117, 120, 125, 130)

Use sample mean & variance

What would be an appropriate conjugate prior distribution for θ? If σ is assumed to be known, what would you assume its value to be? What is the posterior distribution of θ? Provide a 95% quantile-based posterior confidence interval for θ.

What would be a more accurate representation of your information?

Example: Car speeds



95% posterior quantile based confidence interval for θ is (107.45 115.22)

Joint inference for the mean and variance

We want to evaluate:

$$p(\theta, \sigma^2 | y_1, ..., y_n) = \frac{p(y_1, ..., y_n | \theta, \sigma^2) p(\theta, \sigma^2)}{p(y_1, ..., y_n)}$$

Let's develop a simple class of conjugate prior distributions. Recall:

$$p(\theta, \sigma^2) = p(\theta | \sigma^2) p(\sigma^2)$$

Let $\theta | \sigma^2 \sim \mathrm{Norm}(\mu_0, \tau_0 = \frac{\sigma}{\sqrt{\kappa_0}})$.

Interpretation of μ_0 and κ_0?? (hypothetically)

mean & sample size from a set of prior observations

Joint inference for the mean and variance

What about $p(\sigma^2)$?? What is the required support of $p(\sigma^2)$?? The gamma family is a conjugate class of densities for $1/\sigma^2$ (the precision). Then σ^2 has an *inverse-gamma* distribution.

$$\text{precision} = 1/\sigma^2 \sim \text{gamma}(\alpha, \beta)$$

$$\text{variance} = \sigma^2 \sim \text{inverse} - \text{gamma}(\alpha, \beta)$$

$p(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)}e^{-\beta/\sigma^2}$, $\sigma^2 > 0$.

$E[\sigma^2] = \frac{\beta}{\alpha-1}$; $Mode[\sigma^2] = \frac{\beta}{\alpha+1}$; $Var[\sigma^2] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$

Joint inference for the mean and variance

Let $\alpha = \frac{\nu_0}{2}$ and let $\beta = \frac{\nu_0}{2}\sigma_0^2$. So we have:

$$\sigma^2 \sim \text{inverse} - \text{gamma}(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2)$$

sample mean of current & prior observation

$$\theta|\sigma^2 \sim \text{Norm}(\mu_0, \frac{\sigma}{\sqrt{\kappa_0}})$$

σ^2 / total # of obs (both prior & current)

$$Y_1, ..., Y_n|\theta, \sigma^2 \overset{\text{iid}}{\sim} \text{Norm}(\theta, \sigma^2)$$

What is the form of the posterior density $p(\theta|y_1, ..., y_n, \sigma^2)$?? Find $E[\theta|y_1, ..., y_n, \sigma^2]$ and $Var[\theta|y_1, ..., y_n, \sigma^2]$.

normal$(\mu_n, \sigma^2/\kappa_n)$

$\kappa_n = \kappa_0 + n$, $\mu_n = \dfrac{(\kappa_0/\sigma^2)\mu_0 + (n/\sigma^2)\bar{y}}{\kappa_0/\sigma^2 + n/\sigma^2} = \dfrac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}$

Joint inference for the mean and variance

What about $p(\sigma^2|y_1,....,y_n)$??

$$p(\sigma^2|y_1,...,y_n) \propto p(\sigma^2)p(y_1,...,y_n|\sigma^2)$$
$$= p(\sigma^2)\int p(y_1,...,y_n|\theta,\sigma^2)p(\theta|\sigma^2)d\theta$$

With tedious algebra:

$$\left\{\sigma^2|y_1,...,y_n\right\} \sim \mathrm{inv}-\mathrm{gamma}(\nu_n/2, \nu_n\sigma_n^2/2)$$

where

$$\nu_n = \nu_0 + n$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y}-\mu_0)^2\right]$$

$(\kappa_n = \kappa_0 + n)$. Interpret the terms in σ_n^2.

Car speeds example - continued

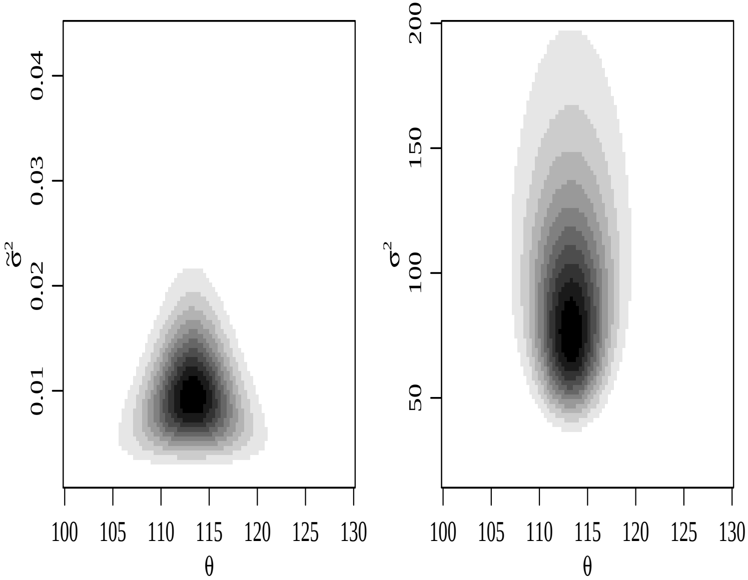
Let $\mu_0 = 110$ and $\sigma_0^2 = 100$. If we set κ_0=ν_0=1, then our prior distributions are only weakly centered around these estimates.

$\bar{y} = 113.6$; $s^2 = 105.6$.

Calculate the parameters of the posterior distribution $p(\theta, \sigma^2 | y_1, ..., y_n)$

Create a contour plot of the bivariate posterior density of $(\theta, \tilde{\sigma}^2)$ where $\tilde{\sigma}^2 = 1/\sigma^2$.

Car speeds example - continued



Monte-Carlo sampling

We might be interested in the marginal posterior distribution of θ given the data and calculate quantities like $E[\theta|y_1, ..., y_n]$; $Var[\theta|y_1, ..., y_n]$.

How would you generate Monte Carlo posterior samples of θ in the case where both θ and σ^2 are unknown.

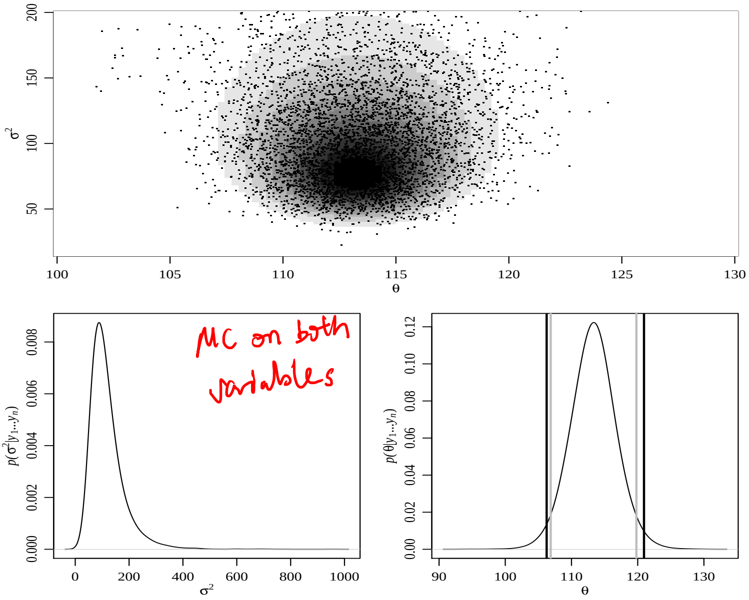
Cars speed example - continued

```
> S<-10000
> s2.postsample<-1/rgamma(S,  (nu0+n)/2, s2n*(nu0+n)/2 )
> theta.postsample<-rnorm(S, mun, sqrt(s2.postsample/(k0+n)))
> quantile(theta.postsample, c(.025,.975))
    2.5%    97.5%
106.8607 119.8012
```

Compare the Bayesian interval to a frequentist confidence interval

```
> c(ybar-1.96*sqrt(s2/n),ybar+1.96*sqrt(s2/n))
[1] 107.2308 119.9692
```

Cars speed example - continued



Estimation from two independent samples.

Exercise 1: Suppose we are interested in learning about the completion times for men and women between ages 20 and 29 who are running the New York marathon. We observed the times for 20 men and 20 women. The 20 measurements in the mens group had a sample mean of 278 minutes and a sample standard deviation of 49.5 minutes. The 20 measurements in the womens group had a sample mean of 291 minutes and a standard deviation of 56.2 minutes.

Find the posterior density of the difference in mean race times between men and women aged 20-29 in the New York marathon. Is there sufficient evidence to conclude that males have an average race time that is faster than the average race time for females?

Improper priors

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_0 + n} \longrightarrow \bar{y}$$

$$\sigma_n^2 = \frac{1}{\nu_0 + n}\left[\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2\right] \longrightarrow \frac{n-1}{n}s^2$$

$$= \frac{1}{n}\sum (y_i - \bar{y})^2$$

Let κ_0, $\nu_0 \to 0$. What happens to μ_n and σ_n^2??

Let $p(\theta, \sigma^2) = 1/\sigma^2$. Is this a proper prior? What are the posterior densities for θ and σ^2?

Derive the marginal posterior distribution of θ.

$$p(\theta, \sigma^2) = \frac{1}{\sigma^2} \iff p(\theta, \log\sigma) \propto 1$$

Bias, variance and mean squared error

$y_1, \dots, y_n \sim N(\theta, \sigma^2)$

$MSE = Bias^2 + Var$

$\hat{\theta}_{MLE} = \bar{y}$

$Var(\hat{\theta}_{MLE}) = \frac{\sigma^2}{n} = MSE(\hat{\theta}_{MLE})$

$E(\hat{\theta}_{MLE}) = \theta$ (unbiased)

Exercise 2: IQ scores

Scoring on IQ tests is designed to produce a normal distribution with a mean of 100 and a standard deviation of 15 (a variance of 225) when applied to the general population.

Let θ be the mean IQ score in a town of population size N. Suppose we take a sample of n individuals and meausure their IQ scores to come up with a sample estimate of θ.

What would be the maximum likelihood estimator of θ using the data? What is the Bayesian estimator of θ given the data? Compare the bias and mean squared error of both estimators.

(For your mean squared error calculations, assume the true mean and standard deviation for the town are $\theta = 112$ and $\sigma = 13$).

posterior inference by sample data. is dominated

$n \to \infty$

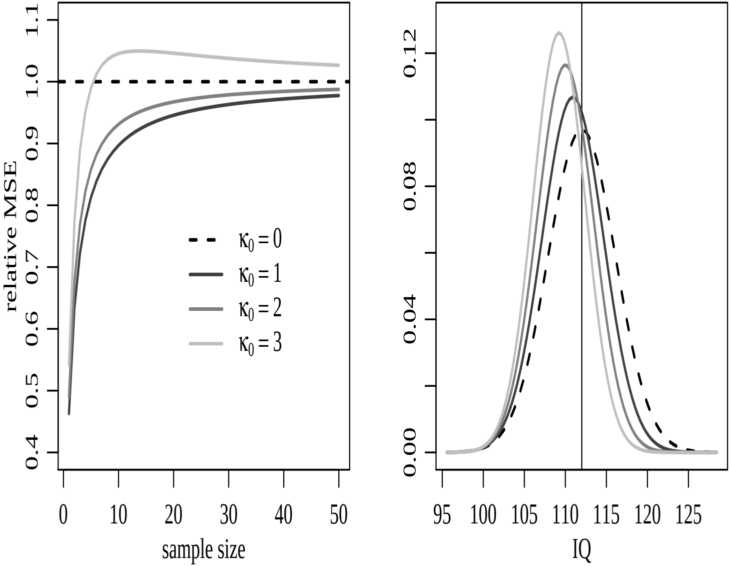
$w \to 1$

$\hat{\theta}_{bayes} = E(\theta|y) = \frac{n}{k_0+n}\bar{y} + \frac{k_0}{k_0+n}\mu_0$

$Bias(\hat{\theta}_{Bayes}) \neq 0$, $MSE(\hat{\theta}_{Bayes}) = w^2 \times \frac{\sigma^2}{n} + (1-w^2)(\mu_0 - \theta_0)^2$

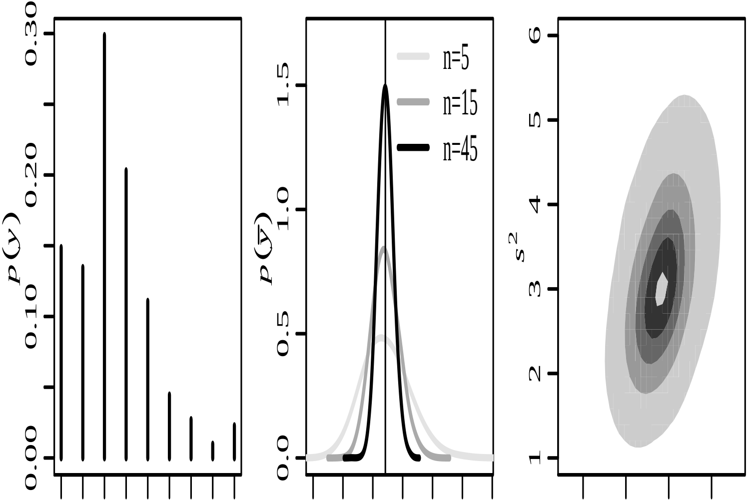
Bias, variance and mean squared error

Example: IQ scores



The normal model for non-normal data

Exercise 3: How reliable is the normal model in Bayesian modelling to estimate population quantities when the data are not normal? (Example: General social survey 1998, variable of interest: number of children per woman)



Poseterior predictive checking - note

Recall the difference in our notation between the predictive outcomes:

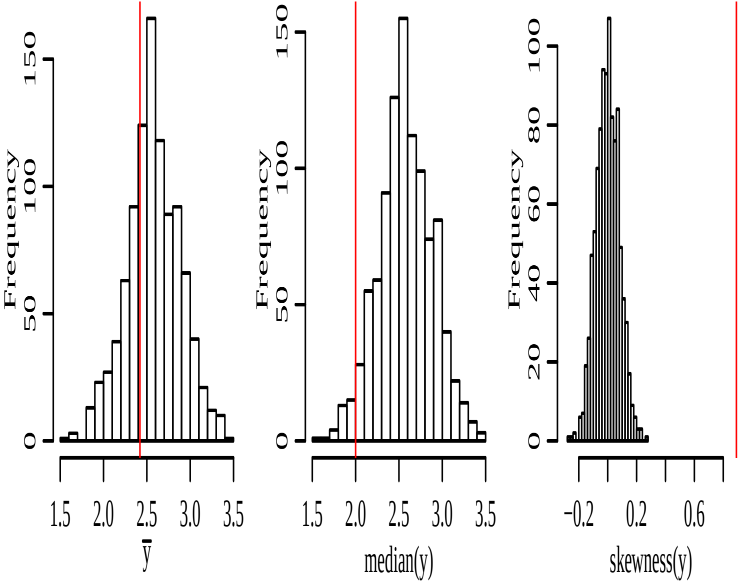
- y_{rep} (replicated data that *could have been observed* for the same units in the observed data set or the data *we would see tomorrow* if the experiment were repeated under the same conditions); and
- $\tilde{y}$ (any future observable value(s) for any unit (either observed or unobserved)).

We generate both y_{rep} and $\tilde{y}$ from the posterior predictive distribution, that is, $p(y_{rep}|y)$ or $p(\tilde{y}|y)$.

We use y_{rep} to calculate posterior - predictive p-values to check our sampling model assumptions (see lecture slides from Ch 4)

The normal model for non-normal data

Posterior predictive checking:



Multinomial model for categorical data

We can generalise the binomial distribution to allow more than two possible outcomes. Let $\mathbf{y}$ be the vector of counts of the number of observations of each of k outcomes.

$$p(\mathbf{y}|\boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{y_j}$$

where $\sum_{j=1}^{k} \theta_j = 1$.

The conjugate prior is a multivariate generalisation of the beta distribution known as the Dirichlet,

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

The resulting posterior for the θ_j's is Dirichlet with parameters $\alpha_j + y_j$.

Interpretation of prior parameters α_j?? Uniform prior density??

Multinomial model for categorical data

use monte carlo sample

Example: Pre-election polling

A survey was conducted on n=1447 adults to find out their preferences in the upcoming election. $y_1 = 727$ people supported candidate A, $y_2 = 583$ people supported candidate B, and y_3 people supported candidate B or expressed no opinion.

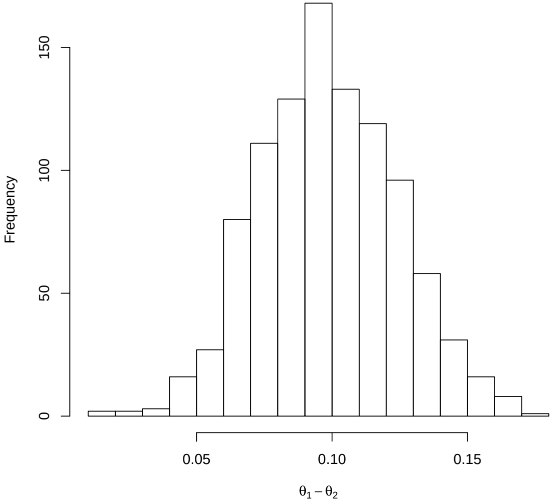
Assuming exchangeability and simple random sampling, the data (y_1, y_2, y_3) follow a multinomial distribution with parameters $(\theta_1, \theta_2, \theta_3)$. An estimand of interest is $\theta_1 - \theta_2$, the population difference in support of the two major candidates.

Multinomial model for categorical data

Example: Pre-election polling

With a noninformative prior distribution on $\boldsymbol{\theta}$, the posterior distribution is Dirichlet(728,584,138).

A is preferable.



Discrete (grid) approximations

IDEA: Construct a posterior distribution over a grid of parameter values.

Two parameter example: Suppose we are interested in the joint posterior distribution of two parameters, θ and σ^2. That is, we want $p(\theta, \sigma^2 | y_1, ..., y_n)$.

set appropriate range for the two values.

We know

$p(\theta, \sigma^2 | y_1, ..., y_n) \propto p(\theta, \sigma^2) p(y_1, ..., y_n | \theta, \sigma^2) = p(\theta, \sigma^2, y_1, ..., y_n)$.

Let's evaluate $p(\theta, \sigma^2, y_1, ..., y_n)$ on a two dimensional grid of values $\{\theta, \sigma^2\}$. Let $\{\theta_1, ..., \theta_G\}$ and $\{\sigma_1^2, ..., \sigma_H^2\}$ be sequences of equally spaced parameter values. To each pair $\{\theta_l, \sigma_k^2\}$ on the grid, we evaluate:

basically, evaluate posterior density.

$$\boxed{p(\theta_l, \sigma_k^2 | y_1, ..., y_n) = \frac{p(\theta_l, \sigma_k^2, y_1, ..., y_n)}{\sum_{g=1}^{G} \sum_{h=1}^{H} p(\theta_g, \sigma_h^2, y_1, ..., y_n)}}$$

"grid points density" $\Rightarrow$ approx of true target density

Discrete (grid) approximations

Let's apply the grid approximation to the cars speed data.

Recall: $n = 10$, $\bar{y} = 113.6$; $s^2 = 105.6$

The conjugate prior distribution $p(\theta|\sigma^2) = \mathrm{dnorm}(\theta, \mu_0, \sigma/\sqrt{\kappa_0})$. Prior uncertainty on θ is driven by the value of σ^2. To remove this constraint, lets use a semiconjugate prior distribution.

$$p(\theta) = \mathrm{dnorm}(\theta, \mu_0 = 110, \tau_0 = 2.5)$$

$$p(\sigma^2) = \mathrm{dinv\text{-}gamma}(\sigma^2, \frac{1}{2}, \frac{1}{2} \times 100)$$

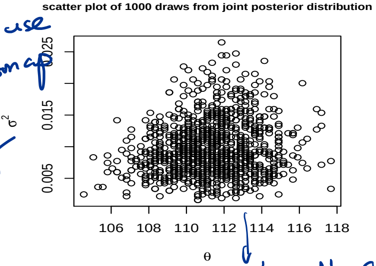
100x100

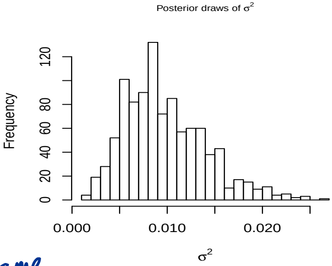
For your grid, use the range $(\theta, 1/\sigma^2) \in (105, 125) \times (0.001, 0.03)$ and divide each interval up into 100 evenly spaced points.

How would you obtain the marginal posterior distributions from your grid approximation? How would you obtain posterior draws of θ and σ^2 from your grid approximation.

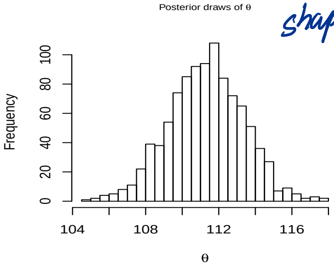
Discrete approximation- cars speed example

better use a heatmap to ~~do~~ display





share the same shape with previous contour plot.



Advantages of grid approximation? Disadvantages?