Bayesian Inference and Decision Theory

Unit 5: The Normal Model

v5.4

Learning Objectives for Unit 5

- Describe the conjugate prior distribution for the normal distribution with:
  - Unknown mean and known variance
  - Unknown mean and unknown variance
- Extend techniques from previous units to infer the posterior distribution for the mean, and the variance if unknown, of a normal distribution from a sample of observations
- Use Monte Carlo sampling to estimate posterior quantities from a normal distribution
- Explain why methods for normally distributed data are often used with non-normal data
  - State conditions under which this is justified
  - State conditions under which this practice can lead to misleading results
- Analyze accuracy of point estimators: Mean squared error, bias and variance

The Normal (Gaussian) Distribution

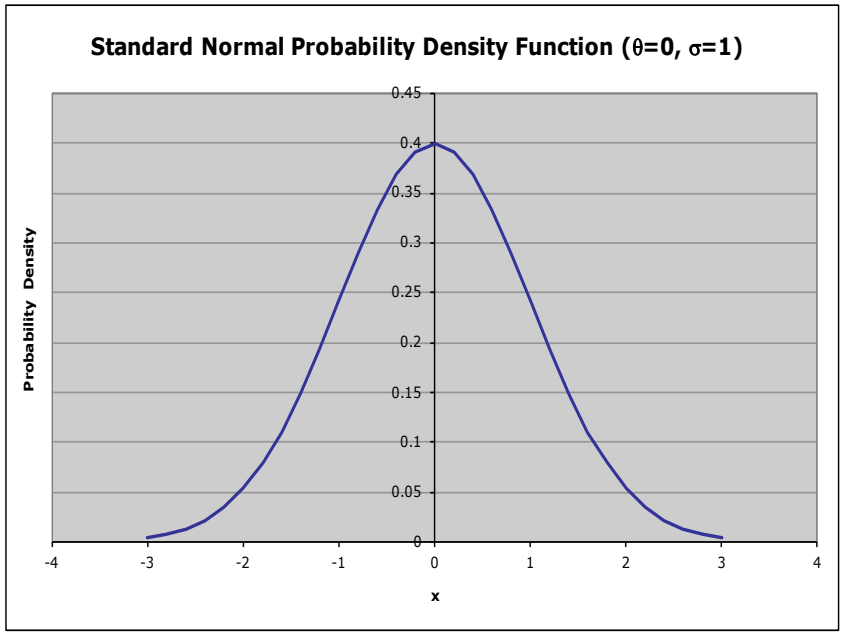
- Most studied and most applied distribution in statistics
- Probability density function is the familiar bell-shaped curve

$$f(x|\theta,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2\right\}$$

Note: $\exp\{x\}$ *means* e^x

  - Symmetric around θ
  - 95% of probability lies within about 2σ of θ
- Linear combinations of normally distributed random variables are normally distributed
- **Central Limit Theorem:**

  Under very general conditions, the sample mean of n iid observations from a distribution with mean θ and variance σ^2 tends, as n grows large, to a normal distribution with mean θ and variance σ^2/n



Conjugate Prior Distribution for Normal Observations: Unknown Mean, Known Standard Deviation

- The observations $X_1, \ldots, X_n$ are a random sample from a normal distribution with unknown mean Θ and known standard deviation σ:

$$f(\underline{x}|\theta,\sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma}\right)^2\right\}$$

Note: $\exp\{x\}$ *means* e^x

- Conjugate prior distribution for Θ is a normal distribution with mean μ and standard deviation τ

$$g(\theta|\mu,\tau) = \frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{1}{2}\left(\frac{\theta-\mu}{\tau}\right)^2\right\}$$

- Joint gpdf for $(\underline{X}, \Theta)$:

$$f(\underline{x}|\theta,\sigma)g(\theta|\mu,\tau) = \frac{1}{\sqrt{2\pi}\tau}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2}\left[\left(\frac{\theta-\mu}{\tau}\right)^2 + \sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma}\right)^2\right]\right\}$$

Notice the quadratic terms inside the exponent

Finding the Posterior Distribution

- Prior times likelihood:

$$f(\underline{x} \mid \theta,\sigma)g(\theta \mid \mu,\tau) = \frac{1}{\sqrt{2\pi}\tau}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2}\left[\left(\frac{\theta - \mu}{\tau}\right)^2 + \sum_{i=1}^{n}\left(\frac{x_i - \theta}{\sigma}\right)^2\right]\right\}$$

Posterior distribution of Θ is normal with mean μ and standard deviation τ**

- Reexpress terms inside square brackets:
  - Terms involving θ^2: $\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)\theta^2$
  - Terms involving θ: $-2\left(\frac{\mu}{\tau^2} + \frac{\sum x_i}{\sigma^2}\right)\theta$
  - Remaining terms: $\frac{\mu^2}{\tau^2} + \sum_i \frac{x_i^2}{\sigma^2}$
  - Add and subtract: $\dfrac{\left(\frac{\mu}{\tau^2} + \frac{\sum x_i}{\sigma^2}\right)^2}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$
  - Define: $\tau^* = \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-\frac{1}{2}}$ $\quad \mu^* = \dfrac{\frac{\mu}{\tau^2} + \frac{\sum x_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$
- Expression in brackets becomes:

$$\left[\left(\frac{\theta - \mu^*}{\tau^*}\right)^2 + \left(\frac{\mu^2}{\tau^2} + \sum_i \frac{x_i^2}{\sigma^2} - \frac{\left(\frac{\mu}{\tau^2} + \frac{\sum x_i}{s^2}\right)^2}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right)\right]$$

Exponent in a normal density — *Constant that does not depend on θ*

- Posterior pdf for Θ:

$$g(\theta \mid \underline{x},\mu,\tau) = g(\theta \mid \mu^*,\tau^*) = \frac{1}{\sqrt{2\pi}\tau^*}\exp\left\{-\frac{1}{2}\left(\frac{\theta - \mu^*}{\tau^*}\right)^2\right\}$$

Summary: Prior to Posterior
Normal / Normal Conjugate Pair (known variance)

IF Observations $X_1, \ldots, X_n$ are a random sample from Normal(Θ, σ) and prior distribution for Θ is Normal(μ, τ)

THEN Posterior distribution for Θ is Normal(μ^*, τ^*), another member of the conjugate family. The posterior hyperparameters are given by:

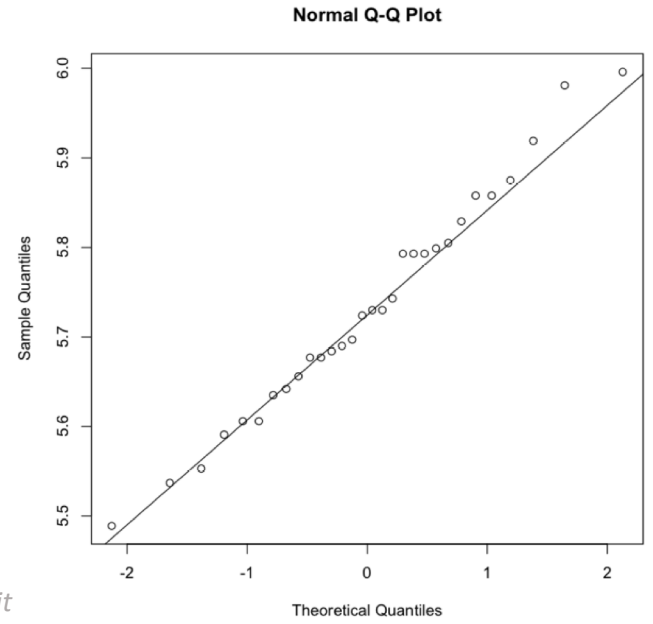
$$\mu^* = \frac{\frac{\mu}{\tau^2} + \frac{\sum x_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \qquad \tau^* = \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-\frac{1}{2}}$$

A Useful re-Parameterization

- Define the precision as the inverse of the variance:
  - X_i has precision $\rho = 1/\sigma^2$
  - Θ has precision $\lambda = 1/\tau^2$
  - Precision is always positive and is measured in the inverse square units of X
- Parameters for the posterior distribution of Θ :
  - $\lambda^* = \lambda + n\rho$
  - $\mu^* = \frac{\lambda\mu + \rho\sum X_i}{\lambda + n\rho} = \frac{\lambda\mu + (n\rho)\left(\frac{1}{n}\sum X_i\right)}{\lambda + n\rho}$
- Each observation increases the precision of the posterior distribution by the precision ρ of one observation
- The posterior mean is a weighted average of the observations and the prior mean:
  - The prior mean receives a relative weight of λ, the prior precision of the mean
  - Each observation receives a relative weight of ρ, the precision of the data distribution
- $\sum_i X_i$ is sufficient for Θ

Example: Reaction Time Data

- Gelman, et al. analyze data on reaction times for 11 non-schizophrenic and 6 schizophrenic subjects. The data set can be found at: http://www.stat.columbia.edu/~gelman/book/data/schiz.asc
- Gelman et al model log reaction times of the 11 non-schizophrenic subjects as normally distributed with subject-dependent mean and common standard deviation
- The normal distribution fits well for some subjects and less well for others
- The chart shows a normal Q-Q plot for log reaction times for the first subject in the study
- We will use this subject's data to illustrate conjugate normal updating with known standard deviation
  - Assume standard deviation is known and equal to sample standard deviation for this subject
  - Value of sample standard deviation is $\sigma = 0.127$
- Later we will consider unknown standard deviation

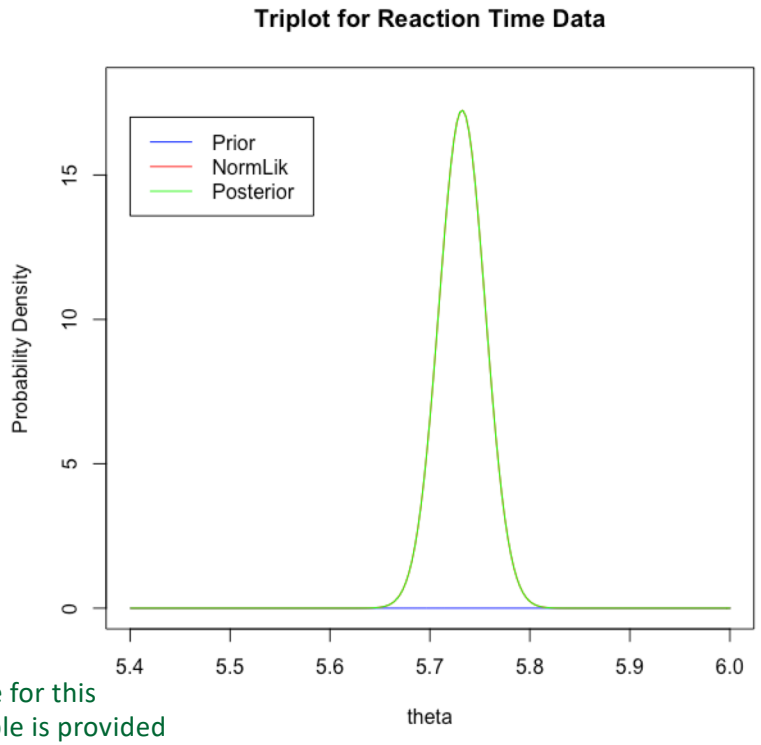


Reaction Times: Posterior Distribution

- The observations are normal with mean Θ and standard deviation $\sigma = 0.127$ (assumed known and equal to sample standard deviation)
- Assume uniform prior distribution for Θ
  - Limit of a Normal(0, τ) distribution as τ tends to infinity
  - This is the Jeffreys prior
  - It is an improper distribution (integrates to ∞)
- Data - log reaction time of first non-schizophrenic subject – 30 observations, sample mean 5.73
- The posterior distribution is normal with:
  - Posterior mean $\mu^* = \left(\frac{\frac{0}{\infty^2} + \frac{1}{0.127^2}\sum_i X_i}{\frac{1}{\infty^2} + \frac{30}{0.127^2}}\right) = \frac{1}{30}\sum_i X_i = 5.73$
  - Posterior standard deviation:

$$\tau^* = \left(\frac{1}{\infty^2} + \frac{30}{0.127^2}\right)^{-1} = 0.023$$

  - Posterior 95% posterior credible interval for Θ: $[5.69, 5.78]$



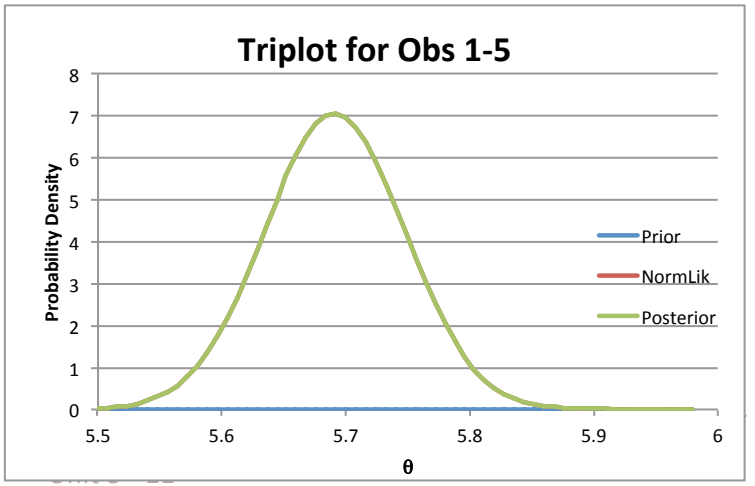
R code for this example is provided

Normal Conjugate Updating: Sequential Processing of Observations

- We will use the reaction time data to illustrate sequential Bayesian updating as we receive new batches of observations
  - Receive observations in batches of 5
  - Update distribution for mean after each batch
  - Assume observations are normal with unknown mean and known standard deviation (equal to sample standard deviation of observations)
- After each batch we update our distribution for Θ and predict the sample mean of the next batch of observations
- To predict the next batch of observations we will need to know the predictive distribution (marginal likelihood) of the sample mean

First Batch of Five Reaction Times

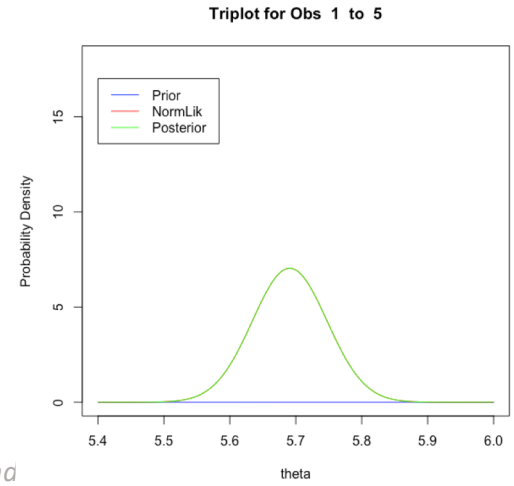
- The observations are normal with mean Θ and standard deviation $\sigma = 0.127$ (assumed known and equal to sample standard deviation)
- Assume uniform prior distribution for Θ
  - Limit of a Normal(0, τ) distribution as τ tends to infinity
  - This is the Jeffreys prior
  - It is an improper distribution (integrates to ∞)
- Data: 5.743, 5.606, 5.858, 5.656, 5.591 (average is 5.691)
- The posterior distribution is normal with:
  - Posterior mean $\mu_1 = \left(\frac{\frac{0}{\infty} + \frac{1}{0.127^2} \sum_i X_i}{\frac{1}{\infty} + \frac{5}{0.127^2}} \right) = \frac{1}{5} \sum_i X_i = 5.69$
  - Posterior standard deviation:
    $$\tau_1 = \left(\frac{1}{\infty^2} + \frac{5}{0.127^2} \right)^{-1} = 0.057$$
  - Posterior 95% credible interval for Θ: $[5.58, 5.80]$
- To predict the next batch of five reaction times, we will need the marginal likelihood for the normal / normal conjugate pair



Compare: First 5 Observations with Full Data

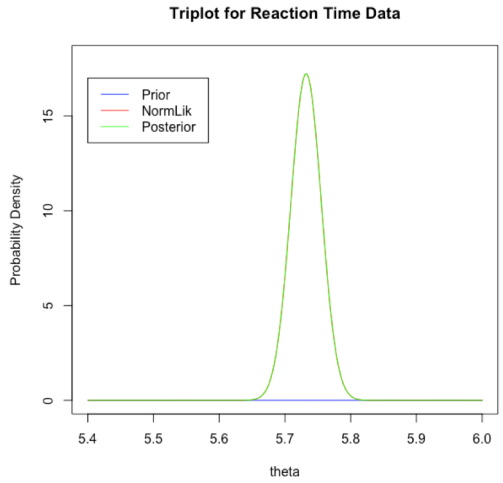
First 5 Observations

- Posterior mean $\mu_1 = 5.69$
- Posterior SD $\tau_1 = 0.057$
- 95% Interval for Θ: $[5.58, 5.80]$



All 30 Observations

- Posterior mean $\mu^* = 5.73$
- Posterior SD $\tau^* = 0.023$
- 95% Interval for Θ: $[5.69, 5.78]$



Marginal Likelihood for Normal / Normal Conjugate Pair (single observation)

- Joint gpdf of (X, Θ):
  - $f(x|\theta,\sigma)g(\theta|\mu,\tau) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\theta}{\sigma}\right)^2\right]\right\}\frac{1}{\sqrt{2\pi}\tau}\exp\left\{\left(\frac{\theta-\mu}{\tau}\right)^2\right\} = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\theta}{\sigma}\right)^2 + \left(\frac{\theta-\mu}{\tau}\right)^2\right]\right\}$

    $= \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2+\tau^2}\right\}\exp\left\{-\frac{1}{2}\left(\frac{\theta-\mu^*}{\tau^*}\right)^2\right\}$ *Use algebraic manipulation to re-express the joint likelihood*
- Integrate over θ to find the marginal likelihood:
  - $f(x|\mu,\tau,\sigma) = \int_\theta f(x|\theta,\sigma)g(\theta|\mu,\tau)d\theta = \frac{1}{\sqrt{2\pi}\tau}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2+\tau^2}\right\}\int_\theta \exp\left\{-\frac{1}{2}\left(\frac{\theta-\mu^*}{\tau^*}\right)^2\right\}d\theta$

    $= \frac{\sqrt{2\pi}\tau^*}{\sqrt{2\pi}\tau}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2+\tau^2}\right\} = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2+\tau^2}\right\}$ *Normal density function*
- Marginal likelihood for X is a normal distribution with mean μ and variance $\sigma^2 + \tau^2$
  - Mean of the predictive distribution for X is μ, the mean of the prior distribution for Θ
  - Variance of the predictive distribution for X is the sum of the prior variance of Θ and the variance of X given Θ
  - Standard deviation of the predictive distribution for X is $\sqrt{\sigma^2 + \tau^2}$

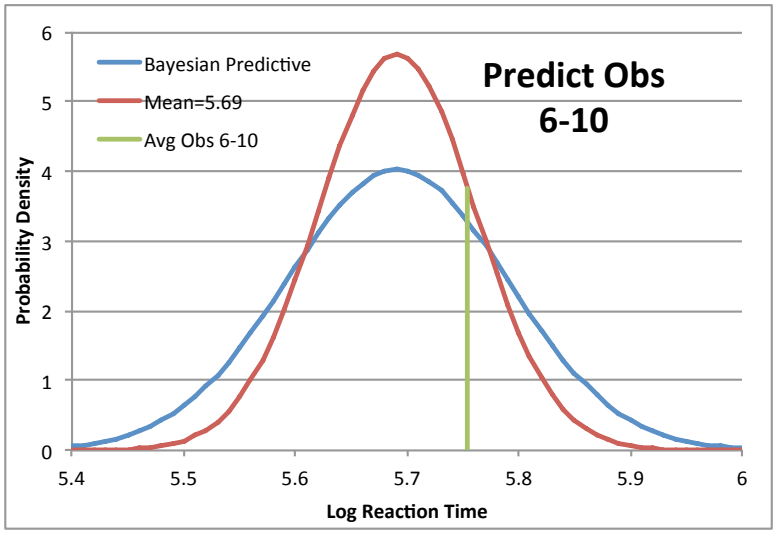
Marginal Likelihood for Sample Mean

- Assume $X_1, \ldots, X_n$ have a normal distribution with mean Θ and variance σ^2
- Assume the standard deviation σ is known, and Θ has a normal prior distribution with mean μ and variance τ^2
- From properties of the normal distribution, we find:
  - Conditional on $\Theta = \theta$, the sample mean $(\sum_i X_i)/n$ is normally distributed with mean θ and variance σ^2/n
- Marginally, integrating over θ:
  - The sample mean $(\sum_i X_i)/n$ is normally distributed with:
    - Mean μ
    - Variance $\sigma^2/n + \tau^2$
    - Standard deviation $\sqrt{\sigma^2/n + \tau^2}$

Using the Normal-Normal Marginal Likelihood to Predict the Mean of the Next 5 Observations

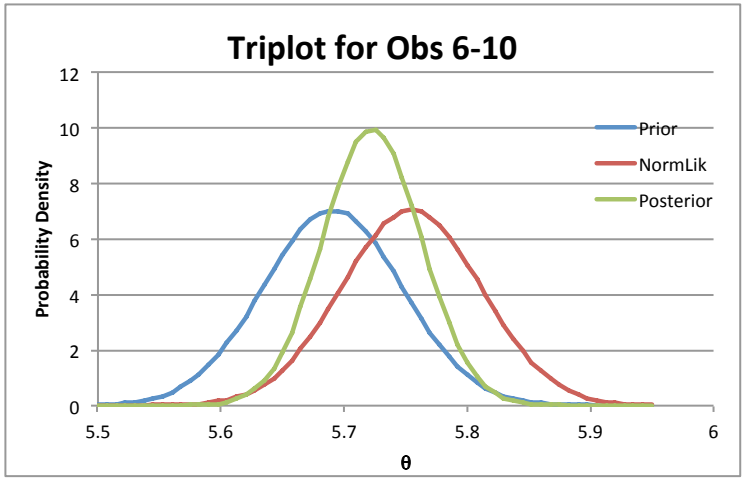
- Our current information about Θ is expressed by a normal distribution with mean μ_1 = 5.69, variance τ_1^2 = 0.127²/5 = 0.0032, std. dev τ_1 = 0.057
- The predictive distribution for the sample mean of the next 5 observations is a normal distribution with:
  - Mean μ_1 = 5.69
  - Variance $\sigma^2/n + \tau_1^2$ = 0.127²/5 + 0.0032 = 0.0064
  - Standard deviation 0.08
- A 95% predictive credible interval for the sample mean of the next 5 observations is [5.53, 5.85]
  - This interval includes both uncertainty about Θ and uncertainty about the sample mean given Θ
- Average of next 5 observations is 5.754
  - This value falls well within the expected range

Comparison of Bayesian predictive distribution with prediction based on point estimate



Updating After Next Batch of Reaction Times

- The observations are normal with unknown mean Θ and standard deviation σ = 0.127 (assumed known and equal to sample SD)
- We use the posterior distribution from our first batch of observations as our prior distribution – Normal(μ_1, τ_1)
  - Prior mean: μ_1 = 5.69
  - Prior standard deviation: τ_1 = 0.057
- Data: 5.793, 5.697, 5.875, 5.677, 5.730 (average is 5.754)
- The posterior distribution is normal with:
  - Posterior mean: $\mu_2 = \frac{\frac{5.69}{\tau_1^2} + \frac{\sum_i X_i}{\sigma^2}}{1/\tau_1^2 + 5/\sigma^2} = 5.723$
  - Posterior standard deviation $\tau_2 = \left(\sqrt{\frac{1}{\tau_1^2} + \frac{5}{\sigma^2}}\right)^{-1}$ =0.040
  - Posterior 95% credible interval for Θ: [5.59, 5.86]



Finding the Normalized Likelihood (for Triplot)

- The likelihood function: $f(\underline{x}|\theta,\sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^n \left(\frac{x_i-\theta}{\sigma}\right)^2\right\}$
- Re-express the sum of squares term:
  - $\sum_{i=1}^n \left(\frac{x_i-\theta}{\sigma}\right)^2 = \sum_{i=1}^n \frac{(x_i^2-2x_i\theta+\theta^2)}{\sigma^2} = \sum_{i=1}^n \frac{x_i^2}{\sigma^2} + \sum_{i=1}^n \frac{(-2x_i\theta+\theta^2)}{\sigma^2}$
    $= \sum_{i=1}^n \frac{x_i^2}{\sigma^2} - 2n\bar{x}/\sigma^2 + n\theta^2/\sigma^2$
    $= \sum_{i=1}^n \frac{x_i^2}{\sigma^2} - n\bar{x}^2/\sigma^2 + n(\bar{x}-\theta)^2/\sigma^2$ (complete the square)
- Re-express the likelihood: $\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^n \left(\frac{x_i-\theta}{\sigma}\right)^2\right\}$
  - $f(\underline{x}|\theta,\sigma) = \exp\left\{\sum_{i=1}^n \frac{x_i^2}{\sigma^2} - n\bar{x}^2/\sigma^2\right\} \times \exp\left\{-\frac{1}{2}\sum_{i=1}^n \left(\frac{\theta-\bar{x}}{\sigma/\sqrt{n}}\right)^2\right\}$
    (does not depend on θ) (proportional to $N(\bar{x}, \sigma/\sqrt{n})$)
- The likelihood function is proportional to a normal distribution with mean $\bar{x}$ and standard deviation $\sigma/\sqrt{n}$
- Therefore, the normalized likelihood is a normal density function with mean $\bar{x}$ and standard deviation $\sigma/\sqrt{n}$

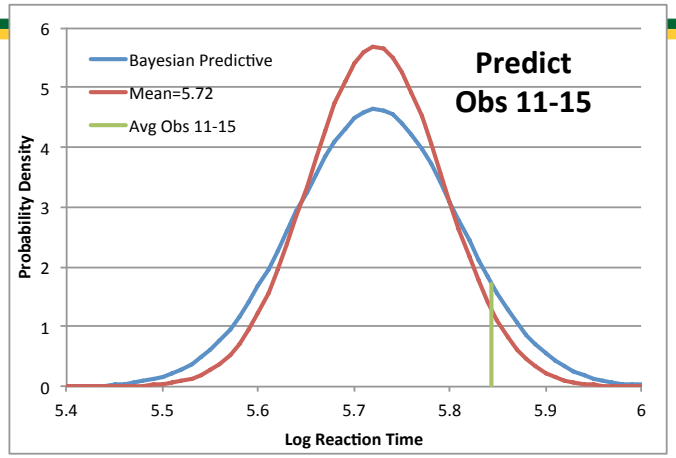
Sequential Prediction

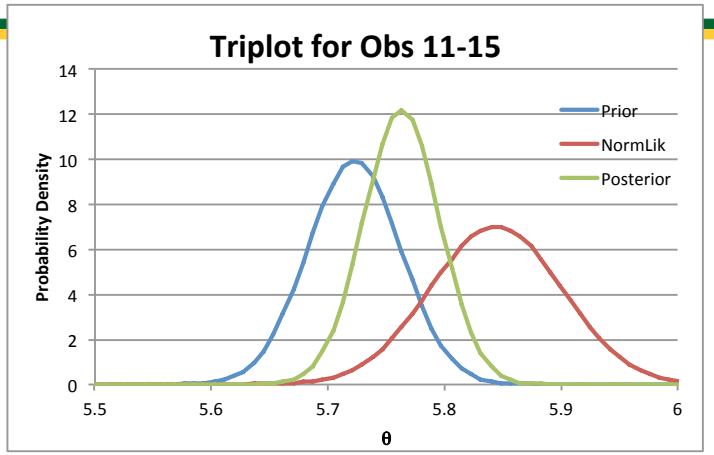
- Observations 1-5:
  - Predictive distribution for sample mean: uniform
  - Data: 5.743, 5.606, 5.858, 5.656, 5.591 (average is 5.691)
  - Posterior distribution: Normal(μ_1, τ_1); μ_1 = 5.69, τ_1 = 0.057
- Observations 6-10:
  - Predictive distribution for sample mean: Normal(μ_1, ω_1); $\omega_1^2 = \sigma^2/5 + \tau_1^2$; ω_1 = 0.0802
  - Data: 5.793, 5.697, 5.875, 5.677, 5.730 (average is 5.754)
  - Posterior distribution: Normal(μ_2, τ_2); μ_2 = 5.72, τ_2 = 0.040
- Observations 11-15:
  - Predictive distribution for sample mean: Normal(μ_2, ω_2); $\omega_2^2 = \sigma^2/5 + \tau_2^2$; ω_2 = 0.0692
  - Data: 5.690, 5.919, 5.981, 5.996, 5.635 (average is 5.844)
  - Posterior distribution: Normal(μ_3, τ_3); μ_3 = 5.76, τ_3 = 0.033
- Observations 16-20:
  - Predictive distribution for sample mean: Normal(μ_3, ω_3); $\omega_3^2 = \sigma^2/5 + \tau_3^2$; ω_3 = 0.0652
  - Data: 5.799, 5.537, 5.642, 5.858, 5.793 (average is 5.726)
  - Posterior distribution: Normal(μ_4, τ_4); μ_4 = 5.75, τ_4 = 0.028
- Observations 21-25:
  - Predictive distribution for sample mean: Normal(μ_4, ω_4); $\omega_4^2 = \sigma^2/5 + \tau_4^2$; ω_4 = 0.0632
  - Data: 5.805, 5.730, 5.677, 5.553, 5.829 (average is 5.719)
  - Posterior distribution: Normal(μ_5, τ_5); μ_5 = 5.74, τ_5 = 0.025
- Observations 26-30:
  - Predictive distribution for sample mean: Normal(μ_5, ω_5); $\omega_5^2 = \sigma^2/5 + \tau_5^2$; ω_5 = 0.0622
  - Data: 5.489, 5.724, 5.793, 5.684, 5.606 (average is 5.659)
  - Posterior distribution: Normal(μ_6, τ_6); μ_6 = 5.73, μ_6 = 0.023

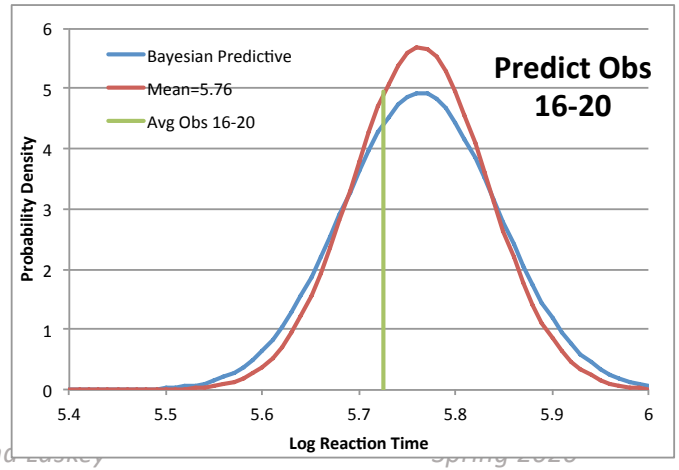
Reaction Time Data Sequential Updating with Unknown Mean and Known Variance

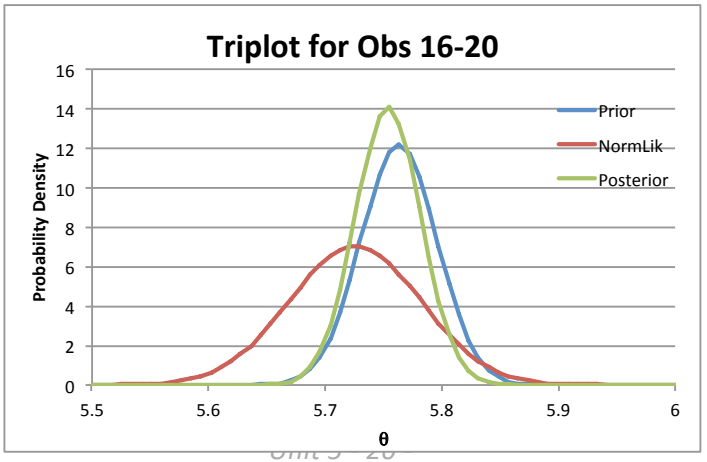
	Prior	Obs 1-5	Obs 6-10	Obs 11-15	Obs 16-20	Obs 21-25	Obs 26-30
Data		5.743	5.793	5.690	5.799	5.805	5.489
		5.606	5.697	5.919	5.537	5.730	5.724
		5.858	5.875	5.981	5.642	5.677	5.793
		5.656	5.677	5.996	5.858	5.553	5.684
		5.591	5.730	5.635	5.793	5.829	5.606
μ	0	5.69	5.72	5.76	5.75	5.75	5.73
τ	∞	0.057	0.040	0.033	0.028	0.025	0.023
Θ 2.5%	$-\infty$	5.58	5.64	5.70	5.70	5.70	5.69
Θ 97.5%	∞	5.80	5.80	5.83	5.81	5.80	5.78
$\bar{X}$ 2.5%	$-\infty$	5.53	5.59	5.63	5.63	5.62	
$\bar{X}$ 97.5%	∞	5.85	5.86	5.89	5.88	5.87	

Sequential Prediction of Next 2 Batches

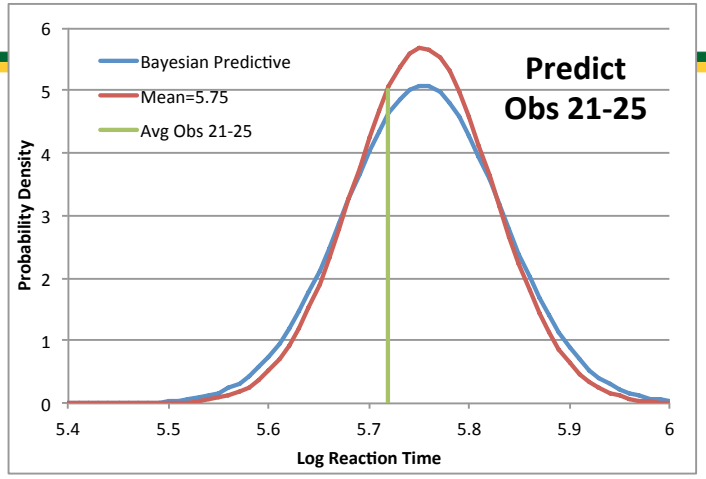


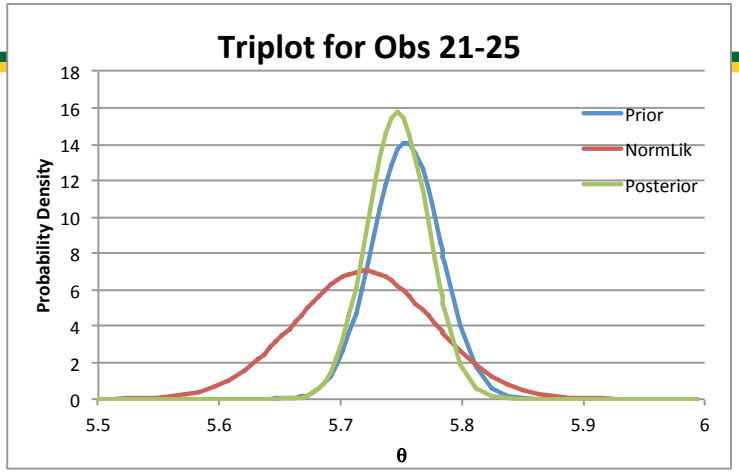


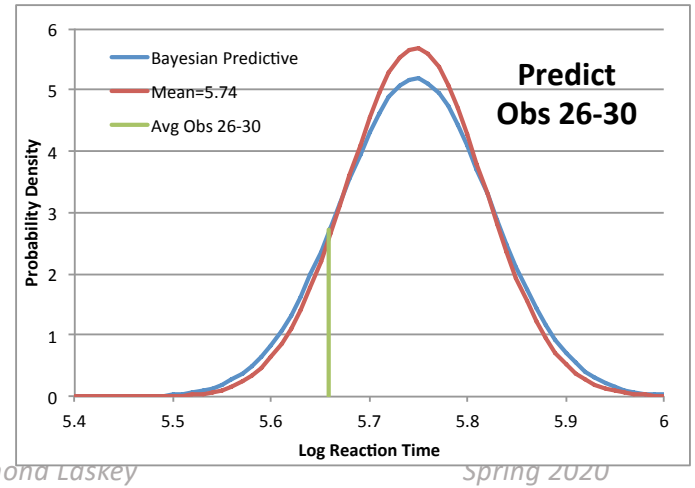


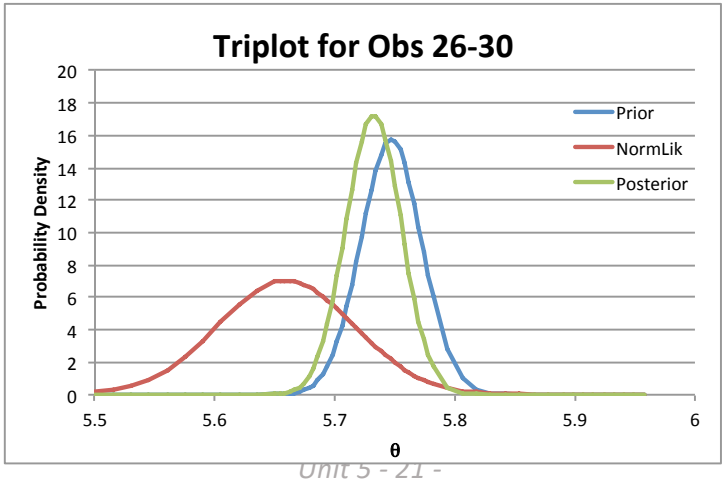


Sequential Prediction of Last 2 Batches







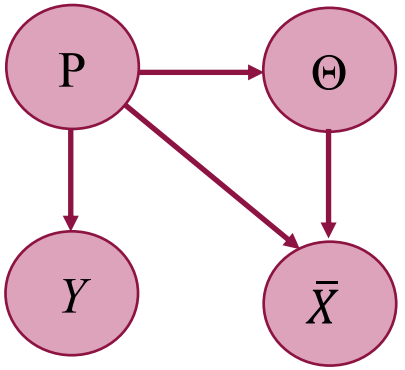


The Case of Unknown Standard Deviation

- Typically, the standard deviation is not known
  - We did not really know the standard deviation in our reaction time example
  - We treated the standard deviation as known to simplify the analysis
- If we are interested only in inferences about the mean, and if the sample size is not too small, we can get a reasonable approximation to the posterior distribution by treating the standard deviation as known and equal to the sample standard deviation
- A more accurate representation of our knowledge should account for the unknown standard deviation (or equivalently, precision)
- There is also a conjugate family for the joint distribution of the mean and precision

Sufficient Statistic for Parameters of Normal Distribution

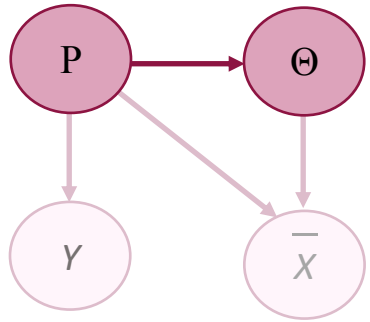
- Data $X_1, \ldots, X_n$ are a random sample from a normal distribution with unknown mean Θ and unknown precision P
- Sufficient statistic for (Θ, P):
  - Sample average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - Sample sum of squared deviations $Y = \sum_{i=1}^{n} (X_i - \bar{X})^2$



- Posterior distribution for (Θ, P) given $X_1, \ldots, X_n$ depends only on $\bar{X}$ and Y
- Conditional distribution of $(\bar{X}, Y)$ given (Θ, P)
  - $\bar{X}$ and Y are independent given Θ and P
  - $\bar{X}$ is normally distributed with mean Θ and precision $n\mathrm{P}$
  - Y has a Gamma distribution with parameters $(n-1)/2$ and $2/\mathrm{P}$

The Normal-Gamma Conjugate Prior for (Θ, P)

- When we do not know the precision, we can use the normal-gamma conjugate prior for (Θ, P)
  - Observations are normally distributed with mean Θ and precision P
  - Prior distribution for (Θ, P) is normal-gamma with hyperparameters:
    - μ (the center)
    - k (the precision multiplier)
    - α (the shape)
    - β (the scale)
- If (Θ, P) has a normal-gamma(μ, k, α, β) distribution, then:
  - P has a gamma(α, β) distribution (therefore, the variance $1/\mathrm{P}$ has an inverse-gamma(α, β) distribution)
  - Conditional on P, the mean Θ has a normal distribution with mean μ and precision $k\mathrm{P}$
- Joint prior density function for (Θ, P)
  - $g(\theta, \rho) = \left(\frac{k\rho}{2\pi}\right)^{1/2} \exp\left\{-\frac{k\rho}{2}(\theta-\mu)^2\right\} \frac{1}{\beta^\alpha \Gamma(\alpha)} \rho^{\alpha-1} e^{-\rho/\beta}$



Finding the Posterior Distribution for (Θ, P)

- The prior density is Normal-Gamma: $g(\theta,\rho)=\left(\frac{k\rho}{2\pi}\right)^{1/2}\exp\left\{-\frac{k\rho}{2}(\theta-\mu)^2\right\}\frac{1}{\beta^\alpha\Gamma(\alpha)}\rho^{\alpha-1}e^{-\rho/\beta}$
- The likelihood function is normal: $f(\underline{x}\,|\,\theta,\rho)=\left(\frac{\rho}{2\pi}\right)^{n/2}\exp\left\{-\frac{\rho}{2}\sum_i(x_i-\theta)^2\right\}$

$$\mu^*=\frac{k\mu+n\bar{x}}{k+n}$$
$$k^*=k+n$$
$$\alpha^*=\alpha+n/2$$
$$\beta^*=\left(\beta^{-1}+\tfrac{1}{2}\sum_i(x_i-\bar{x})^2+\frac{nk}{2(n+k)}(\bar{x}-\mu)^2\right)^{-1}$$

- Prior times likelihood:

$$f(\underline{x}|\theta,\rho)g(\theta,\rho)=\left(\frac{\rho}{2\pi}\right)^{n/2}\exp\left\{-\frac{\rho}{2}\sum_i(x_i-\theta)^2\right\}\left(\frac{k\rho}{2\pi}\right)^{1/2}\exp\left\{-\frac{k\rho}{2}(\theta-\mu)^2\right\}\frac{1}{\beta^\alpha\Gamma(\alpha)}\rho^{\alpha-1}e^{-\rho/\beta}$$
$$\propto\rho^{\frac{n}{2}+\frac{1}{2}+\alpha-1}\exp\left\{-\rho\left(\frac{1}{\beta}+\frac{k}{2}(\theta-\mu)^2+\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\right\}$$

- Algebraic manipulation of term inside parentheses:

$$\frac{1}{\beta}+\frac{k}{2}(\theta-\mu)^2+\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2=\frac{1}{\beta}+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{nk}{2(n+k)}(\bar{x}-\mu)^2+\frac{k+n}{2}\left(\theta-\frac{k\mu+n\bar{x}}{k+n}\right)^2=\frac{1}{\beta^*}+\frac{k^*}{2}(\theta-\mu^*)^2$$

- Re-express prior times likelihood:

$$f(\underline{x}|\theta,\rho)g(\theta,\rho)\propto\left(\rho^{1/2}\exp\left\{-\frac{k^*\rho}{2}(\theta-\mu^*)^2\right\}\right)\left(\rho^{\alpha^*-1}e^{-\rho/\beta^*}\right)$$

- *The first factor is proportional to a Normal density function with mean μ* and precision k*ρ*
- *The second factor is proportional to a Gamma density function with shape α* and scale β**

Updating Equations for Normal / Normal-Gamma Conjugate Pair

- Data $X_1, \ldots, X_n$ are a random sample from a normal distribution with unknown mean Θ and unknown standard deviation Σ, precision $\mathrm{P} = 1/\Sigma^2$
- (Θ, P) have normal-gamma joint distribution with hyperparameters μ, k, α, and β
  - Distribution for Θ given $\mathrm{P} = \rho$ is normal with mean μ and precision $k\rho$
  - Distribution for P is Gamma with parameters α and β
- Posterior distribution for Θ and P is Normal – Gamma μ^*, k^* α^*, and β^*
  - Posterior distribution for P is Gamma with hyperparameters

$$\alpha^* = \alpha + n/2 \text{ and } \beta^* = \left(\frac{1}{\beta} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{nk}{2(n+k)}(\bar{x} - \mu)^2\right)^{-1}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$n\bar{x} = \sum_i x_i$$

  - Posterior distribution for Θ given $\mathrm{P} = \rho$ is normal
    - Mean $\mu^* = (k\mu + \sum_i x_i)/(k+n)$
    - Precision $k^*\rho$, where $k^* = k + n$

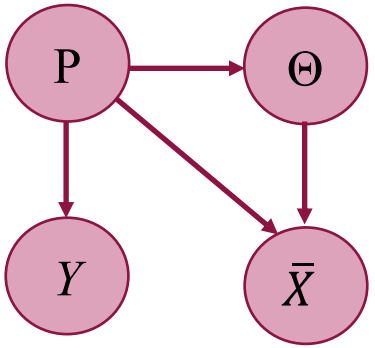
- *The case of known mean θ corresponds to infinite precision multiplier $k \rightarrow \infty$*
- *The case of known precision $\mathrm{P}=\rho$ corresponds to $\alpha \rightarrow \infty$ and $\beta \rightarrow 0$ while $\alpha\beta=\rho$*

Summary: Prior to Posterior Normal / Normal-Gamma Conjugate Pair

IF $X_1, \ldots, X_n$ are a random sample from Normal distribution with unknown mean Θ and precision P, and joint prior distribution for (Θ, P) is Normal-Gamma with center μ, precision multiplier k, shape α, and scale β

THEN Joint posterior distribution for (Θ, P) is Normal-Gamma with:

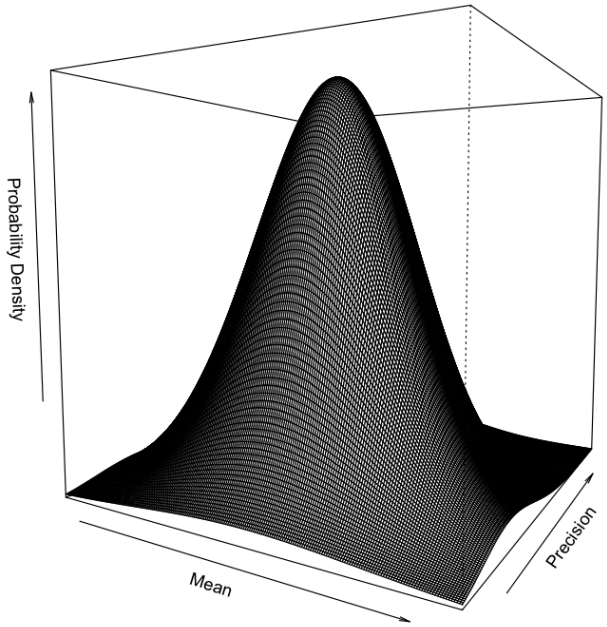
- Center $\mu^* = \frac{k\mu + n\bar{x}}{k+n}$,
- Precision multiplier $k^* = k + n$,
- Shape $\alpha^* = \alpha + n/2$, and
- Scale $\beta^* = \left(\frac{1}{\beta} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{nk}{2(n+k)}(\bar{x} - \mu)^2\right)^{-1}$



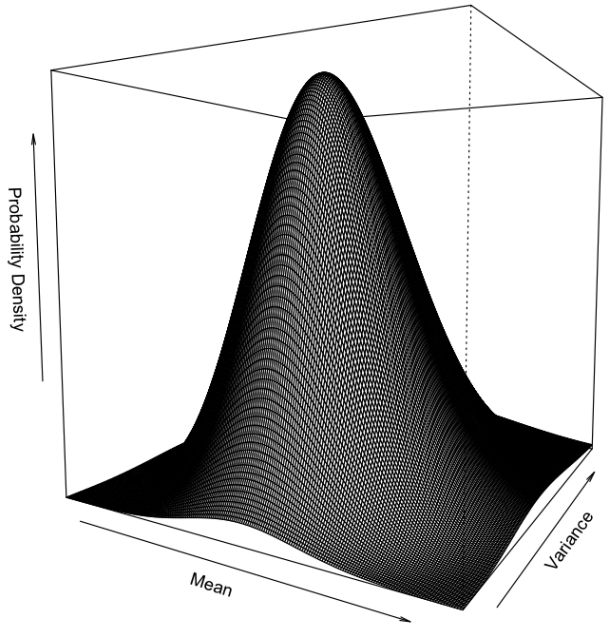
Reaction Times Revisited

- Assume both mean and precision are unknown
- Prior distribution: $g(\theta, \rho) \propto \rho^{-1}$ (uniform on θ, decreasing in ρ)
  - Commonly used reference prior; improper (integrates to ∞)
  - Proportional to a normal-gamma density with $\mu = 0$, $k = 0$, $\alpha = -\frac{1}{2}$, $\beta = \infty$
- Data:
  - Number of observations: $n = 30$
  - Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 5.73$
  - Sum of squared deviations: $Y = \sum_{i=1}^{n} (X_i - \overline{X})^2 = 0.465$
- Posterior distribution for (Θ, P) after 30 observations is normal-gamma with:
  - Posterior center: $\mu^* = \bar{x} = 5.73$
  - Posterior precision multiplier $k^* = n = 30$
  - Posterior shape: $\alpha^* = -1/2 + 30/2 = 14.5$
  - Posterior scale: $\beta^* = \left(\frac{1}{2}0.465\right)^{-1} = 4.30$

Reaction Time Data: Joint Posterior Density for Mean and Precision / Variance

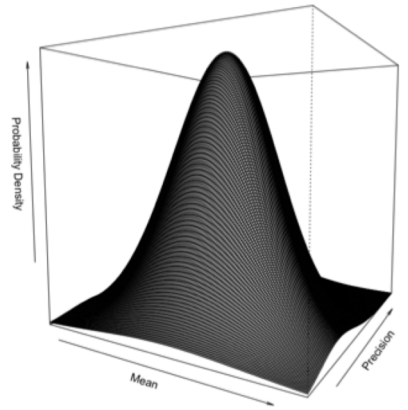


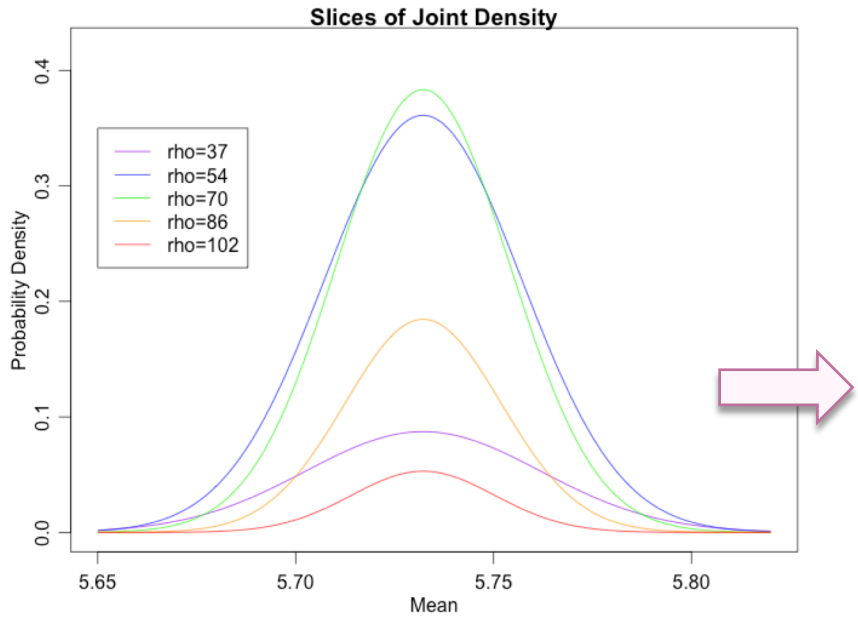
3D Perspective Plot of Bivariate Posterior Density for Mean and Precision

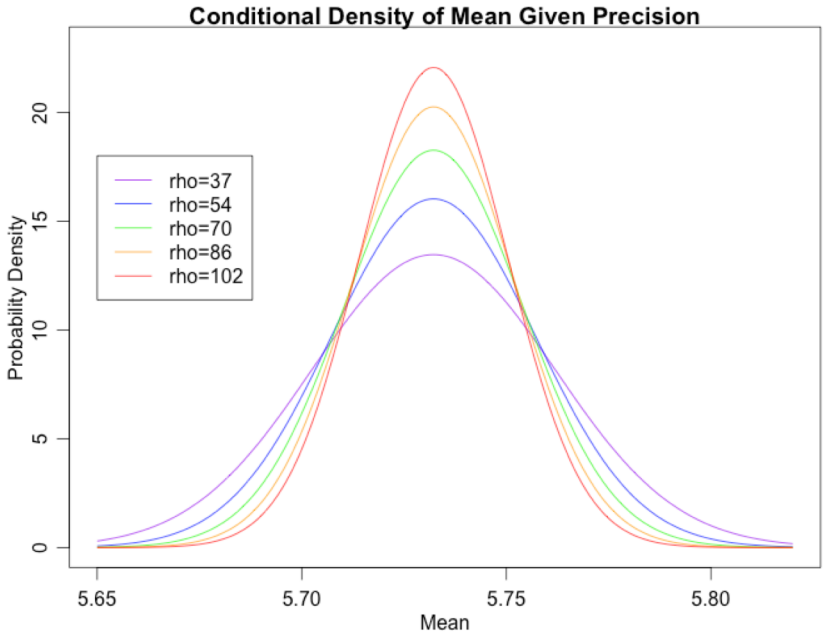


3D Perspective Plot of Bivariate Posterior Density for Mean and Variance

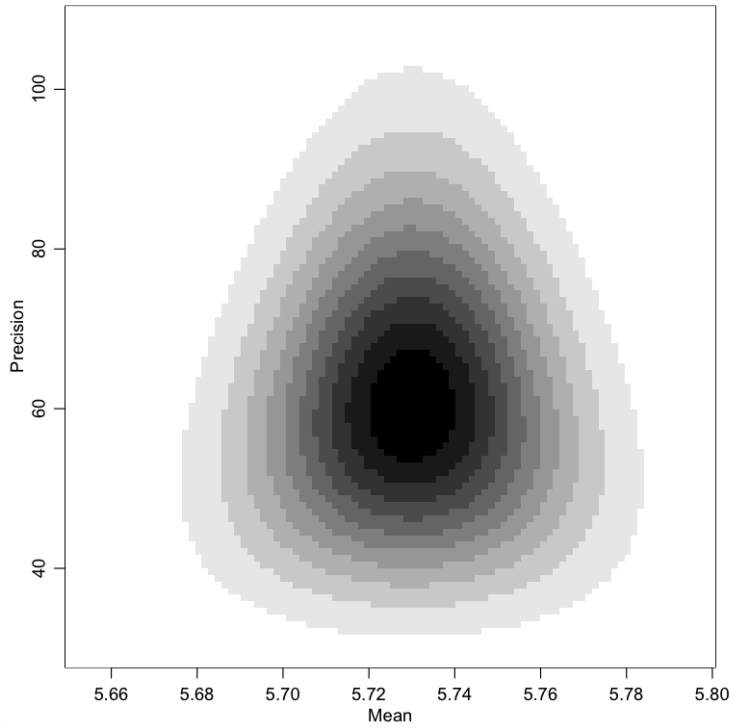
Conditional Distribution of Θ for Different Values of P



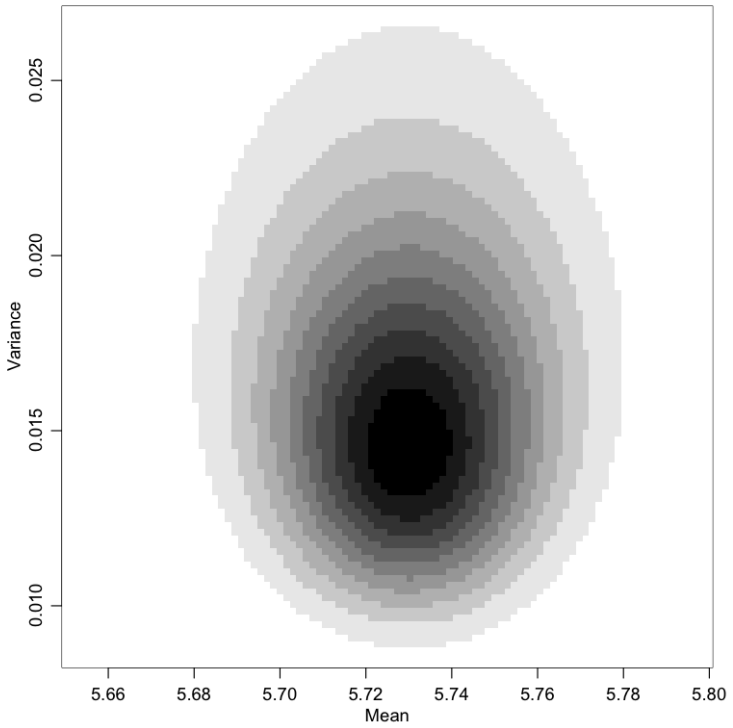




Reaction Time Joint Posterior Density Contour Plots



Contour Plot of Bivariate Posterior Density for Mean and Precision



Contour Plot of Bivariate Posterior Density for Mean and Variance

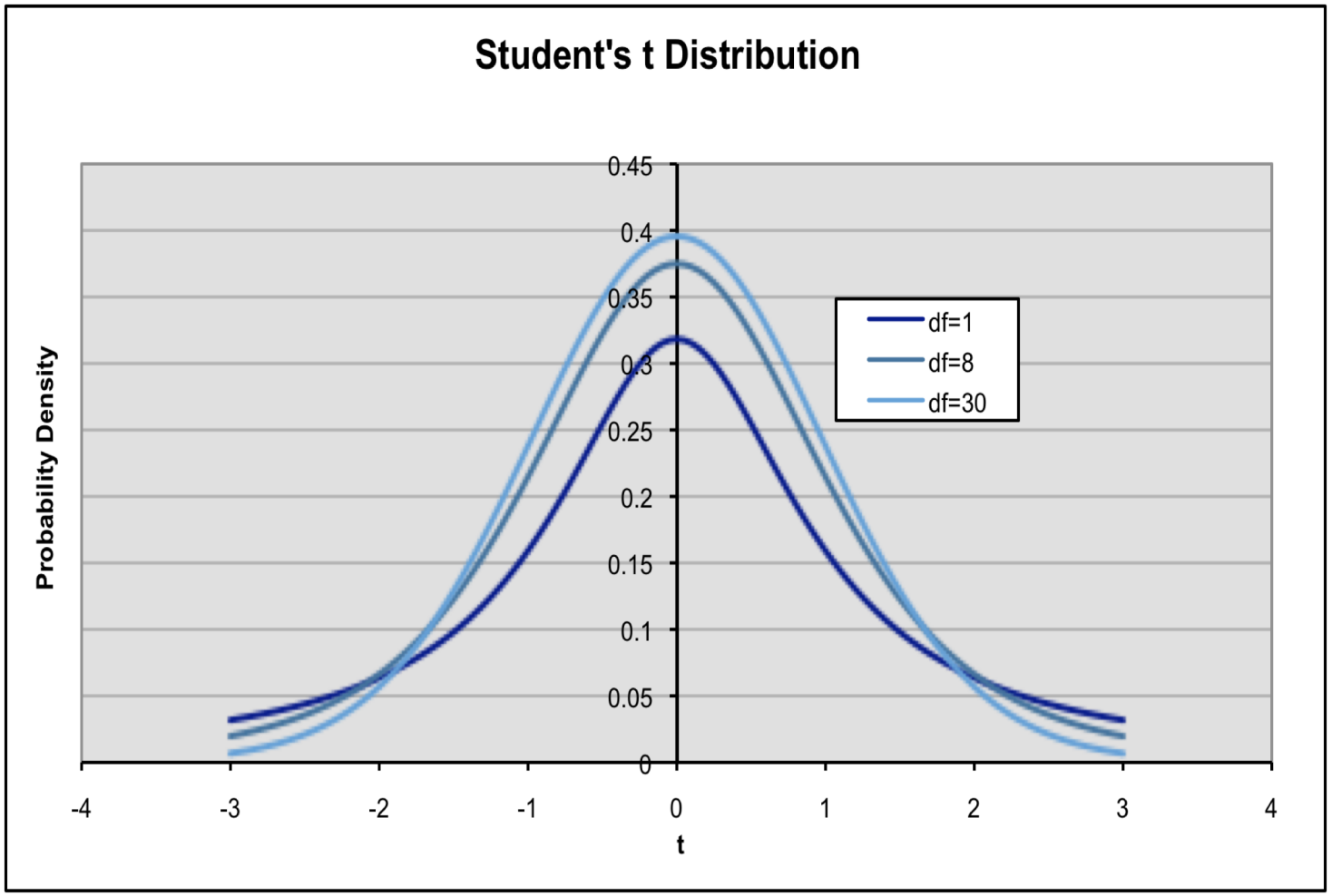
Credible Intervals for Parameters of Normal-Gamma(μ, k, α, β) Distribution

- To find credible interval for the precision P of a normal-gamma distribution:
  - P has a gamma distribution with shape α and scale β
  - Use gamma quantiles to find credible interval
  - A symmetric 100(1-c)% credible interval for P is $[g_{c/2}, g_{1-c/2}]$
    - Endpoints are the $c/2$ and $1-c/2$ quantiles of a gamma(α, β) distribution
- To find credible interval for the variance $\mathrm{V} = \mathrm{P}^{-1}$:
  - A symmetric 100(1-c)% credible interval for V is $[(g_{1-c/2})^{-1}, (g_{c/2})^{-1}]$
- To find credible interval for the standard deviation $\Sigma = \mathrm{P}^{-1/2}$:
  - A symmetric 100(1-c)% credible interval for Σ is $[(g_{1-c/2})^{-1/2}, (g_{c/2})^{-1/2}]$
- To find credible interval for mean Θ:
  - Find marginal distribution of Θ by integrating over P
  - Use quantiles of marginal distribution to find credible interval for Θ

Marginalizing Out the Unknown Precision

- If (Θ, P) has a normal - gamma distribution with hyperparameters μ, k, α, and β then:
  - $\sqrt{k\alpha\beta}(\Theta - \mu)$ has a t distribution with 2α degrees of freedom
- We say Θ has nonstandard t distribution with:
  - Center μ
  - Spread $\left(\sqrt{k\alpha\beta}\right)^{-1}$
  - Degrees of freedom 2α

Marginal distribution for Θ



t distribution with 1 df is called a Cauchy distribution

Non-Standard Student's *t* Distribution

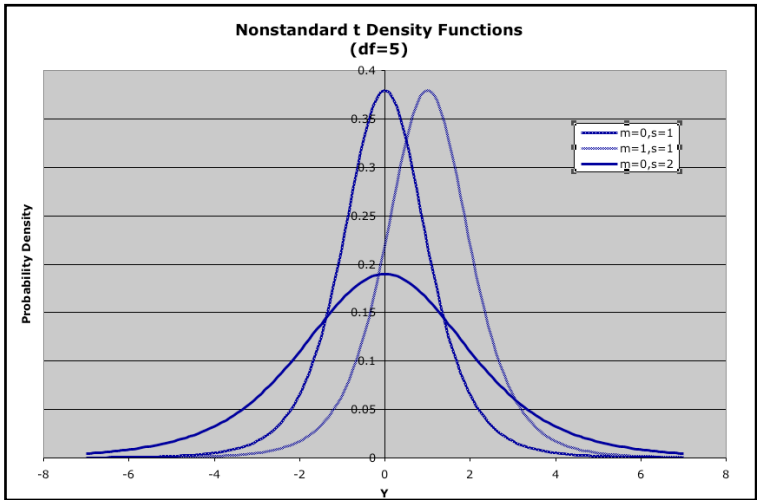
- Non-standard Student's t with center m and spread s and degrees of freedom ν:
  - $Y = sT + m$ where T has the standard Student t distribution with degrees of freedom ν
- Density function for non-standard Student t:

$$f(y \mid m, s, \nu) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)s}\left(1 + \frac{\left((y-m)/s\right)^2}{\nu}\right)^{-(\nu+1)/2}$$

- t distribution in R has optional non-centrality parameter but no scale parameter

$Y = sT + m$ is a location and scale transformation of a standard t random variable. Therefore, the density function $g(y)$ for y is equal to the t density function evaluated at $(y\text{-}m)/s$ divided by s.

Do you know why we divide by s?



Marginal Distribution for Precision, Variance and Standard Deviation

- If (Θ, P) has a normal - gamma distribution with hyperparameters μ, k α, and β then:
  - Distribution for P is gamma with hyperparameters α and β

$$g_\rho(\rho|\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \rho^{\alpha-1} e^{-\rho/\beta}$$

  - Distribution for $V = 1/\mathrm{P}$ is inverse-gamma with hyperparameters α and β

$$g_v(v \mid \alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} v^{-(\alpha+1)} e^{-1/(v\beta)}$$

  - Distribution for $\Sigma = \mathrm{P}^{-1/2}$ has density function

$$g_\sigma(\sigma \mid \alpha,\beta) = \frac{2}{\beta^\alpha \Gamma(\alpha)} \sigma^{-2\alpha-1} e^{-1/(\sigma^2\beta)}$$

Can you derive the density functions for V and Σ?

If X has density function $f_x(x)$ and $y=r(x)$ is a continuous transformation that is either strictly increasing or strictly decreasing, then $Y = r(X)$ has density function:

$$f_y(y) = f_x(s(y))\left|\frac{ds(y)}{dy}\right|$$ where $s(y) = r^{-1}(y)$ is the inverse function of $r(x)$

Credible Intervals for Parameters: Summary

- Suppose (Θ, P) has a Normal - Gamma distribution with hyperparameters μ, k, α, and β
- A symmetric 100(1-c)% credible interval for P is $[g_{c/2}, g_{1-c/2}]$
  - Endpoints are the $c/2$ and $1 - c/2$ quantiles of a gamma(α, β) distribution
- A symmetric 100(1-c)% credible interval for Σ is $[\left(g_{1-c/2}\right)^{-1/2}, \left(g_{c/2}\right)^{-1/2}]$
- To find credible interval for mean Θ:
  - Use quantiles of marginal t distribution
  - Find $t_{1-c/2}$, the $1 - c/2\%$ point for the t distribution with 2α degrees of freedom
  - Let $s = 1/(k\alpha\beta)^{1/2}$
  - The symmetric 100(1-c)% credible interval for Θ is $[\mu - t_{1-c/2}s, \mu + t_{1-c/2}s]$

Credible Intervals for Parameters of Reaction Time Distribution

- The posterior distribution for (Θ, P) is normal-gamma with parameters:
  - Posterior center: μ_1 = 5.73
  - Posterior precision multiplier: k_1 = 30
  - Posterior shape: α_1 = 14.5
  - Posterior scale: β_1 = 4.30
- 95% credible interval for Θ and P
  - The 0.975 quantile of the t distribution with 4 degrees of freedom is `qt(0.975,4)` = 2.045
  - The spread is s = $1/(30 \times 14.5 \times 4.30)^{1/2}$ = 0.0231
  - $[\mu_1 - t_{0.975}s, \mu_1 + t_{0.975}s]$ = [5.73 – 2.045x0.0231, 5.73 + 2.045x0.0231] = [5.68, 5.78]
  - The 0.025 and 0.975 quantiles of the gamma(14.5, 4.30) distribution are 34.52 and 98.36
  - Credible interval for Θ is [5.68, 5.78] and credible interval for P is [34.52, 98.36]
- 95% credible interval for Σ
  - $[(g_{0.975})^{-1/2}, (g_{0.025})^{-1/2}]$ = $[(98.36)^{-1/2}, (34.52)^{-1/2}]$ = [0.101, 0.170]

Hoff's Conjugate Prior

- The Hoff text uses an alternate parameterization for a less general form of the Normal-Gamma prior:
  - μ_0 (the center)
  - ν_0 (the degrees of freedom)
  - σ_0 (the spread)
- The prior distribution:
  - Mean given precision: $\Theta \mid \mathrm{P} \sim \text{Normal}(\mu_0, 1/\sqrt{\mathrm{P}})$
  - Precision: $\mathrm{P} \sim \text{Gamma}(\nu_0/2, 2/(\nu_0\sigma_0^2))$
    (note: scale is $2/(\nu_0\sigma_0^2)$; rate is $(\nu_0\sigma_0^2)/2$
- This is a special case of the our Normal-Gamma prior:
  - Center: μ_0
  - Precision multiplier: ν_0
  - Shape: $\nu_0/2$
  - Scale: $2/(\nu_0\sigma_0^2)$

In the Hoff conjugate prior the precision multiplier is always twice the shape

Sequential Prediction of Batches (Revisited)

- We will revisit the sequential prediction problem of log reaction times when both mean and precision are unknown
  - Receive observations in batches of 5
  - Update distribution for mean and precision after each batch
  - Assume observations are normal with unknown mean and precision
- After each batch we update our distribution for (Θ, P) and predict the sample mean of the next batch of observations
  - Use normal / normal-gamma conjugate updating
- To predict the next batch of observations we will need to know the predictive distribution (marginal likelihood) of the sample mean

Marginalizing Out the Unknown Precision

- If $X_1, \ldots, X_n$ has a normal distribution with mean Θ and precision P, and (Θ, P) has a normal - gamma distribution with hyperparameters μ, k, α, and β then:
  - $\sqrt{\frac{kn}{k+n}\alpha\beta}(\bar{X} - \mu)$ has a t distribution with 2α degrees of freedom
- The marginal distribution of $\bar{X}$ is a nonstandard t distribution with:
  - Center μ
  - Spread $\left(\sqrt{\frac{kn}{k+n}\alpha\beta}\right)^{-1}$
  - Degrees of freedom 2α

Marginal likelihood for $\bar{X}$

Same center and degrees of freedom as distribution for Θ but spread is larger because we are uncertain about both Θ and $\bar{X}$ given Θ

- Compare with marginal distribution of Θ, also a nonstandard t distribution:
  - Center μ
  - Spread $\left(\sqrt{k\alpha\beta}\right)^{-1}$
  - Degrees of freedom 2α

Marginal distribution for Θ

First Batch: Posterior and Predictive Distributions

- Prior distribution
  - Normal-Gamma(μ, k, α, β) distribution with $\mu = 0,\ k = 0,\ \alpha = -½,\ \beta = \infty$
  - Improper reference prior $g(\theta, \rho \mid \mu, k, \alpha, \beta) \propto \rho^{-1}$
- First batch of data: 5.74, 5.61, 5.86, 5.66, 5.59
- Posterior distribution for (Θ, P) is normal-gamma with:
  - Posterior center: $\mu_1 = 5.69$
  - Posterior precision multiplier: $k_1 = 5$
  - Posterior shape: $\alpha_1 = 2$
  - Posterior scale: $\beta_1 = 40.78$
- Marginal distribution of P is gamma with shape $\alpha_1 = 2$ and scale $\beta_1 = 40.78$
- Marginal distribution of Θ is nonstandard t with center $\mu_1 = 5.69$, spread $1/(k_1\alpha_1\beta_1)^{-1/2} = 0.050$, and degrees of freedom 4
- Predictive distribution for sample mean of next batch is nonstandard t with center $\mu_1 = 5.69$, spread $s_1 = 1/\left(\left(\frac{5k_1}{k_1+5}\right)\alpha_1\beta_1\right)^{-1/2} = 0.070$, and degrees of freedom 4

Sequential Prediction

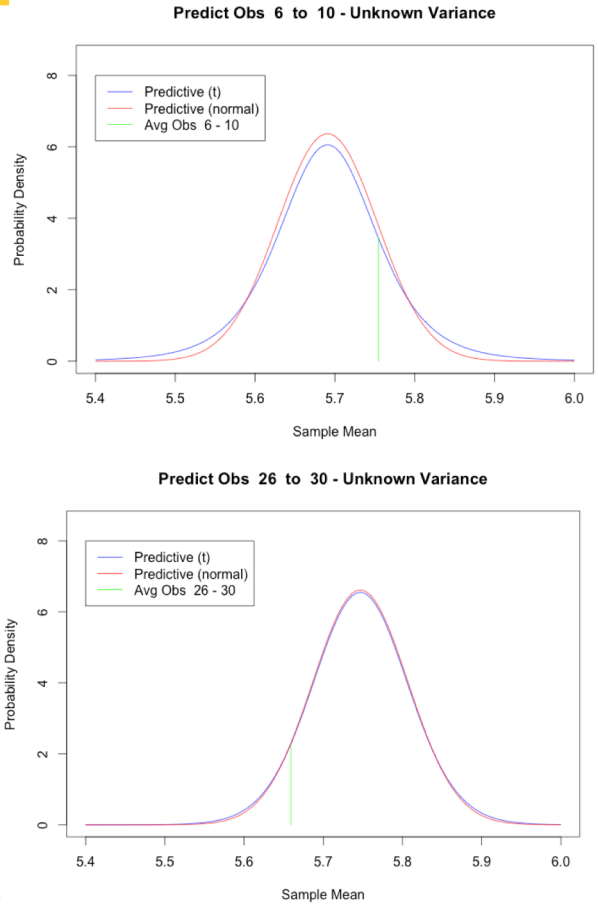
- Observations 1-5:
  - Predictive distribution for sample mean: uniform
  - Data: 5.743, 5.606, 5.858, 5.656, 5.591 (average is 5.691)
  - Posterior distribution: Normal-gamma($\mu_1, k_1, \alpha_1, \beta_1$); $\mu_1 = 5.69, k_1$=5, $\alpha_1 = 2, \beta_1 = 40.78$
- Observations 6-10:
  - Predictive distribution for sample mean: $t(\mu_1, s_1, d_1)$; $\mu_1 = 5.69, s_1$=0.070, d_1=4
  - Data: 5.793, 5.697, 5.875, 5.677, 5.730 (average is 5.754)
  - Posterior distribution: Normal-gamma($\mu_1, k_1, \alpha_1, \beta_1$); $\mu_2 = 5.72, k_2$=10, $\alpha_2 = 4.5, \beta_2 = 23.51$
- Observations 11-15:
  - Predictive distribution for sample mean: $t(\mu_2, s_2, d_2)$; $\mu_2 = 5.72, s_2$=0.053, d_2=9
  - Data: 5.690, 5.919, 5.981, 5.996, 5.635 (average is 5.844)
  - Posterior distribution: Normal-gamma($\mu_3, k_3, \alpha_3, \beta_3$); $\mu_3 = 5.76, k_3$=15, $\alpha_3 = 7, \beta_3 = 8.024$
- Observations 16-20:
  - Predictive distribution for sample mean: $t(\mu_3, s_3, d_3)$; $\mu_3 = 5.76, s_3$=0.069, d_3=14
  - Data: 5.799, 5.537, 5.642, 5.858, 5.793 (average is 5.726)
  - Posterior distribution: Normal-gamma($\mu_4, k_4, \alpha_4, \beta_4$); $\mu_4 = 5.75, k_4$=20, $\alpha_4 = 9.5, \beta_4 = 6.163$
- Observations 21-25:
  - Predictive distribution for sample mean: $t(\mu_4, s_4, d_4)$; $\mu_4 = 5.72, s_4$=0.065, d_4=19
  - Data: 5.805, 5.730, 5.677, 5.553, 5.829 (average is 5.719)
  - Posterior distribution: Normal-gamma($\mu_5, k_5, \alpha_5, \beta_5$); $\mu_5 = 5.75, k_5$=25, $\alpha_5 = 12, \beta_5 = 5.286$
- Observations 26-30:
  - Predictive distribution for sample mean: $t(\mu_5, s_5, d_5)$; $\mu_5 = 5.72, s_5$=0.062, d_5=24
  - Data: 5.489, 5.724, 5.793, 5.684, 5.606 (average is 5.659)
  - Posterior distribution: Normal-gamma($\mu_6, k_6, \alpha_6, \beta_6$); $\mu_6 = 5.73, k_6$=30, $\alpha_6 = 14.5, \beta_6 = 4.302$

Reaction Time Data Sequential Updating with Unknown Mean and Unknown Variance

	Prior	Obs 1-5	Obs 6-10	Obs 11-15	Obs 16-20	Obs 21-25	Obs 26-30
Data		5.743	5.793	5.690	5.799	5.805	5.489
		5.606	5.697	5.919	5.537	5.730	5.724
		5.858	5.875	5.981	5.642	5.677	5.793
		5.656	5.677	5.996	5.858	5.553	5.684
		5.591	5.730	5.635	5.793	5.829	5.606
μ	0	5.69	5.72	5.76	5.75	5.75	5.73
k	0	5	10	15	20	25	30
α	-0.5	2	4.5	7	9.5	12	14.5
β	∞	40.78	23.51	8.02	6.16	5.29	4.30
Θ 2.5%	$-\infty$	5.55	5.65	5.69	5.69	5.69	5.68
Θ 97.5%	∞	5.83	5.78	5.84	5.81	5.80	5.78
Σ 2.5%	0	0.066	0.067	0.098	0.099	0.098	0.101
Σ 97.5%	∞	0.318	0.177	0.210	0.191	0.175	0.170
$\bar{X}$ 2.5%	$-\infty$	5.50	5.60	5.62	5.62	5.62	
$\bar{X}$ 97.5%	∞	5.89	5.84	5.91	5.89	5.86	

MASON

Predicting Mean of Next 5 Observations: Comparison Between Normal and t



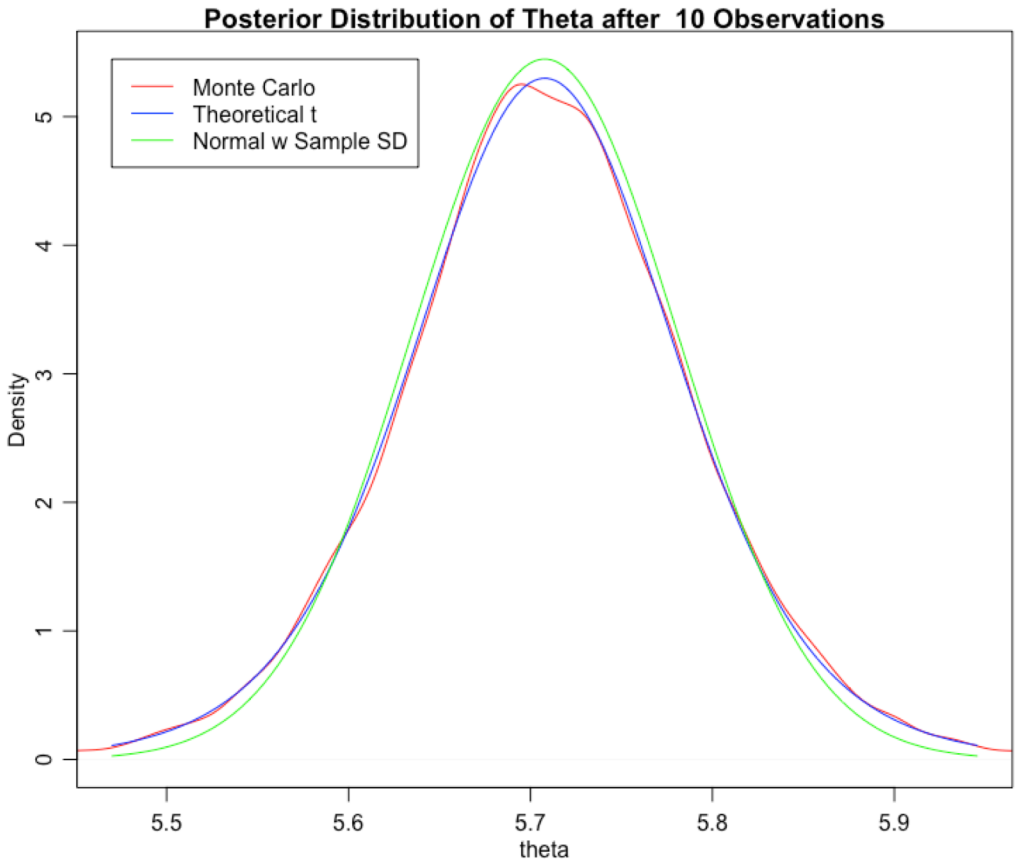
- We compare two predictive distributions for the sample mean of the next 5 observations:
  - Normal predictive distribution with point estimate for variance equal to sample SD
  - t predictive distribution
- Predictions of 2 batches are shown:
  - Predicting Obs 6-10: df=4
  - Predicting Obs 26-30: df=24
- The t distribution has heavier tails with smaller degrees of freedom
- With about 30 degrees of freedom, the predictive distributions are nearly identical

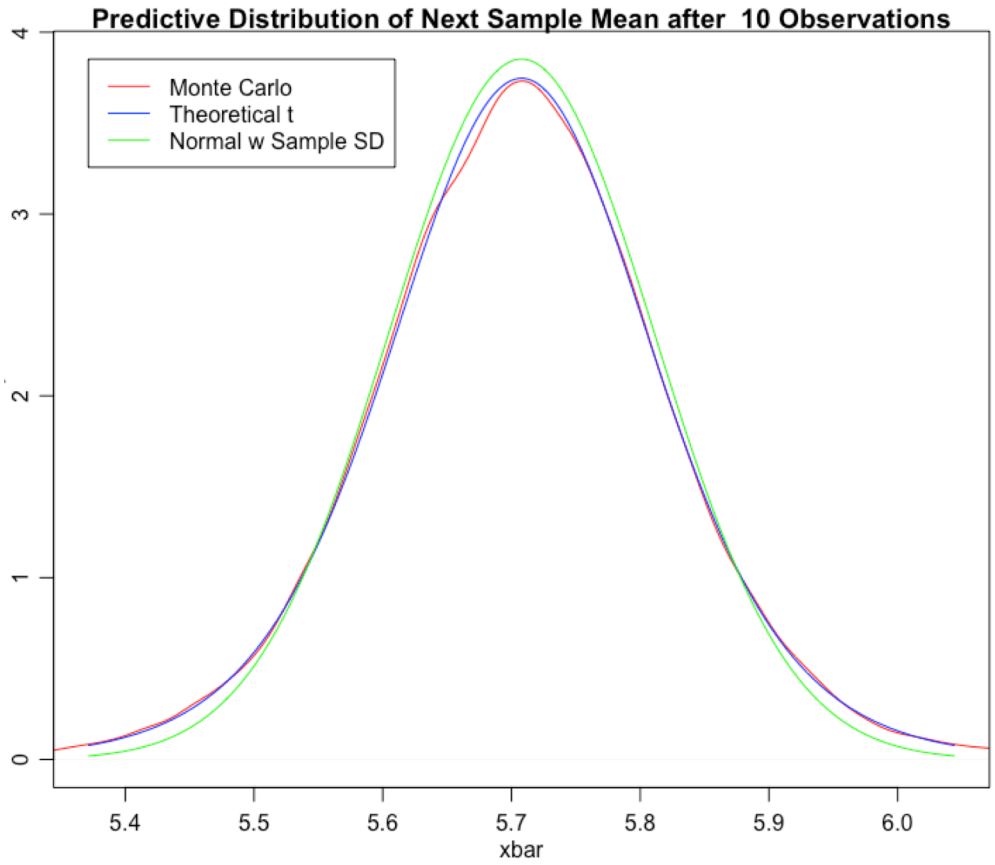
Which Conjugate Prior to Use?

- If we are interested in drawing inferences about an unknown variance, then we need to treat both mean and variance as uncertain
- If our interest is inference about the mean or predicting a future sample mean, and we have more than 30 observations, and we have vague prior information about the standard deviation, then we will get nearly the same results by assuming the variance is known and equal to the sample variance

Sampling from Posterior and Predictive Distributions

- The posterior distribution for the mean and precision (Θ, P) of the reaction time distribution is a normal-gamma(μ^*, k^*, α^*, β^*) distribution
- We can sample from the posterior distribution of the mean and standard deviation as follows:
  - Sample precisions $\rho_1, \rho_2, \ldots, \rho_m$ from a Gamma(α^*, β^*) distribution
  - Sample means θ_i from a Normal(μ^*, $(k^*\rho_i)^{-1/2}$) distribution, for $i = 1, \ldots, m$
  - Calculate $\sigma_i = \sqrt{1/\rho_i}$ for $i = 1, \ldots, m$
  - We now have a sample (θ_1, σ_1), (θ_2, σ_2), …, (θ_m, σ_m) from the posterior distribution for the mean and standard deviation (Θ, Σ)
- Using this sample, we can sample *n* observations from the posterior predictive distribution for $\bar{X}$ as follows:
  - For $i = 1, \ldots, m$, sample X_{ij} from a Normal(θ_i, σ_i) distribution, for $j = 1, \ldots, n$ and calculate the sample average





Comparison of kernel density estimate based on 20,000 direct MC draws, theoretical t density, and normal density using sample standard deviation (R code provided)

R Code for Exact Posterior and Monte Carlo Estimates

```
# Read the reaction time data into a table
reaction<-read.table("NonSchizReactionTime.txt")

# Focus on first non-schizophrenic subject
x <- reaction[,1]
xbar <- mean(x)
n <- length(x)

# Assume non-informative prior distribution
# Normal-Gamma with mu0=0, k0=0, alpha0=-1/2, beta0=infinity
# Posterior hyperparameters mu1, k1, alpha1, beta1
mu1 <- xbar
k1 <- n
alpha1 <- -1/2 + n/2
beta1 <- 1/(0.5*sum((x-xbar)^2))
spread1 <- sqrt(1/(k1*alpha1*beta1))

#Theoretical marginal density for theta
thetaVals <- 5.64+(0:100)/500
stdVals <- (thetaVals - mu1)/spread1
thetaMargDens <- dt(stdVals,df=2*alpha1)/spread1

#Set simulation sample size
numSim <- 10000

# Simulate directly from the posterior distribution
rhoDirect <- rgamma(numSim,shape=alpha1,scale=beta1)
thetaDirect <- rnorm(numSim,mean=mu1,sd=1/sqrt(k1*rhoDirect))

#Plot theoretical and Monte Carlo density functions
plot(density(thetaDirect),col="darkgreen",lty=2,main="",xlab="Theta")
lines(thetaVals,thetaMargDens,col="red")
legend(5.76,15,c("Monte Carlo","Theoretical t"),col=c("darkgreen","red"),lty=c(2,1))
```

Point Estimators: Bias and Variance

- A point estimator is a transformation of the data into a single element of the parameter space
  - Examples: sample mean; posterior mean
- We consider the frequency properties of different estimators by conditioning on the parameter and treating the data as random
- An *unbiased* estimator of Θ has expected value equal to Θ
  - The sample mean is an unbiased estimator of the population mean

    $$E[\bar{X}|\theta] = \theta$$
  - In the normal conjugate model, the posterior mean is biased

    $$E[\mu^*|\theta] = E[w\mu + (1-w)\bar{X}|\theta] = w\mu + (1-w)\theta$$

    $w = \frac{1/\tau^2}{1/\tau^2 + n/\sigma^2}$ or $\frac{k}{k+n}$
- Sample mean has a higher variance than posterior mean

$$\mathrm{Var}\left[\bar{X} \mid \theta,\sigma^2\right] = \mathrm{E}\left[(\bar{X}-\theta)^2 \mid \theta,\sigma^2\right] = \sigma^2/n$$

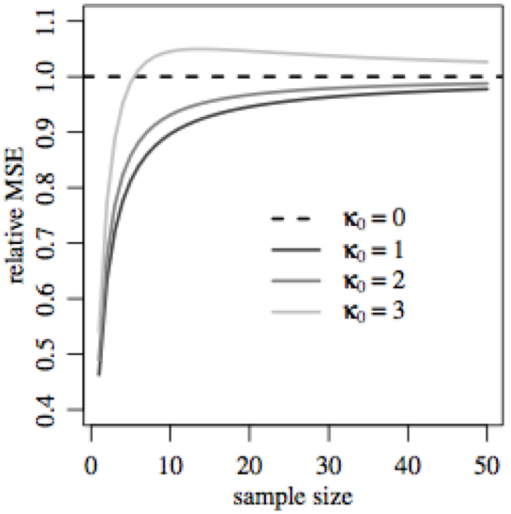
$$\mathrm{Var}\left[\mu^* \mid \theta,\sigma^2\right] = \mathrm{E}\left[\left(\mu^* - \left(E[\mu^*]\right)\right)^2 \mid \theta,\sigma^2\right]$$

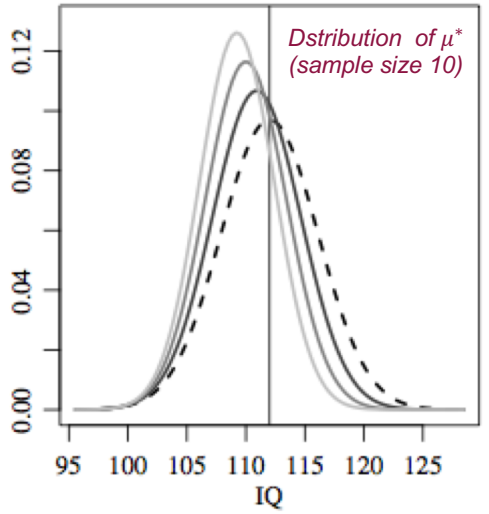
$$= \mathrm{E}\left[\left((1-w)(\bar{X}-\theta)\right)^2 \mid \theta,\sigma^2\right] = (1-w)^2\sigma^2/n$$

Point Estimators: Mean Squared Error

- Statisticians often use mean squared error (MSE) to evaluate accuracy of an estimator $\hat{\theta}$
  - $\mathrm{MSE}\left[\hat{\theta} \mid \theta\right] = \mathrm{E}\left[(\hat{\theta} - \theta)^2\right] = \mathrm{E}\left[(\hat{\theta} - \mathrm{E}[\hat{\theta}])^2 + (\mathrm{E}[\hat{\theta}] - \theta)^2\right]$
  - MSE = Variance + Bias^2
- An estimator with low bias and low variance may have better MSE than an unbiased estimator with a higher variance
- If we can make a reasonable guess at the mean, a Bayesian estimator is better even from the frequentist viewpoint than the sample mean
- This difference is more pronounced for high-dimensional parameter spaces

IQ example from Hoff (p. 82-83): For small virtual sample size the posterior mean has smaller MSE than the sample mean, with greater difference for small sample sizes





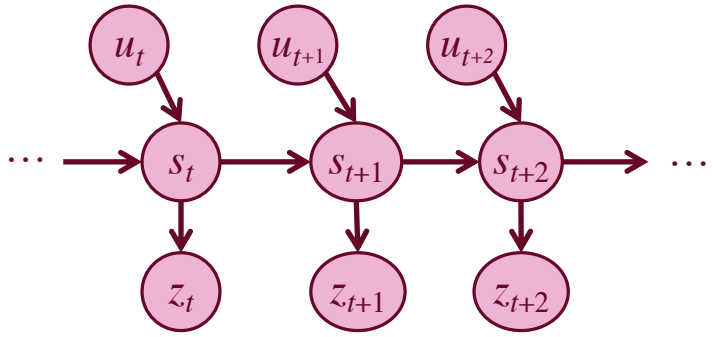
$X_i \sim_{iid} Normal(\theta, 15)$, $\theta \sim Normal\left(100, \sqrt{15^2/\kappa_0}\right)$

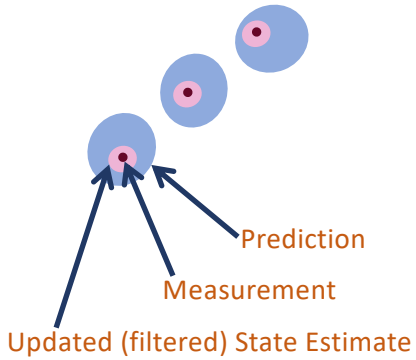
Shrinkage Estimators

- A shrinkage estimator seeks to improve a standard estimator by moving ("shrinking") it toward a value suggested by other available information
  - Bayesian estimators "shrink" toward the prior
  - Other common shrinkage estimators can be interpreted as approximately Bayesian
- The James-Stein estimator is a popular shrinkage estimator
  - It can be shown that the sample mean is inadmissible as an estimator of the mean of a multivariate normal distribution of dimension greater than 2
    - This means there exists an estimator that has uniformly smaller mean squared error
- In general:
  - Maximum likelihood and other standard estimators can perform very poorly on high-dimensional problems, especially with small to moderate sample sizes
  - Well-constructed shrinkage estimators can perform much better
  - Bayesian methods can be used to construct shrinkage estimators

Application of Bayesian Normal Model: Kalman Filter

- Widely applied model for time-varying continuous process measured at regular time intervals
  - "Filter" noise to find best estimate of current state given measurements
  - Predict state at time of next measurement
- Unobservable system state s_t at time t is a real vector
- Measurement z_t at time t depends on system state
- Control input u_t at time t affects movement
- At each time step a recursive algorithm applies Bayesian inference with normal model to:
  - Estimate current state s_t given observations $z_1, \ldots, z_t$
  - Predict next state





Uses of Kalman Filter

- Kalman filter is applied to a wide range of problems where we need to track moving objects
  - Tracking airplanes, missiles, ships, vehicles …
  - Fitting and predicting economic time series
  - Robot navigation
  - Tracking hands, faces, heads in video imagery
- The Kalman filter can be formulated as a model of Bayesian updating and sequential prediction with the normal model

Details: Simple 1-Dimensional Kalman Filter with no Control

- State: position and velocity $s_t = (x_t, v_t)$
- Initial state (x_1, v_1) is known
- Evolution equations:
  - $v_t \mid v_{t-1} \sim \text{Normal}(v_{t-1}, \tau)$
  - $x_t \mid x_{t-1}, v_{t-1} \sim \text{Normal}(x_{t-1} + v_{t-1}, \sigma)$
  - $z_t \mid x_t \sim \text{Normal}(x_t, \xi)$
- At time $t > 1$, conditional on $z_{1:t-1}$ the current state variables are normally distributed
  - $v_t \mid z_{1:t-1}$ has mean m_t and variance ϕ_t^2
  - $x_t \mid vt, z_{1:t-1}$ has mean $a_t + b_t v_t$ and variance ψ_t^2
  - We can use normal-normal conjugate updating to develop recursive updating equations for m_t, a_t, b_t, ϕ_t and ψ_t

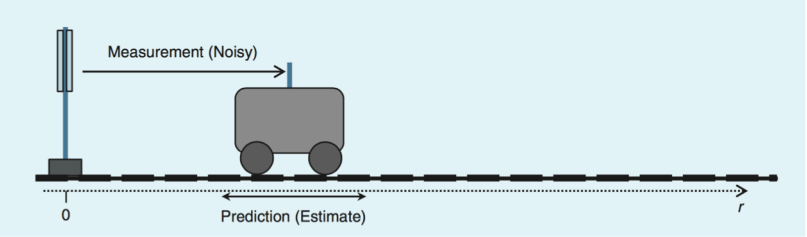
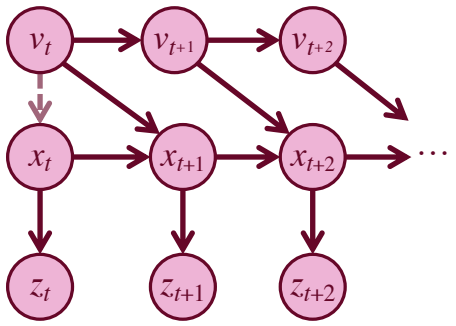
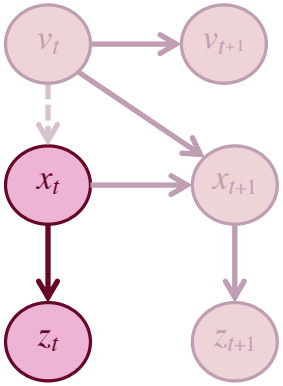


Diagram source: Faragher, Understanding the Basis of the Kalman Filter, *IEEE Signal Processing*, 128, 2012

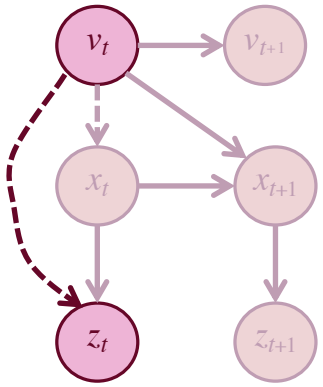


Bayesian Updating: Position

- Evolution equations:
  - $v_t \mid v_{t-1} \sim \text{Normal}(v_{t-1}, \tau)$
  - $x_t \mid x_{t-1}, v_{t-1} \sim \text{Normal}(x_{t-1} + v_{t-1}, \sigma)$
  - $z_t \mid x_t \sim \text{Normal}(x_t, \xi)$
- Current prediction: $x_t \mid vt, z_{1:t-1} \sim \text{Normal}(a_t + b_t v_t, \psi_t)$
- Use Bayesian conjugate updating to find new conditional distribution for position given velocity and $z_{1:t}$
  - Distribution of x_t given $z_{1:t}$, v_t is normal
    - Mean $\dfrac{\frac{a_t + b_t v_t}{\psi_t^2} + \frac{z_t}{\xi_t^2}}{\frac{1}{\psi_t^2} + \frac{1}{\xi_t^2}} = a_t^* + b_t^* v_t$, where $a_t^* = \dfrac{\frac{a_t}{\psi_t^2} + \frac{z_t}{\xi_t^2}}{\frac{1}{\psi_t^2} + \frac{1}{\xi_t^2}}$ and $b_t^* = \dfrac{\frac{b_t}{\psi_t^2}}{\frac{1}{\psi_t^2} + \frac{1}{\xi_t^2}}$
    - Standard deviation $\psi_t^* = \left(\dfrac{1}{\psi_t^2} + \dfrac{1}{\xi_t^2} \right)^{-1/2}$



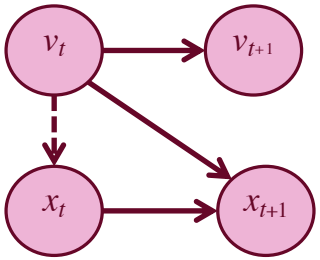
Bayesian Updating: Velocity



- Evolution equations:
  - $v_t \mid v_{t-1} \sim \text{Normal}(v_{t-1}, \tau)$
  - $x_t \mid x_{t-1}, v_{t-1} \sim \text{Normal}(x_{t-1} + v_{t-1}, \sigma)$
  - $z_t \mid x_t \sim \text{Normal}(x_t, \xi)$
- Use Bayesian conjugate updating to find new distribution for velocity
  - Integrate out x_t to get distribution of z_t given v_t and $z_{1:t-1}$
    - $z_t \mid v_t, z_{1:t-1} \sim \text{Normal}(a_t + b_t v_t, \sqrt{\xi^2 + \psi_t^2})$
  - Define transformed measurement $y_t = (z_t - a_t)/b_t$
  - Conditional distribution of y_t given v_t and previous measurements $z_{1:t-1}$ is
    - $y_t \mid v_t, z_{1:t-1} \sim \text{Normal}(v_t, \sqrt{(\xi^2 + \psi_t^2)/ b_t^2})$
  - Distribution of v_t given $z_{1:t-1}$, y_t is normal
    - Mean $m_t^* = \dfrac{\dfrac{m_t}{\varphi_t^2} + \dfrac{y_t b_t^2}{\xi_t^2 + \psi_t^2}}{\dfrac{1}{\varphi_t^2} + \dfrac{b_t^2}{\xi_t^2 + \psi_t^2}}$, Standard deviation $\varphi_t^* = \left(\dfrac{1}{\varphi_t^2} + \dfrac{b_t^2}{\xi_t^2 + \psi_t^2} \right)^{-1/2}$
  - This is also the distribution of v_t given $z_{1:t}$

Predicting the Next Step

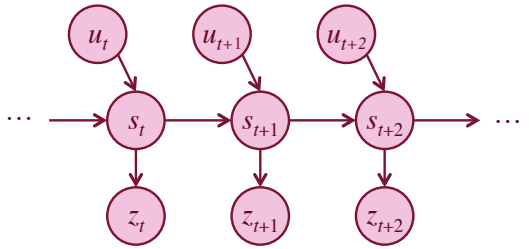
- Distribution of $s_t = (x_t, v_t)$ given $z_{1:t}$ is normal
  - $v_t \mid z_{1:t} \sim \text{Normal}(m_t^*, \varphi_t^*)$
  - $x_t \mid v_t, z_{1:t} \sim \text{Normal}(a_t^* + b_t^* v_t, \psi_t^*)$
- Distribution of $s_{t+1} = (x_{t+1}, v_{t+1})$ given (x_t, v_t) is independent of $z_{1:t}$ and normal
  - $v_{t+1} \mid v_t \sim \text{Normal}(v_t, \tau)$
  - $x_{t+1} \mid x_t, v_t \sim \text{Normal}(x_t + v_t, \sigma)$

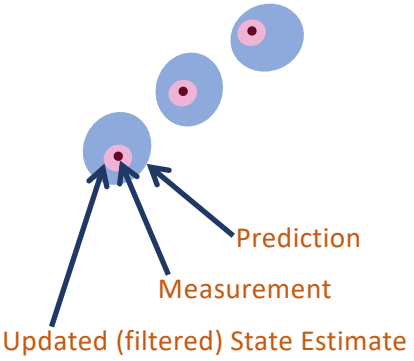


- Marginalizing out v_t and x_t gives predictive distribution for (x_{t+1}, v_{t+1}) given $z_{1:t}$
  - $v_{t+1} \mid z_{1:t} \sim \text{Normal}(m_{t+1}, \varphi_{t+1})$, where $m_{t+1} = m_t^*$ and $\varphi_{t+1}^2 = (\varphi_t^*)^2 + \tau^2$
  - $x_{t+1} \mid v_t, z_{1:t} \sim \text{Normal}(a_{t+1} + b_{t+1} v_{t+1}, \psi_{t+1})$, where $a_{t+1} = a_t^*$, $b_{t+1} = b_t^*$, and $\psi_{t+1}^2 = (\psi_t^*)^2 + \sigma^2$

Summary: Kalman Filter

- The Kalman filter was invented by Rudolf Kalman in 1960-61
- It is widely applied to model time-varying real-valued process measured at regular time intervals
  - "Filter" noise to find best estimate of current state given measurements
  - Predict state at time of next measurement
- Updating equations use Bayesian conjugate updating for normal distribution
- We examined a simple 1-dimensional problem with no control input
- The algorithm operates recursively as follows
  - Filtering: From prediction of current state (prior given measurements prior to current time) and measurement (likelihood) use conjugate Bayesian updating to find posterior distribution given measurements up to and including current time
  - Prediction: Use marginalization to find predictive distribution of next state given measurements up to and including current state





Summary: Bayesian Inference about Parameters of Normal Distribution

- We studied two conjugate families for inferences from normally distributed data:
  - Normal / Normal conjugate pair for inference about unknown mean of normal observations with known precision
  - The Normal / normal-gamma conjugate pair for inference about unknown mean and precision of normal observations
- It is a reasonable approximation to use the known precision model and set the precision equal to the inverse sample variance if:
  - We have more than about 30 observations
  - We do not have strong prior information about the precision
  - We are interested in inference about the mean or the sample mean of future observations
  - We are not directly interested in the posterior distribution of the precision

Normal Model for Non-Normal Data

- The normal model is often applied in situations where we know the observations are not normally distributed
  - Availability of software
  - Familiarity with methods
- This practice is often justified by the Central Limit Theorem, which states that (under fairly general conditions) the sample mean is approximately normal if the sample size is large
- When we are concerned only with inferences about the mean of the distribution, this practice may be reasonable
- Methods based on normal distribution give poor results for inference about quantities other than the mean

Summary: Normal Conjugate Pairs

Data $\underline{X}=(X_1,\ldots,X_n)$	**Prior**	**Sufficient Statistic**	**Posterior**	**Marginal likelihood for sufficient statistic**
$\underline{X}\|\Theta,\sigma \sim$ Normal(Θ,σ)	$\Theta\|\mu,\tau \sim$ Normal(μ,τ^2)	$\bar{X}=\frac{1}{n}\sum_i X_i$	$\Theta\|\underline{X},\sigma \sim$ Normal(μ^*,τ^*) $\mu^*=\frac{\frac{\mu}{\tau^2}+\frac{n\bar{X}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}$, $\tau^*=\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-\frac{1}{2}}$	$f(\bar{x}\mid\mu,\tau,\sigma)=$ $\frac{1}{\sqrt{2\pi(\tau^2+\sigma^2/n)}}\exp\left\{-\frac{1}{2}\frac{(y-\mu)^2}{\tau^2+\sigma^2/n}\right\}$
$\underline{X}\|\Theta,P \sim$ Normal$(\Theta,P^{-1/2})$	$\Theta,T\|\mu,\tau \sim$ Normal-Gamma(μ,k,α,β)	$\bar{X}=\frac{1}{n}\sum_i X_i$, $Y=\sum_i (X_i-\bar{X})^2$	$\Theta,P\|\underline{X},\sigma \sim$ Normal-gamma$(\mu^*,k^*,\alpha^*,\beta^*)$ $\mu^*=\frac{k\mu+n\bar{X}}{k+n}$, $k^*=k+n$, $\alpha^*=\alpha+n/2$; $\beta^*=\left(\beta^{-1}+\frac{1}{2}\sum_i (x_i-\bar{x})^2+\frac{kn(\bar{x}-\mu)^2}{2(k+n)}\right)^{-1}$	$\bar{X} \sim$ nonstd-t$\left(\mu,\frac{1}{\sqrt{\left(\frac{kn}{k+n}\right)\alpha\beta}},2\alpha\right)$ $Y \mid \bar{X}$ has a Gamma-Gamma distribution

Summary and Synthesis

- The normal model is the most studied and the most applied model in statistics
- We studied two conjugate families for inferences from normally distributed data
  - Unknown mean, known precision
  - Unknown mean, unknown precision
- We examined the concept of mean squared error to measure accuracy of a point estimator
  - MSE = bias^2 + variance
  - The Bayesian estimator is biased, but if we can make a reasonable guess at the population mean, it has lower MSE than the sample mean (especially for small sample sizes)
- Methods based on the normal distribution can be useful for inferences from large samples about the mean of non-normal populations, but can be very misleading for inferences about other quantities
- We briefly examined the Kalman filter, a widely applied model for tracking moving objects and predicting their time evolution