



BAYESIAN ESTIMATION

Until now discussion about estimation has assumed a frequentist approach, namely:

- The parameter of the population distribution is unknown but fixed (not random);
- The inference procedures are based not only on the observed sample but also on the population of samples that could have been observed.
- The Bayesian approach assumes that our lack of knowledge about the parameters value should be translated using probability distributions (consequently **unknown parameters are treated as random variables**) and that **only the observed data** (and not the population of samples) **is relevant** to make statistical inference.

Bayesian Inference

$$\left. \begin{array}{l} \text{Model } f_{X|\Theta}(x|\theta) \\ \text{Sample } (x_1, x_2, \dots, x_n) \end{array} \right\} \rightarrow \left. \begin{array}{l} \text{Model distribution } f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) \\ \text{Prior distribution } \pi(\theta) \end{array} \right\} \begin{array}{l} \text{Bayes} \\ \rightarrow \\ \text{Theorem} \end{array} \text{Posterior distribution } \pi_{\Theta|\mathbf{x}}(\theta|\mathbf{x})$$



Prior distribution

- **Definition 15.1** – The **prior distribution** is a probability distribution over the space of possible parameter values. It is denoted $\pi(\theta)$ and represents our opinion concerning the relative chances that various values of θ are the true value of the parameter.
- **Comments:**
 - The existence of a **prior** for θ (scalar or vector) is the core of **Bayesian inference**. From a theoretical point of view it raises important questions about the concept of probability.
 - From a practical point of view, the determination of the prior is a **major problem** of Bayesian methods. In many situations we have some insights about possible parameter values but the main difficulty is translating this knowledge into a probability distribution.
 - Due to the difficulty of finding a prior, we often use an **improper prior** distribution (vague prior) or we take advantage of **conjugate priors**.



- **Definition 15.2** – An **improper prior distribution** is one for which the probabilities (or probability density function) are nonnegative but their sum (integral) is infinite.
- **Comments:**
 - The improper prior is one possible solution when we have minimal knowledge about the parameter behavior.
 - Universal agreement on the best way to construct a vague (or non-informative) prior does not exist.
 - However the use of the improper prior $\pi(\theta) = 1/\theta, \theta > 0$ as a **vague prior** for a scale parameter is quite consensual.
- **Definition 15.17** – A prior distribution is said to be a conjugate prior distribution for a given model if the resulting posterior distribution is from the same family as the prior (but perhaps with different parameters).



Model distribution

- **Theoretical model for the population:** for instance Bernoulli, normal,
- **Instead of considering a random sample** $\mathbf{X} = (X_1, X_2, \dots, X_n)$ – usually the sampling process generates i.i.d. observations – we look at the observed sample $\mathbf{x} = (x_1, x_2, \dots, x_n)$
- **Definition 13.9** – The **model distribution** is the probability distribution for the data as collected given a particular value for the parameter. Note that this matches definition 13.4 for the likelihood function. However, consistent with Bayesian notion, the model pdf is denoted $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$, where vector notation for \mathbf{x} is used to remind us that all the data appear here.
- **Comments:**
 - If the observations are i.i.d., then $L(\theta|\mathbf{x}) = f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = \prod_{i=1}^n f_{\mathbf{X}|\theta}(x_i|\theta)$.
 - Only the likelihood of the **observed sample** is relevant



Bayes Theorem: How to obtain the posterior distribution?

- **Definition 15.6** – The **posterior distribution** is the conditional probability distribution of the parameters, given the observed data. It is denoted $\pi_{\Theta|\mathbf{X}}(\theta | \mathbf{x})$.

- **Theorem 15.8** – (Part a) The posterior distribution can be computed as

$$\pi_{\Theta|\mathbf{X}}(\theta | \mathbf{x}) = \frac{f_{\mathbf{X}|\Theta}(\mathbf{x} | \theta) \times \pi(\theta)}{\int f_{\mathbf{X}|\Theta}(\mathbf{x} | \theta) \times \pi(\theta) d\theta}$$

- **Comment:**

- This is the central purpose of Bayesian analysis: The posterior distribution represents our beliefs about theta once the sample has been observed (and for a given prior).
- In most situations we determine the posterior up to a normalizing constant. This constant can be determined using the condition $\int \pi_{\Theta|\mathbf{X}}(\theta | \mathbf{x}) d\theta = 1$ but it is obtained more easily when the posterior belongs to a known family of distributions. In such cases we identify the core of the family and then we get the constant (using for instance Appendix A or B of the book).

- Remember Bayes's formula: Partition $\{A_1, A_2, \dots\}$, event B , then $P(A_i | B) = \frac{P(B | A_i) \times P(A_i)}{\sum_i P(B | A_i) \times P(A_i)}$



The predictive distribution

- **Definition 15.7** – The predictive distribution is the conditional distribution of a new observation y given the data \mathbf{x} . It is denoted $f_{Y|\mathbf{X}}(y|\mathbf{x})$

- **Theorem 15.8** – (Part b) The predictive distribution can be computed as

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \int f_{Y|\Theta}(y|\theta) \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) d\theta$$

Where $f_{Y|\Theta}(y|\theta)$ is the pdf of the new observation, given the parameter value.

Other definitions (less important)

- **Definition 15.4** – The joint distribution has pdf $f_{\mathbf{X},\Theta}(\mathbf{x}|\theta) \times \pi(\theta)$
- **Definition 15.5** – The marginal distribution of \mathbf{x} has pdf $f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X},\Theta}(\mathbf{x}|\theta) \times \pi(\theta) d\theta$



Example 15.1 – The following amounts were paid on a hospital liability policy

125 132 141 107 133 319 126 104 145 223.

The amount of a single payment has the single-parameter Pareto distribution with $\theta = 100$ and α unknown. The prior is a gamma distribution with parameters $\alpha = 2$ and $\theta = 1$. Determine all of the relevant Bayesian quantities.

Prior: $\pi(\alpha) = \alpha e^{-\alpha}$, $\alpha > 0$

This means that $\alpha \sim \gamma(2,1)$, $E(\alpha) = \text{var}(\alpha) = 2$

Likelihood:

$$L(\alpha | \mathbf{x}) = \prod_{i=1}^n f(x_i | \alpha) = \prod_{i=1}^n \frac{\alpha 100^\alpha}{x_i^{\alpha+1}} \quad x_i > 100$$

$$= \alpha^{10} \prod_{i=1}^{10} \frac{100^\alpha}{x_i^\alpha} \prod_{i=1}^{10} \frac{1}{x_i} = \alpha^{10} \times 0.022346^\alpha \times \frac{1}{\prod_{i=1}^{10} x_i} \propto \alpha^{10} \times 0.022346^\alpha$$



Posterior:

$$\begin{aligned}
 \pi_{\mathbf{A}|\mathbf{X}}(\alpha | \mathbf{x}) &\propto L(\alpha | \mathbf{x}) \times \pi(\alpha) \propto \alpha^{10} \times 0.022346^\alpha \times \alpha \times e^{-\alpha} \\
 &= \alpha^{11} \times \exp(\alpha \ln 0.022346) \times e^{-\alpha} = \alpha^{11} e^{-\alpha(1 - \ln 0.022346)} \\
 &= \alpha^{11} e^{-4.80112\alpha} \quad \alpha > 0
 \end{aligned}$$

We get the *core* of a gamma distribution with parameters 12 and $1/4.80112$ and then we know the normalizing constant which is $4.80112^{12} / \Gamma(12) = 3.757995$. As the posterior belongs to the same family of the prior we said that we are using a conjugate prior for this model.

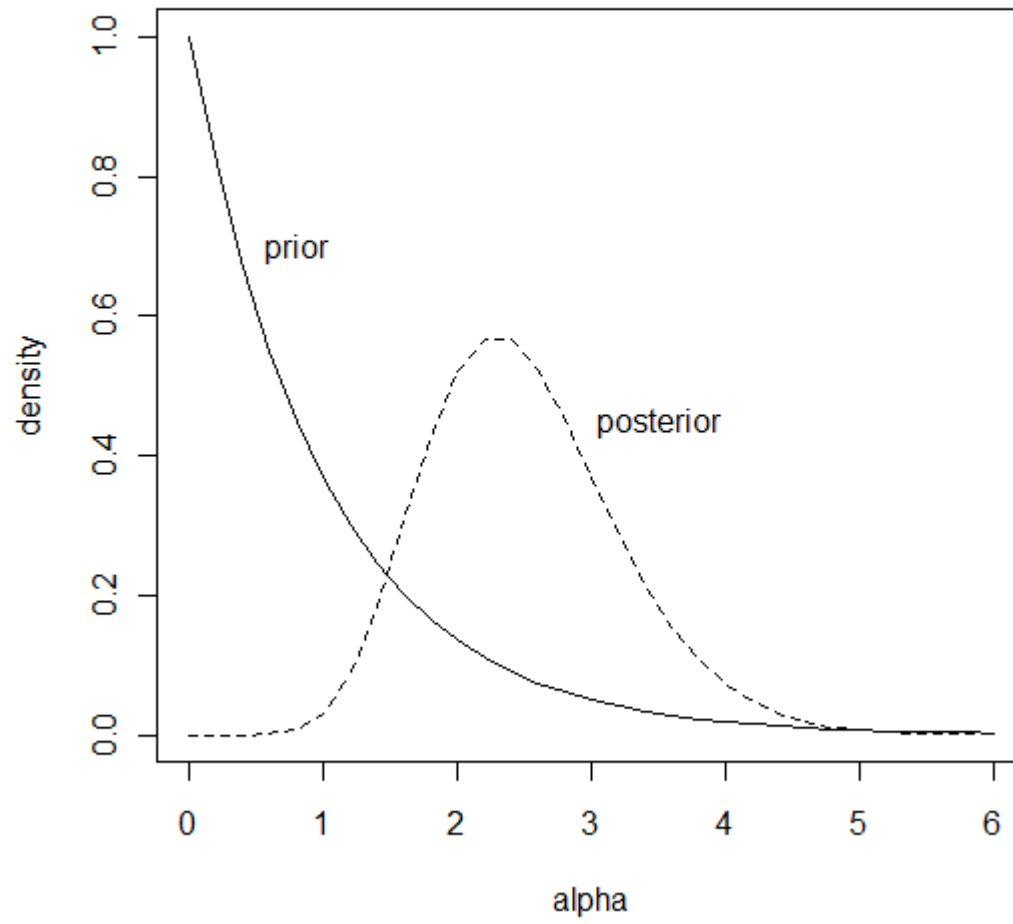
The point here is that the observed samples leads us to change our believes about α from a $\gamma(2,1)$ to a $\gamma(12,0.20828)$ and now $E(\alpha | \mathbf{x}) = 2.49942$ and $\text{var}(\alpha | \mathbf{x}) = 0.52059$.

We can draw both densities on the same graph to visualize the differences:

```

> x=seq(0,6,by=0.2)
> plot(x,dgamma(x,shape=1,scale=1),type="l",ylab="density",xlab="alpha")
> y=dgamma(x,shape=12,scale=0.208285) # posterior
> lines(x,y,type="l",lty=2)
> text(3.5,0.45,"posterior"); text(0.8,0.7,"prior")

```



Predictive:

$$\begin{aligned}
 f_{Y|X}(y|\mathbf{x}) &= \int_0^{\infty} f_{Y|A}(y|\alpha) \pi_{A|X}(\alpha|\mathbf{x}) d\alpha = \int_0^{\infty} \frac{\alpha 100^{\alpha}}{y^{\alpha+1}} 3.757995 \alpha^{11} e^{-\alpha/0.20828} d\alpha \\
 &= \frac{3.757995}{y} \int_0^{\infty} \alpha^{12} e^{-\alpha/0.20828 + \alpha \ln 100 - \alpha \ln y} d\alpha = \frac{3.757995}{y} \int_0^{\infty} \alpha^{12} e^{-\alpha(1/0.20828 - \ln 100 + \ln y)} d\alpha \\
 &= \frac{3.757995}{y} \int_0^{\infty} \alpha^{12} e^{-\alpha(0.195951 + \ln y)} d\alpha \quad y > 100
 \end{aligned}$$

The integrand is the *core* of a gamma density function with parameters 13 and $1/(0.195951 + \ln y)$.

Then we can use the usual normalizing constant to calculate the integral. We get

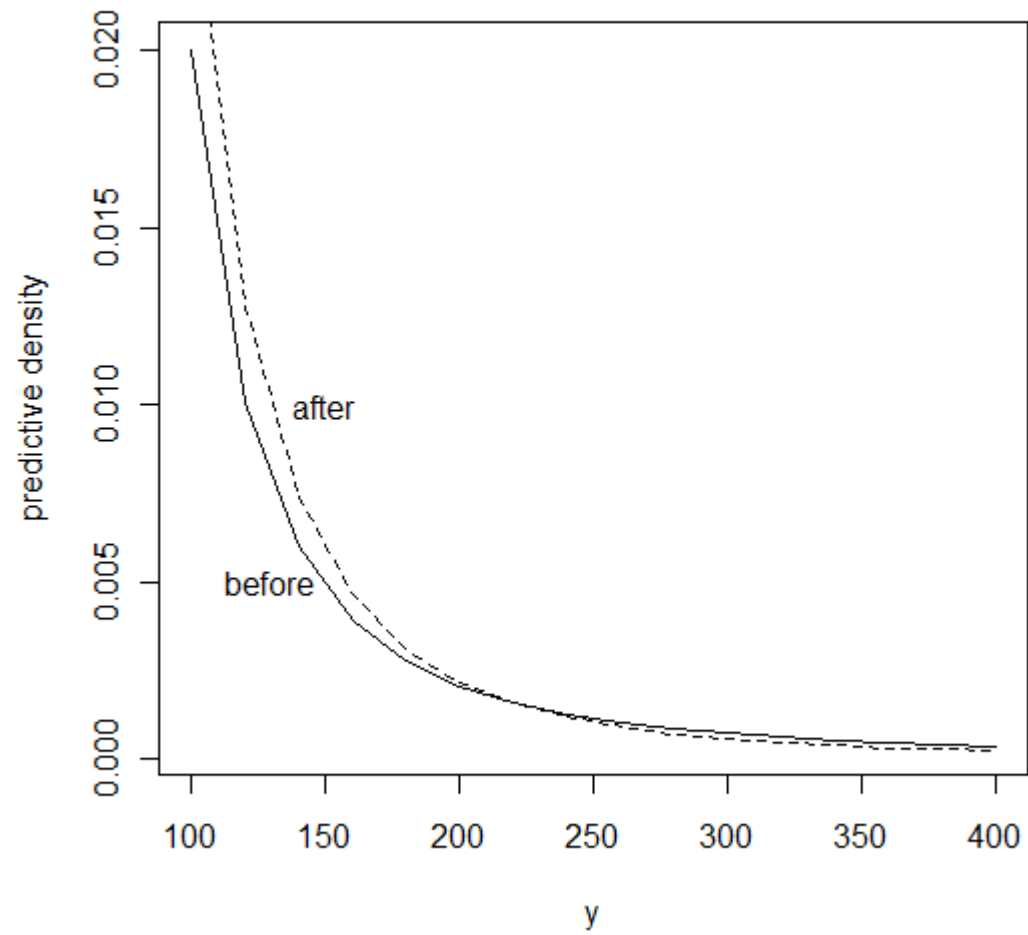
$$f_{Y|X}(y|\mathbf{x}) = \frac{\Gamma(13) \times 3.757995}{(0.195951 + \ln y)^{13} \times y} = \frac{3.757995 \times 12!}{(0.195951 + \ln y)^{13} \times y}, \quad y > 100$$

The density does not look familiar but it can be proved that $\ln Y - \ln 100$ has a Pareto distribution.

```

> y=seq(100, 400, by=20)
> yy1=2*(y^(-1))*((1+log(y/100))^(-3));
> plot(y, yy1, type="l", ylab="predictive density", xlab="y")
> yy2=3.757995*factorial(12)*(y^(-1))*((0.195951+log(y))^(-13));
> lines(y, yy2, type="l", lty=2)
> text(130, 0.005, "before"); text(150, 0.010, "after")

```





Bayesian inference and prediction

From a Bayesian point of view the analysis is complete when we specify the posterior distribution which quantifies our knowledge about θ after the observation of the sample. However, for practical purposes point estimation and/or “confidence interval” are, most of the time, needed. The problem is how to sum up a distribution in one point or using an interval. For point estimation the usual Bayesian solution is to use a loss function.

- **Definition 15.9** – A **loss function** $l_j(\hat{\theta}_j, \theta_j)$ describes the penalty paid by the investigator when $\hat{\theta}_j$ is the estimate and θ_j is the true value of the j th parameter.
- **Comment:** The loss function is random since it depends on θ_j .
- **Definition 15.10** – The **Bayes estimate** for a given loss function is the one that minimizes the expected loss, given the posterior distribution of the parameter in question.
- **Definition 15.11** – For **squared-error** loss, the loss function is (all subscripts are dropped for convenience) $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. For **absolute loss** it is $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$. For **zero-one loss** it is $l(\hat{\theta}, \theta) = 0$ if $\hat{\theta} = \theta$ and 1 otherwise \square .
- **Comment:** Strictly speaking, Definition 13.17 defines the loss functions up to a multiplicative constant.



- **Theorem 15.12** – For squared-loss, the Bayes estimate is the mean of the posterior distribution; for absolute loss it is the median and for zero-one loss it is the mode.

Challenging question: Prove the theorem for the squared loss function (easier) and other functions.

- **Comments:**

- There is no guarantee that the posterior's mean exists (or the mode) or that the median is unique.
- When no otherwise specified, the term Bayes estimate refers to the posterior mean (squared-loss function).

- **Example 15.3** – Determine the three estimates of α (example 15.1 continued)

The posterior is a gamma distribution with parameters 12 and 0.20828. Then $E(\alpha | \mathbf{x}) = 2.49942$, the mode is $11 \times 0.20828 = 2.29132$ and the median has to be determined numerically (2.430342).

- Sometimes the expected value of the predictive distribution is of interest. We can calculate it using the predictive and it can be shown that $E(Y | \mathbf{x}) = \int y f_{Y|\mathbf{X}}(y | \mathbf{x}) dy = \int \pi_{\Theta|\mathbf{X}}(\theta | \mathbf{x}) E(Y | \theta) d\theta$ (see *Loss Models*).



Bayesian HPD credibility set

- **Definition 15.13** – The points $a < b$ define a $100 \times (1 - \alpha)\%$ **credibility interval** for θ_j , provided that $\Pr(a \leq \theta_j \leq b) \geq 1 - \alpha$.
- **Comments:**
 - The term credibility is used to underline the differences between the frequentist (confidence interval) and the Bayesian approaches. This term has no relation with credibility theory.
 - The inequality is due to discrete distribution
 - Definition 15.19 does not produce a unique solution for the credibility interval. Usually we look for the shortest interval.

- **Theorem 15.14** – If the posterior random variable $\theta_j | \mathbf{x}$ is continuous and unimodal, then the $100 \times (1 - \alpha)$ credibility interval with the smallest width, $b - a$, is the unique solution to

$$\int_a^b \pi_{\theta_j | \mathbf{X}}(\theta_j | \mathbf{x}) d\theta = 1 - \alpha \text{ and } \pi_{\theta_j | \mathbf{X}}(b | \mathbf{x}) = \pi_{\theta_j | \mathbf{X}}(a | \mathbf{x})$$

This interval is a special case of a highest posterior density (HPD).

- **Comment:** The posterior cannot have any local maximum except the mode which is unique.



- **Example 15.5** – Determine the shortest 95% credibility interval for the parameter α (example 15.1 continued)

Let us use EXCEL's solver to determine the interval

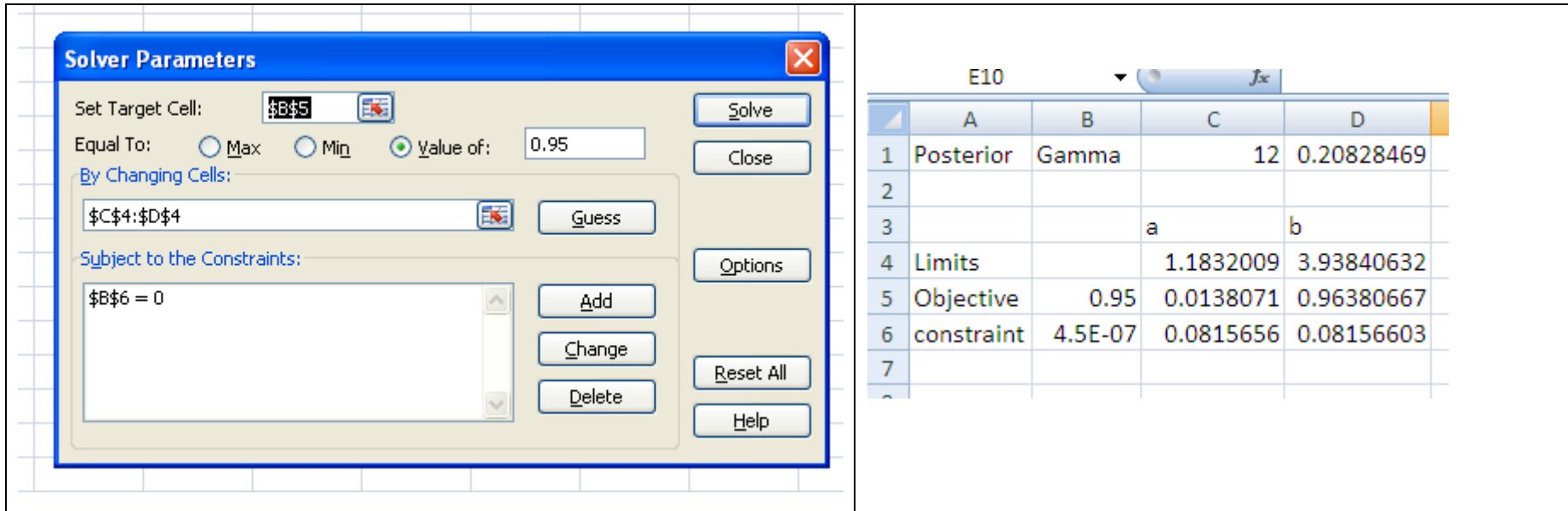
	A	B	C	D
1	Posterior	Gamma	12	0.20828469
2				
3			a	b
4	Limits		0.5	5
5	Objective	0.99748	8.473E-06	0.99748392
6	constraint	0.00672	0.0001664	0.00688849
7				
8				

In cells C4 and D4 we put two initial values for the limits of the interval

In cells C5 and D5 we calculate the value of the distribution function at points a and b respectively. Cell B5 contains the probability of the interval.

In cells C6 and D6 we calculate the value of the posterior density function at points a and b respectively. Cell B6 contains the difference between the density at point b and the density at point a .

Now we want to get the value $1 - \alpha = 0.95$ in cell (5,B) and the value 0 in cell (6,B).



The credibility interval is then (1.1832009; 3.93840632)

To get an approximate solution we place a probability of 0.025 at each end \rightarrow (1.29148; 4.09947).

- **Definition 15.15** – For any posterior distribution the $100 \times (1 - \alpha)\%$ HPD credibility set is the set of parameter values C such that $\Pr(\theta_j \in C) \geq 1 - \alpha$ and $C = \{\theta_j : \pi_{\Theta_j | \mathbf{X}}(\theta_j | \mathbf{x}) \geq c\}$ for some c , where c is the largest value for which the previous inequality holds.
- Sometimes computing posterior probabilities is difficult but computing posterior moments is easier. We can then use the Bayesian central limit theorem.



- **Theorem 15.16 – Bayesian central limit theorem** – If $\pi(\theta)$ and $f_{\mathbf{X}|\Theta}(x|\theta)$ are both twice differentiable in the elements of θ and other commonly satisfied assumptions hold, then the posterior distribution of Θ given $\mathbf{X} = \mathbf{x}$ is asymptotically normal.
- **Comment:** The “commonly satisfied assumptions” are like those presented with Theorem 15.5
- **Example 15.6** – Construct a 95% credibility interval for α using the Bayesian central limit theorem (example 15.1 continued).

The posterior is $\gamma(12, 0.20828)$ and then $E(\alpha|\mathbf{x}) = 2.49942$ and $\text{var}(\alpha|\mathbf{x}) = 0.52059$. The credibility interval is then $2.49942 \pm 1.96 \times \sqrt{0.52059}$, i.e. (1.085238, 3.913594). Note that the method is not appropriate for this example as the sample size is far from large.

$$L(\theta|\mathbf{x}) = \prod_{j=1}^n f_{\mathbf{X}_j|\Theta}(x_j|\theta) = \prod_{j=1}^n \frac{p(x_j) e^{r(\theta)x_j}}{q(\theta)} = \frac{e^{r(\theta)\sum x_j} \prod_{j=1}^n p(x_j)}{q(\theta)^n}$$

$$\text{With } k^* = k + n \text{ and } \mu^* = \frac{\sum x_j + \mu k}{k + n} = \frac{k}{k + n} \times \mu + \frac{n}{k + n} \times \bar{x}.$$

- **Computational issues:** Bayesian analysis proceeds by taking integrals (or sums) and most of the time numerical integration is needed.