L5A: Summarising Posteriors

Overview

A complete summary of a posterior is the posterior distribution. In the case of a single parameter, θ, simply drawing a picture of the pmf (a histogram) or the pdf (a curve) "shows everything" about the parameter. In the case of two parameters, Θ=(θ_1, θ_2), a three dimensional (perspective) plot or contour plot can be drawn. Given that the θs often have some degree of dependency in the joint posterior distribution, e.g., $p(\theta_1, \theta_2|y)$ $\neq p(\theta_1|y)p(\theta_2|y)$, one would want to look at the joint distribution, by looking at contour plots, for example. However, in the case of dozens or even 100s of parameters, sorting through dozens or 100s of histograms or density plots may be burdensome.

Thus while graphical summaries of the posterior distributions are indeed good statistical practice, there is utility in simpler numerical summaries, such as simple point estimates, like posterior mean, measures of spread like posterior standard deviations or interval estimates.

1 Point estimates

Point estimates are single number summaries, such as the mean, or median, or mode. As it turns out, each of these summaries, in the context of *Decision Theory*, can be viewed as optimal estimators for particular types of *Risk*, where risk is the expected value of a *Loss Function*.

1.1 Components of Decision Theory with emphasis on parameter estimation

Decision theory includes parameter estimation, hypothesis testing, and prediction: here we emphasize parameter estimation. We note below that the notation for the loss function is somewhat atypical.

1. *State space, Θ:* There is a true state of nature. In this case this true state is an unknown parameter θ which is contained in a set of possible values called the *state space*, denoted Θ; thus $\theta \in \Theta$. (The state space could be a set of hypotheses.)

2. *Action space, $\mathcal{A}$:* The decision maker will choose a possible action a from a set of possible actions, the action space denoted $\mathcal{A}$; thus $a \in \mathcal{A}$. The action might be viewed as a "decision". In some cases $\mathcal{A}$ is the same as Θ and the action is to choose a value from Θ. (Selecting one hypothesis (H_0) or another (H_1) can be the action.)

3. *Sampling distribution for data, $f(\mathbf{y}|\theta)$:* Conditional on the state of nature, there are data $\mathbf{y}$ which have a distribution that depends upon θ, $f(\mathbf{y}|\theta)$. The data in turn can affect the action taken, i.e., the action taken can be a function of data: a=$\delta(\mathbf{y})$. An example action is to estimate a parameter using the data: $a = \hat{\theta}(\mathbf{y})$.

4. *Loss function, $\mathcal{L}(a|\theta)$:* Given a specific action state of nature, denoted θ, there is an associated loss or "cost" for any action or decision. This loss is denoted $\mathcal{L}(a|\theta)$[1], and in the examples below $\mathcal{L}(\hat{\theta}|\theta)$. An example loss function is $\mathcal{L}(\hat{\theta}|\theta) = (\theta - \hat{\theta})^2$. Note that $\mathcal{L}(\hat{\theta}|\theta)$ is a function of $\hat{\theta}$ as θ is a fixed value.

[1]This is not typical notation, $l(\theta, a)$ or $l(\theta, \hat{\theta})$ is more common but that does not make clear that θ is fixed and a or $\hat{\theta}$ is the variable.

5. *Risk, $R_\theta(a|\mathbf{y})$:* The *Risk* is the expected value of the loss function under some probability distribution for the state of nature, Θ. In a Bayesian setting the probability distribution is the posterior distribution[2] for θ:

$$\text{Risk: } R_\theta(a|\mathbf{y}) = E_\theta\left[\mathcal{L}(a|\theta,\mathbf{y})\right] = \int_{\theta\in\Theta} \mathcal{L}(a|\theta)p(\theta|\mathbf{y})\mathrm{d}\theta$$

Here we focus on the action of estimating a parameter, namely a=$\hat{\theta}$

$$\text{Risk: } R_\theta(\hat{\theta}|\mathbf{y}) = E_\theta\left[\mathcal{L}(\hat{\theta}|\theta,\mathbf{y})\right] = \int_{\theta\in\Theta} \mathcal{L}(\hat{\theta}|\theta)p(\theta|\mathbf{y})\mathrm{d}\theta \tag{1}$$

The subscript θ has been attached to R and E to emphasize that the expectation is being taken with regard to the distribution of θ.

Note that, like loss, risk is a function of $\hat{\theta}$. The distinction from loss is that θ is not part of the risk function as θ has been removed from the function by integration over the support of θ.

6. *Bayes estimator of a parameter:* The *Bayes Estimator* of θ is *defined* as the value of θ that minimizes *Posterior Risk*:

$$\text{Bayes Estimator: } \hat{\theta}_{BE} = \underset{\hat{\theta}\in\Theta}{\operatorname{argmin}}\, R_\theta(\hat{\theta}|\mathbf{y}) \tag{2}$$

For clarity the argmin operation yields the value of $\hat{\theta}$ with the smallest risk. Note that the subscript BE is *not* standard notation.

Comments.

- Determining what loss function to use is not necessarily simple, and can be arrived at via "elicitation", much like priors. (See the 2017 presentation by Soares on Learn in Course Materials, Week 4.)
- Reich & Ghosh (2019) have a tidy 2 page description of Decision Theory, focusing on applications to point estimation, hypothesis testing, and prediction.

Example distinguishing loss and risk. Suppose that θ is restricted to a finite interval, $L \le \theta \le U$, e.g., [2,8], and suppose that the true value of θ is 5. Figure 1 shows plots of $\mathcal{L}(\hat{\theta}|\theta = 5)$ for four different loss functions (to be discussed shortly) against varying values of $\hat{\theta}$. Note that the loss is minimized when $\hat{\theta}$=5, namely θ, for all four loss functions.

In practice one will not know the true value of θ. Uncertainty about the value of θ is reflected by its posterior distribution $p(\theta|\mathbf{y})$. This leads to the risk function calculation. Suppose that the posterior distribution for θ is right-triangular on [2,8] (see Figure 2):

$$p(\theta|\mathbf{y}) = -\frac{1}{9} + \frac{1}{18}\theta, \quad 2 \le \theta \le 8$$

The risk function for the different lost functions and the triangle pdf is calculated using eq'n 1. Plots of the risk versus $\hat{\theta}$ for four loss functions are shown in Figure 3 and the corresponding Bayes Estimators are indicated on the plot. R code to reproduce these plots is in Appendices A and B.

[2] A prior distribution alone could in principle be used, too.

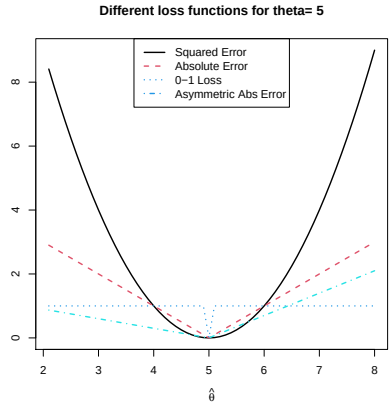


Figure 1: Different loss function values for $\hat{\theta}$ when θ=5.

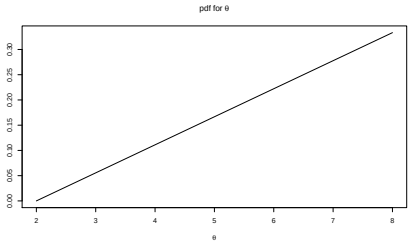


Figure 2: Triangle pdf for θ: $-\frac{1}{9} + \frac{1}{18}\theta$,

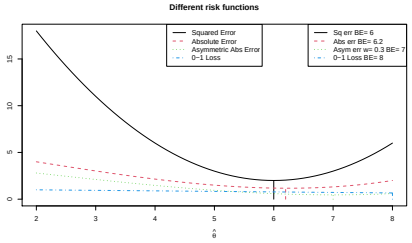


Figure 3: Risk furnction for triangle pdf (on [2,8]) for four loss functions with corresponding θ_{BE} indicated.

1.2 Common loss functions for estimators

1.2.1 Squared error loss

Squared error loss = $\mathcal{L}(\hat{\theta}|\theta)=(\theta-\hat{\theta})^2$. (The notation for $\mathbf{y}$ is omitted.) Here we consider the continuous case. The Bayes Estimator, $\hat{\theta}$, for minimizing expected squared error loss is the *posterior mean*, $E[\theta|\mathbf{y}]$.

Proof:

$$\frac{\mathrm{d}}{d\hat{\theta}}E\left[(\theta-\hat{\theta})^2|\mathbf{y}\right]=\frac{d}{d\hat{\theta}}\int(\theta-\hat{\theta})^2p(\theta|\mathbf{y})d\theta=\int\frac{d}{d\hat{\theta}}(\theta-\hat{\theta})^2p(\theta|\mathbf{y})d\theta=\int-2(\theta-\hat{\theta})p(\theta|\mathbf{y})d\theta$$

Setting the above equal to 0 and solving for $\hat{\theta}$:

$$\begin{aligned}\int(\theta-\hat{\theta})p(\theta|\mathbf{y})d\theta=0 \quad &\Rightarrow \quad \int\theta p(\theta|\mathbf{y})d\theta=\hat{\theta}\int p(\theta|\mathbf{y})d\theta \\ &\Rightarrow \hat{\theta}=E(\theta|\mathbf{y}) \qquad (3)\end{aligned}$$

Thus $E(\theta|\mathbf{y})$ is a critical point. To check that it is a minimum, take the 2nd derivative:

$$\frac{d}{d\hat{\theta}}\int-2(\theta-\hat{\theta})p(\theta|\mathbf{y})d\theta=2\int p(\theta|\mathbf{y})d\theta=2>0$$

Thus $E(\theta|\mathbf{y})$ is a minimum.

Example A. Let θ be the average amount students in Edinburgh will spend on entertainment this month. In this case the sample space is a range of positive numbers; e.g., $\Theta=[£1,\ £1000]$. A random sample of n students will be taken next month and the data are the amounts spent by each sampled student, $\mathbf{y}=y_1,\ldots,y_n$. The "action" or decision is to estimate θ based on the data, $a=\hat{\theta}(\mathbf{y})$. The sampling distribution of the y_i is assumed Normal(θ,50^2). The prior for θ is Normal(30, $\frac{50^2}{5}$)[3]. A random sample of n=100 students was selected and the average spent was £42. The posterior mean[4] is the Bayes Estimator for θ, namely, 41.43.

1.2.2 Absolute error loss

Absolute error loss= $\mathcal{L}(\hat{\theta}|\theta)=|\theta-\hat{\theta}|$. Here again we consider the continuous case. The Bayes Estimator, $\hat{\theta}$, for minimizing expected absolute error loss is the *posterior median*, denoted $\theta_{0.5}$.

Proof:

First re-express the Risk as follows:

$$\begin{aligned}E\left[|\theta-\hat{\theta}||\mathbf{y}\right]&=\int_{-\infty}^{\infty}|\theta-\hat{\theta}|p(\theta|\mathbf{y})\mathrm{d}\theta=\int_{-\infty}^{\hat{\theta}}(\hat{\theta}-\theta)p(\theta|\mathbf{y})\mathrm{d}\theta+\int_{\hat{\theta}}^{\infty}(\theta-\hat{\theta})p(\theta|\mathbf{y})\mathrm{d}\theta \\ &=\hat{\theta}\int_{-\infty}^{\hat{\theta}}p(\theta|\mathbf{y})\mathrm{d}\theta-\int_{-\infty}^{\hat{\theta}}\theta p(\theta|\mathbf{y})\mathrm{d}\theta+\int_{\hat{\theta}}^{\infty}\theta p(\theta|\mathbf{y})\mathrm{d}\theta-\hat{\theta}\int_{\hat{\theta}}^{\infty}p(\theta|\mathbf{y})\mathrm{d}\theta \\ &=\hat{\theta}F_{\theta|\mathbf{y}}(\hat{\theta})-\int_{-\infty}^{\hat{\theta}}\theta p(\theta|\mathbf{y})\mathrm{d}\theta+\int_{\hat{\theta}}^{\infty}\theta p(\theta|\mathbf{y})\mathrm{d}\theta-\hat{\theta}(1-F_{\theta|\mathbf{y}}(\hat{\theta})) \\ &=\hat{\theta}2F_{\theta|\mathbf{y}}(\hat{\theta})-\hat{\theta}-\int_{-\infty}^{\hat{\theta}}\theta p(\theta|\mathbf{y})\mathrm{d}\theta+\int_{\hat{\theta}}^{\infty}\theta p(\theta|\mathbf{y})\mathrm{d}\theta\end{aligned}$$

[3]This is not entirely realistic as θ must be non-negative.

[4]The posterior distribution is Normal$\left(\frac{5*30+100*42}{5+100},\ \frac{50^2}{5+100}\right)$ or Normal(41.43, 7.29^2).

where $F_{\theta|\mathbf{y}}(\theta))$ is the cumulative posterior distribution function for $p(\theta|\mathbf{y})$.

Now differentiate with respect to $\hat{\theta}$[5]:

$$\frac{d}{d\hat{\theta}} E\left[|\theta - \hat{\theta}||\mathbf{y}\right] = 2F_{\theta|\mathbf{y}}(\hat{\theta}) + 2\hat{\theta}p(\hat{\theta}) - 1 - \hat{\theta}p(\hat{\theta}) - \hat{\theta}p(\hat{\theta}) = 2F_{\theta|\mathbf{y}}(\hat{\theta}) - 1$$

Setting the above equal to 0,

$$2F_{\theta|\mathbf{y}}(\hat{\theta}) = 1 \quad \Rightarrow \quad F_{\theta|\mathbf{y}}(\hat{\theta}) = 1/2 \tag{4}$$

Thus $\hat{\theta}$=50th percentile, $\theta_{0.5}$, is a critical value. To check that it is a minimum, take the 2nd derivative:

$$\frac{d}{d\hat{\theta}}\left[2F_{\theta|\mathbf{y}}(\hat{\theta}) - 1\right] = 2p(\hat{\theta})$$

where $p(\hat{\theta})$ is the probability density function which is greater than or equal to 0. If greater than 0, then a=median is a minimum. If equal to 0 then need to check further (we will not worry about this).

Example A (cont). Because the posterior distribution is normal, the median equals the mean, thus the Bayes estimator of absolute loss is again £41.43.

Exercise. Show that the Bayes estimator for the following loss function:

$$\mathcal{L}(\hat{\theta}|\theta) = \begin{cases} (1-\tau)(\hat{\theta} - \theta) & \text{if } \hat{\theta} > \theta \\ \tau(\theta - \hat{\theta}) & \text{if } \hat{\theta} < \theta \end{cases}$$

where $0 < \tau < 1$, is θ_τ, the τ^{th} quantile.

1.2.3 0-1 Loss

0-1 loss, $\mathcal{L}(\hat{\theta}|\theta)$=$I(\theta \neq \hat{\theta})$ is defined as follows:

$$\mathcal{L}(\hat{\theta}|\theta) = \begin{cases} 1, & \text{if } \hat{\theta} \neq \theta. \\ 0, & \text{if } \hat{\theta} = \theta. \end{cases}$$

or $\mathcal{L}(\hat{\theta}|\theta) = I(\theta \neq \hat{\theta})$. We only consider the proof for the discrete case but the results are the same for the continuous case[6]. We want to minimize

$$E\left[\mathcal{L}(\hat{\theta}|\theta)\right] = \sum_{\theta \in \Theta} I(\theta \neq \hat{\theta}) p(\theta|\mathbf{y}) \tag{5}$$

Before giving the general result, suppose there are 3 values of θ: θ_1, θ_2, and θ_3. Thus $\hat{\theta}$ can be chosen to be one of these 3 values. The expected loss (risk) for each choice is then:

$$E\left[\mathcal{L}(\hat{\theta} = \theta_1|\theta))\right] = 0 * p(\theta_1|\mathbf{y}) + 1 * p(\theta_2|\mathbf{y}) + 1 * p(\theta_3|\mathbf{y}) = p(\theta_2|\mathbf{y}) + p(\theta_3|\mathbf{y}) = 1 - p(\theta_1|\mathbf{y})$$
$$E\left[\mathcal{L}(\hat{\theta} = \theta_2|\theta))\right] = 1 * p(\theta_1|\mathbf{y}) + 0 * p(\theta_2|\mathbf{y}) + 1 * p(\theta_3|\mathbf{y}) = p(\theta_1|\mathbf{y}) + p(\theta_3|\mathbf{y}) = 1 - p(\theta_2|\mathbf{y})$$
$$E\left[\mathcal{L}(\hat{\theta} = \theta_3|\theta))\right] = 1 * p(\theta_1|\mathbf{y}) + 1 * p(\theta_2|\mathbf{y}) + 0 * p(\theta_3|\mathbf{y}) = p(\theta_1|\mathbf{y}) + p(\theta_2|\mathbf{y}) = 1 - p(\theta_3|\mathbf{y})$$

[5]Recall the Fundamental theorem of calculus: $\frac{\mathrm{d}}{dx}\int_c^x f(t)dt = f(x)$, and similarly $\frac{\mathrm{d}}{dx}\int_x^c f(t)dt = -f(x)$.

[6]The proof for the continuous case involves the Dirac Delta function; or one can define an even more complicated loss function which has 0-1 loss as a limiting case.

The value of $\hat{\theta}$ minimizing expected loss is the θ_i that has the largest probability, the most likely value or the *posterior mode*[7].

The more general solution given q possible values for θ. Let Θ denote a finite set of possible parameters. The risk is

$$E\left[\mathcal{L}(\hat{\theta}|\theta)\right] = \sum_{\theta\in\Theta} I(\theta \neq \hat{\theta})p(\theta|\mathbf{y}) = 1 - p(\hat{\theta}|\mathbf{y}) \tag{6}$$

Thus the risk is minimized by $\hat{\theta}$ equal to θ with the largest probability, namely the mode.

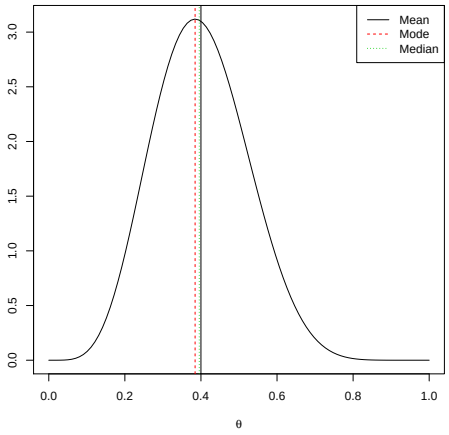
1.3 More Examples

Example B. Binomial θ with a Beta prior. Given a Beta(α, β) prior for a Binomial(n, θ) sampling distribution, the posterior distribution is Beta($\alpha + y$, $\beta + n - y$). Suppose $\alpha = 2$, $\beta = 3$, n=10, and $y = 4$. Then the posterior is Beta(6,9). The Bayes estimators for the loss functions given above:

$$\begin{aligned}
\text{Quadratic loss (posterior mean): } \hat{\theta} &= \frac{\alpha + y}{\alpha + \beta + n} = \frac{2+4}{2+3+10} = 0.4 \\
\text{Absolute error loss (posterior median): } \hat{\theta} &= \texttt{qbeta}(0.5, \texttt{shape1} = \alpha + \texttt{y}, \texttt{shape2} = \beta + \texttt{n} - \texttt{y}) \\
&= \texttt{qbeta}(0.5, 6, 9) = 0.395 \\
\text{0-1 loss (posterior mode): } \hat{\theta} &= \frac{\alpha + y - 1}{\alpha + \beta + n - 2} = \frac{2+4-1}{2+3+10-2} = 0.385
\end{aligned}$$

The three values are shown in Figure 4.

Figure 4: Posterior distribution for Binomial θ, Beta(6,9), with posterior mean, median, and mode.



[7]Note the Bayes estimate would not be unique if there were two or more modes that were the largest and of equal value.

Example C. Poisson θ with a Gamma prior. As shown previously, given a Gamma(α, β) prior for a Poisson(θ) sampling distribution, the posterior distribution (given n=1) is Gamma($\alpha + y$, $\beta + 1$). Suppose $\alpha = 5$, $\beta = 3$, and $y = 3$. The posterior for θ is Gamma(8,4). The Bayes estimators for the loss functions given above:

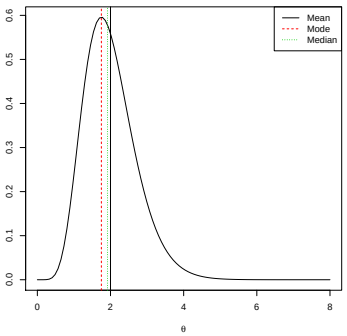
$$\text{Quadratic loss: } \hat{\theta} = \frac{\alpha + y}{\beta + 1} = \frac{5+3}{3+1} = 2$$

$$\text{Absolute error loss: } \hat{\theta} = \text{qgamma}(0.5, \text{shape1}=\alpha + y, \text{ shape2}=\beta + 1) = \texttt{qgamma}(0.5, 8, 4) = 1.917$$

$$\text{0-1 loss: } \hat{\theta} = \frac{\alpha + y - 1}{\beta + 1} = \frac{5+3-1}{3+1} = 1.75$$

Values are shown in Figure 5.

Figure 5: Posterior distribution for Poisson θ, Gamma(8,4), with posterior mean, median, and mode.



2 Interval estimates

Analogous to frequentist *Confidence Intervals* are Bayesian *Credible Intervals* but the latter are more easily interpreted. To begin we just consider the situation of a single parameter, θ, and then discuss intervals, "regions", for two or more parameters.

For a given probability P where P is often expressed as $1 - \alpha$, where $0 < \alpha < 1$, a P% Bayesian Credible interval is *defined* as an interval $[LB, UB]$ where

$$\int_{LB}^{UB} p(\theta|\mathbf{y})d\theta = P \tag{7}$$

Suppose α equals 0.05, then P=1-0.05=0.95. Then a 95% credible interval are those values LB and UB such that

$$\int_{LB}^{UB} p(\theta|\mathbf{y})d\theta = 0.95$$

Similar to one sided-confidence bounds, one can define lower credible bounds, $[LB, \infty]$, and upper credible bounds, $[-\infty, UB]$.

2.1 Example

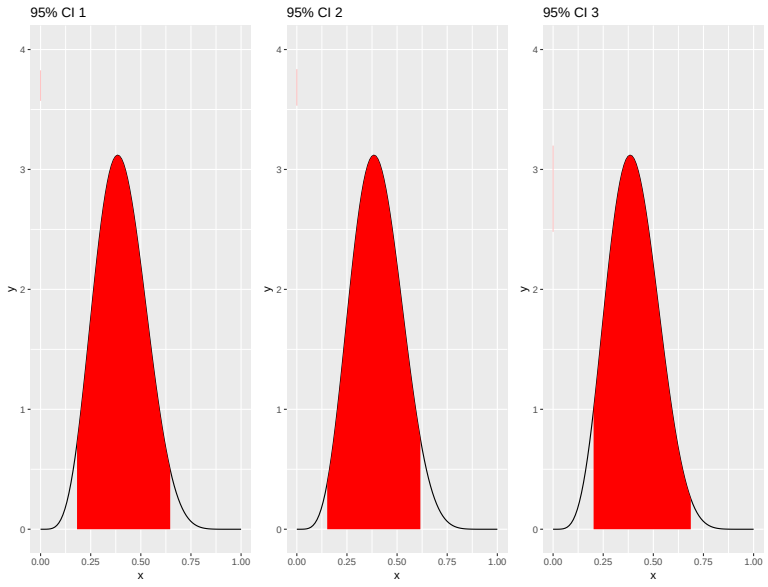
Binomial θ with Beta prior. Suppose the prior for θ is Beta(2,3) and for n=10 Bernoulli trials the observed number of successes is y=4. Then the posterior for θ is Beta(6,9). A 95% credible interval will be any combination of LB and UB such that:

$$\int_{LB}^{UB} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \int_{LB}^{UB} \frac{\Gamma(6+9)}{\Gamma(6)\Gamma(9)}\theta^{6-1}(1-\theta)^{9-1}d\theta = 0.95$$

Note that these bounds are not unique. For example, the following three intervals are all 95% credible intervals (See Figure 6).

$$[0.177, 0.649] \quad [0.146, 0.623] \quad [0.196, 0.692]$$

Figure 6: Three 95% credible intervals for Binomial θ, with posterior Beta(6,9).



Bimodal posteriors. If the posterior distribution is bimodal, then a credible interval composed of two line segments may be more sensible. (Draw a picture.)

2.2 Symmetric credible intervals

A less arbitrary approach to constructing credible intervals is to use symmetric intervals where the probability 1-P, or α, is divided evenly in the tails:

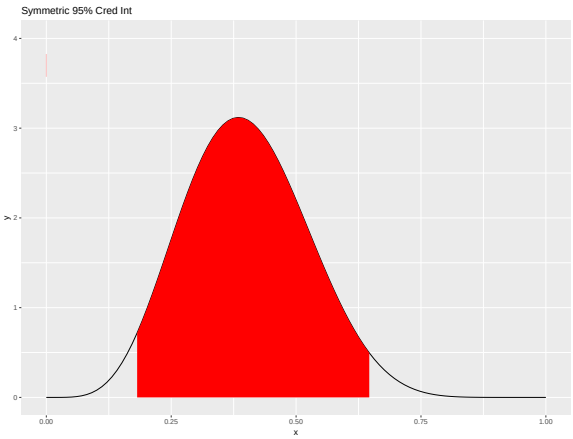
$$\int_{LB_{sym}}^{UB_{sym}} p(\theta|\mathbf{y})d\theta = P \tag{8}$$

where $\Pr(\theta \leq LB_{sym}) = \Pr(\theta \geq UB_{sym}) = \alpha/2$.

- One advantage of the symmetric approach is that for monotonic transformations of θ, $g(\theta)$, the 1-α symmetric interval for $g(\theta)$ is $[g(LB), g(UB)]$ where $[LB, UB]$ are the symmetric bounds for θ.

Referring to the previous Beta example: a 95% symmetric credible interval is [0.177, 0.649], and is shown in Figure 7.

Figure 7: Symmetric 95% credible interval for Binomial θ, with posterior Beta(6,9).



2.3 Highest Posterior Density Intervals

Another non-arbitrary approach is Highest Posterior Density Intervals (HPDIs). These are 1-α intervals where the densities (or probabilities) for values in the interval are higher than for values not in the interval. Another way to say this: "it is the interval with the shortest interval width, U-L, while maintaining appropriate coverage" (Reich and Ghosh, p 26[8]. Formally: The interval $[LB, UB]$ is the 1-α HPDI if

1. $[LB, UB]$ is a 1-α credible interval.
2. For all $\theta' \in [LB, UB]$ and all $\theta'' \notin [LB, UB]$, $p(\theta'|\mathbf{y}) \geq p(\theta''|\mathbf{y})$.

- This will yield the shortest 1-α credible interval.
- If the posterior is unimodal and symmetric, the symmetric and HPDI are the same.
- If the posterior is multimodal, the HPDI may consist of two or more line segments.
- In contrast to symmetric intervals, HPDI's are not invariant to monotonic transformations.

[8]Although bimodal distributions may have two non-overlapping intervals that are shorter in total length than a single interval.

Software for HPDI calculations. The `R` package `HDInterval` has a function `hdi` that can be used to calculate HPDIs for a variety of situations. Below are two examples.

Demonstration with Binomial.

```
#Jeffreys prior
alpha.prior <- 0.5
beta.prior  <- 0.5

#Sampling distribution is Binomial(n=5,theta)
set.seed(1922)
n <- 5
true.theta <- 0.7
y <- rbinom(n=1,size=n,prob=true.theta)
cat("Obs'd data=",y,"\n")
# Obs'd data= 4

#posterior Beta(a+y, b+n-y)
alpha.post <- alpha.prior + y
beta.post  <- alpha.post+n-y
cat("posterior alpha=",alpha.post,"beta=",beta.post,"\n")
#posterior alpha= 4.5 beta= 5.5

x <- seq(0.01,0.99,by=0.01)
plot(x,dbeta(x,alpha.post,beta.post),xlab=expression(theta),
     ylab="",main="Beta Posterior",type="l")

# symmetric and HPDI 99% credible intervals
credible.level <- 0.99
alpha.level    <- 1-credible.level
sym.interval <- qbeta(c(alpha.level/2, 1-alpha.level/2),
                      shape1=alpha.post,shape2=beta.post)
cat("Symmetric interval",sym.interval[1],sym.interval[2],
    "length=",diff(sym.interval),"\n")
# Symmetric interval 0.1147131 0.8191715 length= 0.7044585

# HPDI
hpdi.interval <- hdi(object=qbeta,credMass=credible.level,
                     shape1=alpha.post,shape2=beta.post)
cat("HPD interval",hpdi.interval["lower"],hpdi.interval["upper"],
    "length=",diff(hpdi.interval),"\n")
# HPD interval 0.1092904 0.8129713 length= 0.7036809
```

Note that in this case the symmetric and HPDI are nearly identical.

Demonstration with Poisson.

```
# Prior is Gamma(25.0, 2.5) to yield a
# prior mean of 10 and a CV=0.2
shape.prior <- 25.0
rate.prior  <- 2.5

set.seed(128)
# Sampling distribution
true.theta <- 3
n <- 5
y <- rpois(n=n,lambda=true.theta)
print(y)
# [1] 4 6 4 5 6

# Posterior distribution
```

```
shape.post <- shape.prior+sum(y)
rate.post  <- rate.prior+n

#posterior alpha= 4.5 beta= 5.5
x <- seq(0.1,12,by=0.01)
plot(x,dgamma(x,shape.post,rate.post),type="l")

# symmetric and HPDI 90% credible intervals
credible.level <- 0.90
alpha.level    <- 1-credible.level
sym.interval <- qgamma(c(alpha.level/2, 1-alpha.level/2),
                       shape=shape.post,rate=rate.post)
cat("Symmetric interval",sym.interval[1],sym.interval[2],
    "length=",diff(sym.interval),"\n")
# Symmetric interval 5.195298 8.289474 length= 3.094177

# HPDI
hpdi.interval <- hdi(object=qgamma,credMass=credible.level,
                     shape=shape.post,rate=rate.post)
cat("HPD interval",hpdi.interval["lower"],hpdi.interval["upper"],
    "length=",diff(hpdi.interval),"\n")
# HPD interval 5.113741 8.194088 length= 3.080347
```

In this case the symmetric and HPDI interval endpoints differ slightly but the lengths are nearly identical.

2.4 Credible regions

Given two or more parameters, 1-α, credible regions can be defined as well. There are a variety of ways to construct such regions and the construction can be quite complex. As a simple example suppose that there are two parameters, thus the joint posterior is a surface over the (θ_1, θ_2) plane. A rectangular region can be defined with the four corners being $[(LB_1, LB_2), (LB_1, UB_2), (UB1, LB_2), (UB_1, UB_2)]$ such that

$$\int_{LB_2}^{UB_2} \int_{LB_1}^{UB_1} p(\theta_1, \theta_2|\mathbf{y}) d\theta_1 d\theta_2 = 1 - \alpha$$

The region need not be rectangular, however. An HPD region can be calculated as well, generally by numerical methods, and it will usually not be rectangular. (Draw a picture.)

2.5 Contrast with frequentist intervals

Often, particularly if the data dominates the prior, 1-α frequentist confidence intervals and Bayesian credible intervals can be quite similar. Their interpretation is quite different however.

For example, the sampling model is Normal(θ,1) and the prior for θ is Normal(0,100). One then takes a random sample of n=10 observations and the sample average $\bar{y}$ is 1.54. The posterior distribution is Normal(1.540,0.0999) (we will show why this is in a later lecture).

- Frequentist 95% confidence interval:

$$[\overline{y} - 1.96\sqrt{1/n},\ \overline{y} + 1.96\sqrt{1/n}] = [0.922,\ 2.162]$$

All randomness is in terms of the data as θ is treated as a fixed but unknown quantity.

The interpretation of a 95% confidence interval is that if such intervals are constructed repeatedly based on repeated random samples of the data, the interval will include the unknown parameter θ 95% of the time.

For this given sample the interval either contains θ or it doesn't. Once the sample has been observed there is no more randomness, and one cannot make a probability statement about an observed event *after the fact*. That's like rolling a die and seeing a 5 and then trying to say there's a 90% probability that it is a 3.

- Bayesian 95% credible interval (symmetric and HPDI are the same here), letting $\theta_q|\mathbf{y}$ denote the qth quantile from the Normal(1.540, 0.0999) posterior:

$$[\theta_{0.025}|\mathbf{y},\ \theta_{0.975}|\mathbf{y}] = [0.921,\ 2.160]$$

Thus the credible interval is quite similar to the confidence interval.

However, the interpretation the credible interval is that there is a 95% probability that this particular interval contains the unknown parameter θ.

3 Other summaries

Standard numerical output. Numerical summaries of posterior distributions are routinely produced by software (e.g., JAGS). These include:

Parameter	Min	1st q	Median	Mean	3rd q	Max	StdDev
θ_1	3	12	21	19	31	45	8
θ_2	83	101	122	126	154	198	27

Posterior probabilities for specific events. Given a posterior density one can calculate many quantitative summaries some of which could be quite complex. For example if one has a joint posterior density for two parameters, θ_1 and θ_2, one can calculate:

$$\Pr(\theta_1 + \theta_2 > 7)$$
$$\Pr[(2 \leq \theta_1 \leq 4) \cap (1 \leq \theta_2)]$$
$$\Pr(\exp(\theta_1) < 10)$$

Posterior distributions for functions of parameters. The Schaffer surplus production model is used to model the biomass of a harvested fish stock. Suppose that the sampling model is

$$B_{t+1} \sim \text{Lognormal}\left(\ln\left[B_t + r_{max}B_t\left(1 - \frac{B_t}{K}\right) - C_t\right], \sigma^2\right)$$

where B_t is fish biomass in year t, C_t is the catch and r_{max} and K are unknown parameters. Suppose one has n years of catch and biomass data, and carries out a Bayesian analysis of the unknown parameters, and arrives at a joint posterior distribution for r_{max} and K. The maximum sustainable yield (the maximum harvest that can be taken in "perpetuity") is calculated by

$$MSY = \frac{r_{max}K}{4}$$

Given the joint posterior distribution for r_{max} and K (or a sample from it), one can calculate a joint posterior distribution for MSY.

A R Code for Loss Function Plots

```
# 4 Different Loss Functions:

# Squared Error
L1 <- function(theta,theta.hat) {
  out <- (theta-theta.hat)^2; return(out)
}

# Absolute Error
L2 <- function(theta,theta.hat) {
  out <- abs(theta-theta.hat); return(out)
}

# Asymmetric Error
L3 <- function(theta,theta.hat,w) {
  out <- rep(NA,length(theta))
  ok <- theta >=theta.hat
  out[ok]  <- (1-w)*(theta-theta.hat[ok])
  out[!ok] <- w*(theta.hat[!ok]-theta)
  return(out)
}

# 0-1 Loss
L4 <- function(theta,theta.hat) {
  out <- ifelse(theta==theta.hat,0,1); return(out)
}

theta <- 5; w <- 0.3
lower <- 2; upper <- 8
theta.hat <- seq(lower,upper,by=0.1)
L1.example <- L1(theta,theta.hat)
L2.example <- L2(theta,theta.hat)
L3.example <- L3(theta,theta.hat,w=w)
L4.example <- L4(theta,theta.hat)

my.ylim <- range(c(L1.example,L2.example,L3.example))
my.lwd <- 2
plot(theta.hat,L1.example,type="l",lty=1,
     col=1,ylim=my.ylim,lwd=my.lwd,ylab="",xlab=expression(hat(theta)),
     main=paste("Different loss functions for theta=",theta))
lines(theta.hat,L2.example,lty=2,col=2,lwd=my.lwd)
lines(theta.hat,L3.example,lty=3,col=4,lwd=my.lwd)
lines(theta.hat,L4.example,lty=4,col=5,lwd=my.lwd)
legend("top",legend=c("Squared Error","Absolute Error",
                      "Asymmetric Abs Error","0-1 Loss"),
       lty=1:4,col=1:4,lwd=my.lwd)
```

B R Code for Risk Function Plots

B.1 Triangle pdf code

```
# Calculate triangle pdf parameters and pdf values
 triangle.pdf <- function(theta,lower,upper,verbose=FALSE) {
  b1 <- 2/(upper-lower)^2
  b0 <- -b1*lower
  if(verbose) {
      cat("b0=",b0,"b1=",b1,"\n")
  }
  out <- list(pdf=b0+b1*theta,b0=b0,b1=b1)
  return(out)
}

# Plot the Triangle pdf (and calculate hyper-parameters)
lower <- 2; upper <- 8
theta.hat <- seq(lower,upper,by=0.1)
true.theta <- theta.hat
prob.theta <- triangle.pdf(theta=true.theta,lower=lower,upper=upper,verbose=TRUE)
b0 <- prob.theta$b0
b1 <- prob.theta$b1
plot(true.theta,prob.theta$pdf,type="l",xlab=expression(theta),
     ylab="",main=expression(paste("Triangle pdf")))
```

B.2 Risk function code

```
# Function to calculate the Risk for 4 loss functions
risk.fun <- function(theta.hat,b0,b1,lower,upper,loss.choice=1,
                     w=NULL,verbose=FALSE) {
  n <- length(theta.hat)
  out <- numeric(n)
  for(i in 1:n) {
    theta.hat.val <- theta.hat[i]
    switch(as.character(loss.choice),
           #squared error
           "1"= {
             integrand <-function(theta,theta.hat.val,b0,b1,w) {
               x <-  (b0+b1*theta)*(theta-theta.hat.val)^2
               return(x)
             }
           },
           #absolute error
           "2"= {
             integrand <-function(theta,theta.hat.val,b0,b1,w) {
               x <-  (b0+b1*theta)*abs(theta-theta.hat.val)
               return(x)
             }
           },
           # asymmetric absolute error
```

```
          "3"= {
            integrand <-function(theta,theta.hat.val,b0,b1,w) {
              x <- numeric(length(theta))
              ok <- theta >= theta.hat.val
              x[ok] <- (1-w)*(b0+b1*theta[ok])*(theta[ok]-theta.hat.val)
              x[!ok] <- w*(b0+b1*theta[!ok])*(theta.hat.val-theta[!ok])
              # x <- 2*x  # this will scale to the absolute error loss function
              return(x)
            }
          }
    )
    if(loss.choice !=4) {
      out[i]<- integrate(f=integrand,lower=lower,upper=upper,
                         theta.hat.val=theta.hat.val,b0=b0,b1=b1,w=w)$value
    } else {
      # 0-1 Loss function
      out[i] <- 1-(b0+b1*theta.hat.val)
    }

  }  #end of loop over different values of theta.hat
  return(out)
}
```

B.3 Risk function plots

```
#---Now Calculate and plot Risk functions
w <- 0.3
R1 <- risk.fun(theta.hat,b0=b0,b1=b1,lower=lower,upper=upper,loss.choice=1)
R2 <- risk.fun(theta.hat,b0=b0,b1=b1,lower=lower,upper=upper,loss.choice=2)
R3 <- risk.fun(theta.hat,b0=b0,b1=b1,lower=lower,upper=upper,loss.choice=3,w=w)
R4 <- risk.fun(theta.hat,b0=b0,b1=b1,lower=lower,upper=upper,loss.choice=4)

my.ylim <- range(c(0,R1,R2,R3,R4))
plot(theta.hat,R1,type="l",xlab=expression(hat(theta)),ylab="",
     main="Different risk functions",ylim=my.ylim,lwd=my.lwd)
x <- which(R1==min(R1))
Bayes.L1 <- theta.hat[x]
segments(Bayes.L1,0,Bayes.L1,R1[x],col=1,lwd=my.lwd)
cat("Bayes estimator Sq Err=",Bayes.L1,"\n")
# Bayes estimator Sq Err= 6

lines(theta.hat,R2,lty=2,col=2,lwd=my.lwd)
x <- which(R2==min(R2))
Bayes.L2 <- theta.hat[x]
segments(Bayes.L2,0,Bayes.L2,R2[x],col=2,lty=2,lwd=my.lwd)
cat("Bayes estimator abs error=",Bayes.L2,"\n")
# Bayes estimator abs error= 6.2

lines(theta.hat,R3,lty=3,col=3,lwd=my.lwd)
x <- which(R3==min(R3))
```

```
Bayes.L3 <- theta.hat[x]
segments(Bayes.L3,0,Bayes.L3,R3[x],col=3,lty=3,lwd=my.lwd)
cat("Bayes estimator quantile=",Bayes.L3,"\n")
# Bayes estimator quantile= 7

lines(theta.hat,R4,lty=4,col=4,lwd=my.lwd)
x <- which(R4==min(R4))
Bayes.L4 <- theta.hat[x]
segments(Bayes.L4,0,Bayes.L4,R4[x],col=4,lty=4,lwd=my.lwd)
cat("Bayes estimator 0-1=",Bayes.L4,"\n")
# Bayes estimator 0-1= 8

legend("top",legend=c("Squared Error","Absolute Error","Asymmetric Abs Error",
                      "0-1 Loss"),
       lty=1:4,col=c(1,2,3,4),lwd=my.lwd)

legend("topright",legend=c(paste("Sq err BE=",Bayes.L1),
                           paste("Abs err BE=",Bayes.L2),
                           paste("Asym err w=",w,"BE=",Bayes.L3),
                           paste("0-1 Loss BE=",Bayes.L4)),
       lty=1:4,col=1:4,lwd=my.lwd)
```

October 27, 2020