MONASH University

MONASH BUSINESS SCHOOL

Statistical Thinking (ETC2420/ETC5242)

Associate Professor Catherine Forbes

Week 8: Bayesian inference for numerical data and decision rules

Learning Goals for Week 8

- Review Bayesian statistical thinking
- Apply Bayes theorem with continuous priors
- Consider loss functions and decision rules
- Construct credibility factors

Assigned reading for Week 8:

- Chapter 2 in *Doing Bayesian Data Analysis*, by J. K. Kruschke (same as for Week 7)

Remember where we are...

- Week 7: Transition to Bayesian Thinking
- Bayesian inference is an alternative to Frequentist inference
- Use probability to describe subjective belief,
  - update that belief after observing new information
  - via Bayes theorem:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- We have so far looked at applications where the parameter is defined over a set of possible **discrete** values:
  - Manufacturer who made shirt $M \in \{M_1, M_2, M_3\}$
  - Coin with probability of head $p \in \{p_1, p_2, \ldots, p_K\}$
  - Intended word $W \in \{W_1, W_2, W_3\}$
  - Insurance claims $\theta \in \{\theta_L, \theta_M, \theta_H\}$
- In these cases we normalise *Likelihood* $\times$ *Prior* by making *Posterior* sum to 1

Bayes theorem with a continuous parameter

- Now we consider continuous $\theta \in \Theta \subseteq \mathbb{R}$
- Bayes theorem still holds:

$$f(\theta \mid Data) = \frac{\mathcal{L}_n(\theta) f(\theta)}{\int_\Theta \mathcal{L}_n(\theta) f(\theta)\, d\theta}$$

$$\Rightarrow \quad \text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- How will we compute the normalising constants? (i.e. the integrals)
- We will find our posteriors using:
  - math “tricks” (algebra for conjugate priors)
  - simulation (a Markov chain Monte Carlo technique)
- Aims:
  - Fit simple statistical models using Bayesian method (alternative to the MLE)
  - Obtain posterior probability intervals (alternative to confidence interval from CLT or bootstrap)
  - Construct forecast distribution (alternative to MLE)
- Start with the simple Binomial model under a Uniform(0,1) prior

Bayes theorem for Binomial observation, with a $Uniform(0, 1)$ prior

Now consider prior belief for $p \in (0, 1)$ is **continuous** $Uniform(0, 1)$

- Again assume data $X = x$ (number of heads in n coin tosses)
- Prior density? $f(p) = 1$, for $p \in (0, 1)$
- Calculate **posterior density**:

$$f(p|x) = \frac{P(X = x|p)f(p)}{f(x)} = \frac{\binom{n}{x} p^x (1-p)^{n-x} (1)}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} (1)\, dp}$$

- Notice the denominator **does not depend on** p

$$f(x) = \int_0^1 f(x|p)f(p)dp = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp$$

- And $\binom{n}{x}$ also does not depend on p

Bayes theorem for Binomial observation, with a $\mathit{Uniform}(0,1)$ prior

- So this posterior simplifies to

$$f(p|x) \propto p^x(1-p)^{n-x}\ (\times\ 1)$$

- Notice the symbol $\propto$
  - It means ("is proportional to")
  - $\Rightarrow$ we can drop all factors in $\mathcal{L}(p) \times f(p)$ that **do not depend** on p

$$f(p|x) \propto \mathcal{L}(p)\ f(p)$$

Do you recognize what distribution $f(p|x) \propto p^x(1-p)^{n-x}$ is??

- This is a $\mathit{Beta}(x+1, n-x+1)$ distribution!

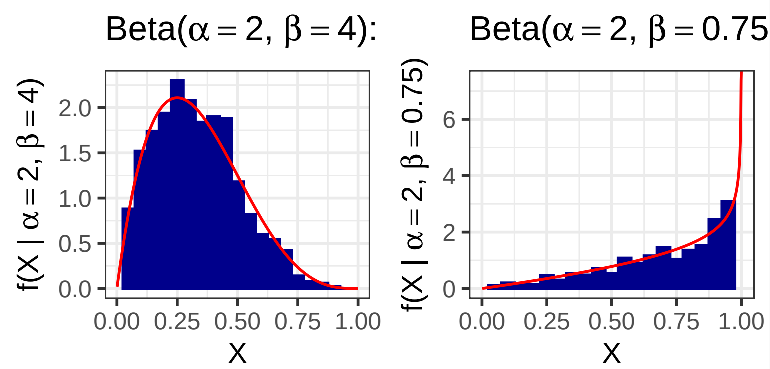
- We didn't actually need to do the integration!
- Just **need to recognize the distribution**!

Beta distribution (from week 6)

If a random variable *X* has a *Beta*(α, β) distribution, the pdf is

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in (0,1) \text{ for } \alpha > 0, \beta > 0$$

- Parameters $\alpha > 0$ and $\beta > 0$
- Generalisation of a continuous uniform on $x \in (0,1)$ (Uniform is *Beta*$(\alpha = 1, \beta = 1)$)



The Beta-Binomial Conjugate Pair

In fact there is a more general result:

- If we assume a $Beta(\alpha,\ \beta)$ **prior** distribution for p in a $Binomial(n,p)$ **model**
- the corresponding **posterior** distribution will be $Beta(\tilde{\alpha} = \alpha + x,\ \tilde{\beta} = \beta + (n - x))$

$$f(p \mid x) \propto \binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

- **NOTE!** Here p is the random variable, and x is fixed!

- In this special situation, the posterior density function and the likelihood function
- Combine to produce a posterior density from the same distributional family as the prior
  - with different hyper-parameter values
- We call such **prior-likelihood** combinations a **conjugate pair**

The Beta-Binomial Conjugate Pair

If you start with the general form of **Bayes' theorem**:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

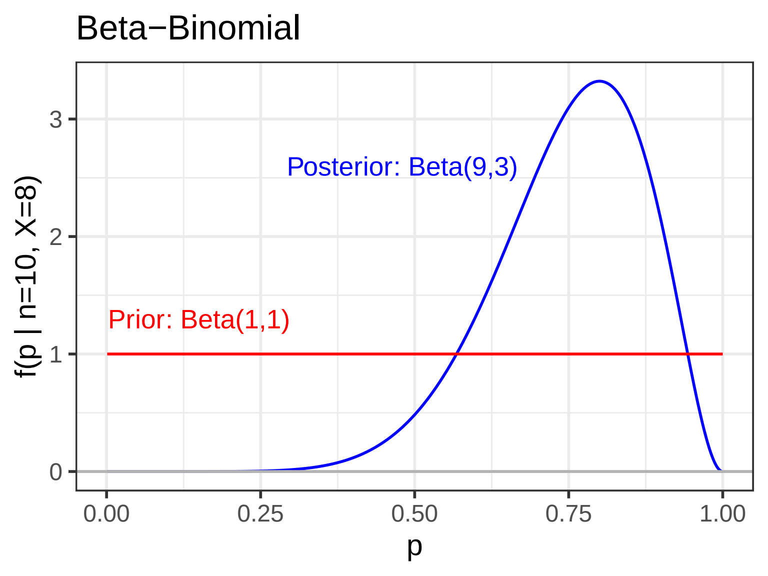
How **to recognise the posterior distribution**?

1. Drop all constants
2. Simplify algebra
3. Look at the remaining functional form
4. Identify hyper-parameter values

$$\begin{aligned} f(p \mid x) &\propto p^x (1-p)^{n-x} \, p^{\alpha-1} (1-p)^{\beta-1} \\ &\propto p^{x+\alpha-1} (1-p)^{n-x+\beta-1} \\ &\propto p^{\tilde{\alpha}-1} (1-p)^{\tilde{\beta}-1} \\ &\propto \text{density of } Beta(\tilde{\alpha} = \alpha + x, \tilde{\beta} = \beta + n - x) \end{aligned}$$

- so if prior is $Beta(1,1)$ and $X \sim Binomial(n,p)$, then the posterior is $Beta(\tilde{\alpha} = 1 + x, \ \tilde{\beta} = 1 + n - x)$
- Under $Uniform(0,1)$ prior, if $x = 8$ *Heads* from $n = 10$ tosses, posterior is $Beta(9,3)$

Beta posterior from Uniform prior and $X = 8$ from $Binomial(n = 10, p)$



Bayes theorem for continuous random variables

- We are interested in the unknown parameter, θ
- Choose **prior density** $f(\theta)$, before we see any data
- Choose a **model** $f(x|\theta)$ that reflects belief about X given θ
- After observing data $X = x$, update belief by calculating **posterior density** $f(\theta|x)$, using Bayes' theorem:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \propto f(x|\theta)f(\theta)$$

- where $f(x) = \int_\theta f(x|\theta)\pi(\theta)d\theta$ (a constant)

- Then **follow steps 1-4 to recognise the posterior distribution** (if possible)

Bayes theorem for continuous parameter and i.i.d. data

- Recall when $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} F_{X|\theta}$,
- the **likelihood function** is given by:

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta), \quad \text{for all } \theta \in \Theta$$

So given a **prior pdf** $f(\theta)$, the posterior pdf satisfies:

$$f(\theta|x_1, x_2, \ldots, x_n) = \frac{L(\theta)f(\theta)}{\int_\Theta L(\theta)f(\theta)\, d\theta} \propto L(\theta)f(\theta)$$

That is, the posterior density satisfies **posterior** $\propto$ **likelihood** $\times$ **prior**

- Note that get same posterior using $X \sim Binomial(n, p)$ or $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(p)$ (**why?**)

Some proposed benefits of the Bayesian approach

The posterior (or prior if no data!) tells us:

- **What are the plausible values for the parameter of interest?**
  - This is precisely what we want to know!
- **Estimate parameters**: Can use any suitable measure of central tendency directly (e.g. mean, median)
  - This type of information is intuitive
- **Quantify uncertainty**: Use **credible (= probability) intervals**
  - Use quantiles of the posterior distribution
  - No (difficult to interpret) confidence intervals!
- Can use Bayes’ theorem to adapt the given “prior” distribution in light of additional evidence
  - Use to **further update posterior** if new information arrives!
  - Just treat first posterior as a new prior
- No need for *p*-values or significance levels as measures of evidence
  - We can directly provide **probabilities** about any hypotheses of interest
  - (Not covered in this unit)

Possible negatives of the Bayesian approach

Parameters **treated as a random variable**, even if they are really a constant

- (Probabilities express uncertainty in the parameter value, so need not truly be “random”)

Probability distributions express **subjective belief**

- not “objective”
- (though there may have been some earlier analysis that has informed this opinion)

Conjugate Priors

Computing posterior distributions can be difficult

- multivariate θ
- high dimensional data X

Special case: **Conjugate Prior**

- A class of special cases where calculation is easy
- Prior and likelihood function share the same **kernel** functional form

Definition: Let $\mathcal{F}$ denote the class of probability density (or mass) functions $f(x \mid \theta)$ indexed by θ. A class $\mathcal{C}$ of prior distributions is a **conjugate family** for $\mathcal{F}$ if the posterior distribution is in the class $\mathcal{C}$ for all $f \in \mathcal{F}$, all priors in $\mathcal{C}$, and all x in the sample space.

- Some (univariate) Prior-Likelihood conjugate pairs
  - Beta-Binomial
  - Beta-Bernoulli
  - Gamma-Poisson
  - Gamma-Exponential
  - Normal-Normal (mean)

Conjugate Prior-Likelihood Pairs

Beta-Binomial

$X =$ number of successes in n Bernoulli trials

$$\begin{array}{lrcl} \textit{Likelihood}: & X \mid \theta & \sim & \textit{Binomial}(n, \theta) \\ \textit{Prior}: & \theta & \sim & \textit{Beta}(\alpha, \beta) \end{array} \quad \Rightarrow \quad \theta \mid X = x \sim \textit{Beta}(\alpha + x, \beta + n - x)$$

Beta-Bernoulli

$$X_i = \begin{cases} 1 & \text{if 'success'} \\ 0 & \text{if 'failure'} \end{cases}$$

$$\begin{array}{rcl} X_1, X_2, \ldots, X_n \mid \theta & \sim & \textit{Bernoulli}(\theta) \\ \theta & \sim & \textit{Beta}(\alpha, \beta) \end{array} \quad \Rightarrow \quad \theta \mid x_{1:n} \sim \textit{Beta}(\alpha + n\overline{x}, \beta + n - n\overline{x})$$

Notes:

1. $x_{1:n} = \{x_1, x_2, \ldots, x_n\}$
2. $n\overline{x} = \sum_{i=1}^{n} x_i$
3. $\Rightarrow$ Same posterior in these two cases (just slightly different notation!)

Beta-Binomial Conjugate Pair

$$
\begin{aligned}
f(\theta|X) &\propto \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\textbf{prior pdf}} \cdot \underbrace{\binom{n}{x}\theta^{x}(1-\theta)^{n-x}}_{\text{likelihood function}} \\
&\propto \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\textbf{prior kernel}} \cdot \underbrace{\theta^{x}(1-\theta)^{n-x}}_{\text{likelihood kernel}} \\
&\propto \underbrace{\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}}_{\textbf{posterior kernel}} \\
&\propto \underbrace{\boxed{\underbrace{\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)}}_{\text{normalising constant}} \cdot \underbrace{\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}}_{\textbf{posterior kernel}}}}_{\textbf{posterior pdf}} \\
&\Rightarrow \underbrace{\boxed{Beta(\alpha+x,\beta+n-x)}}_{\textbf{posterior}}
\end{aligned}
$$

Beta-Bernoulli Conjugate Pair

- Here $X_1, X_2, \ldots X_n \mid \theta \overset{i.i.d}{\sim} Bernoulli(\theta)$

$$
\begin{aligned}
f\left(\theta \mid x_1, x_2, \ldots, x_n\right) \quad & \propto \quad \underbrace{\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}}_{\textbf{prior pdf}} \cdot \underbrace{\prod_{i=1}^{n} \theta^{x_i}\left(1-\theta\right)^{1-x_i}}_{\text{likelihood function}} \\
& \propto \quad \underbrace{\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}}_{\textbf{prior kernel}} \cdot \underbrace{\theta^{\sum_{i=1}^{n} x_i}\left(1-\theta\right)^{n-\sum_{i=1}^{x}}}_{\text{likelihood kernel}} \\
& \propto \quad \underbrace{\theta^{\alpha+n\bar{x}-1}\left(1-\theta\right)^{\beta+n-n\bar{x}-1}}_{\textbf{posterior kernel}} \\
& \propto \quad \underbrace{\boxed{\underbrace{\frac{\Gamma\left(\alpha+\beta+n\right)}{\Gamma(\alpha+n\bar{x})\Gamma(\beta+n-n\bar{x})}}_{\text{normalising constant}} \cdot \underbrace{\theta^{\alpha+n\bar{x}-1}\left(1-\theta\right)^{\beta+n-n\bar{x}-1}}_{\textbf{posterior kernel}}}}_{\textbf{posterior pdf}} \\
& \Rightarrow \quad \underbrace{\boxed{Beta\left(\alpha+n\bar{x}, \beta+n-n\bar{x}\right)}}_{\textbf{posterior}}
\end{aligned}
$$

Other Univariate Conjugate Prior-Likelihood Pairs?

Gamma-Poisson

$$\begin{array}{rcl} \theta & \sim & Gamma(\alpha, \beta) \\ X_1, X_2, \ldots, X_n \mid \theta & \overset{i.i.d.}{\sim} & Poisson(\theta) \end{array} \quad \Rightarrow \quad \theta \mid X_1, \ldots X_n \sim Gamma(\alpha + n\bar{x}, \beta + n)$$

β is a 'rate' parameter.

Gamma-Exponential

$$\begin{array}{rcl} \lambda & \sim & Gamma(\alpha, \beta) \\ X_1, X_2, \ldots, X_n \mid \lambda & \overset{i.i.d.}{\sim} & Exponential(\lambda) \end{array} \quad \Rightarrow \quad \lambda \mid X_1, \ldots X_n \sim Gamma(\alpha + n, \beta + n\bar{x})$$

β is a 'rate' parameter.

Normal-Normal (mean only)

$$\begin{array}{rcl} \mu & \sim & N(\mu_p, \tau^2) \\ X_1, X_2, \ldots, X_n \mid \mu & \overset{i.i.d.}{\sim} & N(\mu, \sigma^2) \end{array} \quad \Rightarrow \quad \mu \mid X_1, \ldots X_n \sim N(\tilde{\mu}_p, \tilde{\sigma}_p^2)$$

Gamma-Poisson Conjugate Pair

$\theta \sim Gamma(\alpha, \beta)$ and $X_1, X_2, \ldots, X_n \mid \theta \overset{i.i.d.}{\sim} Poisson(\theta)$

$$
\begin{aligned}
f(\theta | X_1, X_2, \ldots, X_n) \quad &\propto \quad \underbrace{\frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\theta\beta}}_{\textbf{prior pdf}} \cdot \underbrace{\prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}}_{\text{likelihood function}} \\
&\propto \quad \underbrace{\theta^{\alpha-1} e^{-\theta\beta}}_{\textbf{prior kernel}} \cdot \underbrace{\theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}}_{\text{likelihood kernel}} \\
&\propto \quad \underbrace{\theta^{\alpha+n\bar{x}-1} e^{-(\beta+n)\theta}}_{\textbf{posterior kernel}} \\
&\propto \quad \underbrace{\frac{(\beta+n)^{\alpha+n\bar{x}}}{\Gamma(\alpha+n\bar{x})}}_{\text{normalising constant}} \cdot \underbrace{\theta^{\alpha+n\bar{x}-1} e^{-(\beta+n)x}}_{\textbf{posterior kernel}} \\
&\Rightarrow \quad \underbrace{\boxed{Gamma(\alpha+n\bar{x}, \beta+n)}}_{\textbf{posterior}}
\end{aligned}
$$

Gamma-Exponential Conjugate Pair

$\lambda \sim Gamma(\alpha, \beta)$ and $X_1, X_2, \ldots, X_n \mid \lambda \overset{i.i.d.}{\sim} Exp(\lambda)$

$$
\begin{aligned}
f\left(\lambda | X_1, X_2, \ldots, X_n\right) \quad & \propto \quad \underbrace{\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)} \lambda^{\alpha-1} e^{-\lambda\beta}}_{\textbf{prior pdf}} \cdot \underbrace{\prod_{i=1}^{n} \lambda e^{\lambda x_i}}_{\text{likelihood function}} \\
& \propto \quad \underbrace{\lambda^{\alpha-1} e^{-\lambda\beta}}_{\textbf{prior kernel}} \cdot \underbrace{\lambda^{n} e^{-n\bar{x}\lambda}}_{\text{likelihood kernel}} \\
& \propto \quad \underbrace{\lambda^{\alpha+n-1} e^{-\left(\beta+n\bar{x}\right)\lambda}}_{\textbf{posterior kernel}} \\
& \propto \quad \underbrace{\frac{\left(\beta+n\bar{x}\right)^{\alpha+n}}{\Gamma\left(\alpha+n\right)}}_{\text{normalising constant}} \cdot \underbrace{\lambda^{\alpha+n-1} e^{-\left(\beta+n\bar{x}\right)\lambda}}_{\textbf{posterior kernel}} \\
& \Rightarrow \quad \underbrace{\boxed{Gamma\left(\alpha+n, \beta+n\bar{x}\right)}}_{\textbf{posterior}}
\end{aligned}
$$

Normal-Normal Conjugate Pair

$\mu \sim N(\mu_p, \tau^2)$ and $X_1, X_2, \ldots, X_n \mid \mu \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$

$$
\begin{aligned}
f\left(\mu | X, \sigma^2\right) &\propto \underbrace{\left(2\pi\tau^2\right)^{-1} e^{-\frac{1}{2\tau^2}\left(\mu-\mu_p\right)^2}}_{\textbf{prior pdf}} \cdot \underbrace{\prod_{i=1}^{n}\left(2\pi\sigma^2\right)^{-1} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}}_{\text{likelihood function}} \\
&\propto e^{-\frac{1}{2\tau^2}\left(\mu-\mu_p\right)^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\
&\propto e^{-\frac{1}{2}\left[\frac{1}{\tau^2}\left(\mu-\mu_p\right)^2+\frac{1}{\sigma^2}\left[(n-1)s^2+n\left(\bar{x}-\mu\right)^2\right]\right]} \\
&\propto e^{-\frac{1}{2}\left[\frac{1}{\tau^2}\left(\mu^2-2\mu\mu_p+\mu_p^2\right)+\frac{n}{\sigma^2}\left(\bar{x}^2-2\mu\bar{x}-\mu^2\right)\right]} \\
&\propto e^{-\frac{1}{2}\left[\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)\mu^2-2\mu\left(\frac{\mu_p}{\tau^2}+\frac{n\bar{x}}{\sigma^2}\right)\right]} \\
&\propto e^{-\frac{1}{2}\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)\left[\left(\mu^2-\frac{\left(\frac{\mu_p}{\tau^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)}\right)^2\right]}
\end{aligned}
$$

Normal-Normal (continued)

$$\propto \left(2\pi\tilde{\tau}^2\right)^{-1} e^{-\frac{1}{2\tilde{\tau}^2}(\mu-\tilde{\mu_p})^2}$$

$$\Rightarrow \mu \mid X = \{x_1, x_2, \ldots, x_n\}, \sigma^2 \sim \underbrace{\boxed{N\left(\tilde{\mu}_p, \tilde{\tau}^2\right)}}_{\textbf{posterior}}$$

where

$$\begin{aligned}
\tilde{\mu}_p &= \frac{n\bar{x}\tau^2 + \mu_p\sigma^2}{\sigma^2 + n\tau^2} \\
&= \left(\frac{n\tau^2}{\sigma^2 + n\tau^2}\right)\bar{x} + \left(\frac{\sigma^2}{\sigma^2 + n\tau^2}\right)\mu_p \\
\tilde{\tau}^2 &= \frac{\tau^2\sigma^2}{\sigma^2 + n\tau^2}
\end{aligned}$$

Basic elements of decision theory

Decision theory is concerned with determining the optimal strategies for taking actions.

- θ denotes a **state of nature** (usually unknown)
- Θ is the set of **all possible states of nature**
- Decision a is called an **action**
- $\mathcal{A}$ is the set of **all possible actions**
- Require a **loss function** $L(\theta, a)$ defined over all $(\theta, a) \in \Theta \times \mathcal{A}$.

Use principles of decision theory to determine **how to use** the posterior distribution

- Will depend on the application setting

When estimating a parameter, actions are estimators $a = \widehat{\theta}(X)$ (usually functions of data X)

$\Rightarrow \mathcal{A} \equiv \Theta$.

The **Squared Error Loss** function is given by

$$L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$$

Bayes estimators minimise posterior expected loss

From a Bayesian perspective, θ is treated as random

$\Rightarrow$ we use the posterior probability distribution to characterise belief

- let $f(\theta \mid X)$ denote
  - the posterior pdf, when θ is a continuous random variable, or
  - the posterior probability (mass) function, when θ is a discrete random variable

A **Bayes estimator**, denoted $\widehat{\theta}_{Bayes}$ is the estimator that minimises the posterior expected loss, i.e.

$$\widehat{\theta}_{Bayes} = \arg\min_{\widehat{\theta}\in\Theta} E[L(\theta, \widehat{\theta})] = \arg\min_{\widehat{\theta}\in\Theta} \int_{\Theta} L(\theta, \widehat{\theta}) f(\theta \mid X) d\theta$$

Common loss functions and the corresponding Bayes estimators

Squared error loss: $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$

- Bayes estimator is the **posterior mean** $\widehat{\theta}_{Bayes} = E(\theta \mid X)$

Why?

- Posterior expected loss is

$$\begin{aligned} \varphi(\widehat{\theta}) &= E[(\theta - \widehat{\theta})^2 \mid X] \\ &= \widehat{\theta}^2 - 2\widehat{\theta}E[\theta \mid X] + E[\theta^2 \mid X] \end{aligned}$$

$\Rightarrow$ If we differentiate with respect to $\widehat{\theta}$ and solve for the root of the first derivative...

$$\varphi'(\widehat{\theta}) = 2\widehat{\theta} - 2E[\theta \mid X] + 0$$

$$\Rightarrow \widehat{\theta} = E[\theta \mid X]$$

Credibility Factors

Notice that in the normal-normal problem,

$$\tilde{\mu}_p = \left(\frac{n\tau^2}{\sigma^2 + n\tau^2} \right) \overline{x} + \left(\frac{\sigma^2}{\sigma^2 + n\tau^2} \right) \mu_p$$

It turns out that in many cases (all we consider) we can write the posterior mean as a linear combination of

- a (sensible!) estimator based solely on data (e.g. an MLE), and
- the **prior mean**

This means we can interpret the estimator $\widehat{\theta}_{Bayes}$ as a trade-off between two reasonable alternatives

- a (sensible!) estimator based solely on data (e.g. an MLE), and
- a (**sensible!**) estimator based on **judgement and prior knowledge**

Definition: A so-called **credibility factor** for an estimator that linearly combines a data-based estimator $\widehat{\theta}(X)$ with a non-data-based estimator, $\widehat{\theta}_{prior}$, is the relative weight Z given to the data-based estimator.

Credibility Factor Example 2: Beta-Binomial

Suppose $X \mid \theta \sim Binomial(n, \theta)$

And we take **conjugate prior** $\theta \sim Beta(\alpha, \beta)$, having prior mean $\frac{\alpha}{\alpha+\beta}$

$\Rightarrow$ the **posterior is** $Beta(\tilde{\alpha} = \alpha + x, \tilde{\beta} = \beta + n - x)$

Taking

- the **sample proportion** $\frac{x}{n}$ as the data-based estimator, and
- the **prior mean** $\frac{\alpha}{\alpha+\beta}$ as the estimator based on prior knowledge,

$\Rightarrow$ it can be shown that the posterior mean $\frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}} = \frac{\alpha+x}{\alpha+\beta+n}$ satisfies

$$\frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}} = Z\left(\frac{x}{n}\right) + (1-Z)\left(\frac{\alpha}{\alpha+\beta}\right),$$

when the credibility factor $Z = \left(\frac{n}{\alpha+\beta+n}\right)$

Credibility Factor Example 3: Gamma-Exponential

Suppose $X_1, X_2, \ldots, X_n \mid \lambda \overset{i.i.d}{\sim} Exponential(\lambda)$

And we take **conjugate prior** $\lambda \sim Gamma(\alpha, \beta)$

$\Rightarrow$ **posterior** $\lambda \mid x_1, x_2, \ldots, x_n \sim Gamma(\tilde{\alpha} = \alpha + n, \tilde{\beta} = \beta + n\overline{x})$

Taking

- the MLE $\hat{\lambda}_{MLE} = (\overline{x})^{-1}$ as the data-based estimator, and
- the prior mean $E[\lambda] = \frac{\alpha}{\beta}$ as the estimator based on prior knowledge,

$\Rightarrow$ the posterior mean $\frac{\tilde{\alpha}}{\tilde{\beta}} = \frac{\alpha+n}{\beta+n\overline{x}}$ satisfies

$$\frac{\alpha + n}{\beta + n\overline{x}} = Z\left(\frac{1}{\overline{x}}\right) + (1 - Z)\,\frac{\alpha}{\beta},$$

when the **credibility factor** $Z = \left(\frac{n\overline{x}}{n\overline{x}+\beta}\right)$

Compare with Frequentist?

Frequentist also try to minimise expected squared error loss

But expectation taken with respect to $f(X \mid \theta)$

- Can't find unique solution
- Need to combine with other strategies

e.g. Squared error loss

- Expected loss: (average over X, with θ fixed)

$$
\begin{aligned}
E_X[L(\theta, \widehat{\theta}(X)) \mid \theta] &= E_X[(\theta - \widehat{\theta}(X))^2] \\
&= E_X[(\theta - E_X[\widehat{\theta}(X)] + E_X[\widehat{\theta}(X) \mid \theta] - \widehat{\theta}(X))^2] \\
&= E_X[(\theta - E_X[\widehat{\theta}(X))^2] + E_X[\widehat{\theta}(X) - E_X[\widehat{\theta}(X)])^2] \\
&= \text{Bias}^2 + \text{Variance}
\end{aligned}
$$

$\Rightarrow$ "Bias - Variance Trade-off"

Other loss functions

Absolute loss

$$L(\theta, \widehat{\theta}) = \left|\theta - \widehat{\theta}\right|$$

$\Rightarrow$ Bayes estimator is the posterior median

(Why?)

Asymmetric loss functions also possible