Bayesian Inference, Basics

Professor Wei Zhu

Bayes Theorem

- Bayesian statistics named after Thomas Bayes (1702-1761) -- an English statistician, philosopher and Presbyterian minister.

Bayes Theorem

- Bayes Theorem for probability events A and B

$$P(A|B) = \frac{P(B \bigcap A)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- Or for a set of mutually exclusive and exhaustive events (i.e. $P(\bigcup_i A_i) = \sum_i P(A_i) = 1$), then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

Example – Diagnostic Test

- A new Ebola (EB) test is claimed to have "95% sensitivity and 98% specificity"
- In a population with an EB prevalence of 1/1000, what is the chance that a patient testing positive actually has EB?

Let A be the event patient is truly positive, A' be the event that they are truly negative

Let B be the event that they test positive

Diagnostic Test, ctd.

- We want P(A|B)
- “95% sensitivity” means that P(B|A) = 0.95
- “98% specificity” means that P(B|A’) = 0.02

So from Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')P(A')}$$

$$= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 \times 0.999} = 0.045$$

Thus over 95% of those testing positive will, in fact, not have EB.

Bayesian Inference

In Bayesian inference there is a fundamental distinction between

- Observable quantities *x,* i.e. the data
- Unknown quantities θ

θ can be statistical parameters, missing data, latent variables…

- Parameters are treated as random variables

In the Bayesian framework we make probability statements about model parameters

In the frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data.

Prior distributions

As with all statistical analyses we start by posting a model which specifies f(**x**| θ)

This is the **likelihood** which relates all variables into a '**full probability model**'

However from a Bayesian point of view :

- θ is unknown so should have a probability distribution reflecting our uncertainty about it before seeing the data
- Therefore we specify a **prior distribution** f(θ)

Note this is like the prevalence in the example

Posterior Distributions

Also *x* is known so should be conditioned on and here we use Bayes theorem to obtain the conditional distribution for unobserved quantities given the data which is known as the **posterior distribution**.

$$f(\theta \mid \mathbf{x}) = \frac{f(\theta) f(\mathbf{x} \mid \theta)}{\int f(\theta) f(\mathbf{x} \mid \theta) d\theta} \propto f(\theta) f(\mathbf{x} \mid \theta)$$

The prior distribution expresses our uncertainty about θ **before** seeing the data.

The posterior distribution expresses our uncertainty about θ **after** seeing the data.

Examples of Bayesian Inference using the Normal distribution

Known variance, unknown mean

It is easier to consider first a model with 1 unknown parameter. Suppose we have a sample of Normal data: $X_i \sim N(\mu, \sigma^2), i = 1, ..., n.$

Let us assume we know the variance, σ^2 and we assume a prior distribution for the mean, μ based on our prior beliefs:

$\mu \sim N(\mu_0, \sigma_0^2)$ Now we wish to construct the posterior distribution f(μ|***x***).

Posterior for Normal distribution mean

So we have

$$f(\mu) = (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\mu-\mu_0)^2/\sigma_0^2)$$

$$f(x_i \mid \mu) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(x_i-\mu)^2/\sigma^2)$$

and hence

$$f(\mu \mid \mathbf{x}) = f(\mu) f(\mathbf{x} \mid \mu)$$

$$= (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\mu-\mu_0)^2/\sigma_0^2) \times$$

$$\prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(x_i-\mu)^2/\sigma^2)$$

$$\propto \exp(-\tfrac{1}{2}\mu^2(1/\sigma_0^2 + n/\sigma^2) + \mu(\mu_0/\sigma_0^2 + \sum_i x_i/\sigma^2) + cons)$$

Posterior for Normal distribution mean (continued)

For a Normal distribution with response *y* with mean θ and variance ϕ we have

$$f(y) = (2\pi\phi)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(y-\theta)^2/\phi\}$$

$$\propto \exp\{-\tfrac{1}{2} y^2\phi^{-1} + y\theta/\phi + cons\}$$

We can equate this to our posterior as follows:

$$\propto \exp(-\tfrac{1}{2}\mu^2(1/\sigma_0^2 + n/\sigma^2) + \mu(\mu_0/\sigma_0^2 + \sum_i x_i/\sigma^2) + cons)$$

$$\rightarrow \phi = (1/\sigma_0^2 + n/\sigma^2)^{-1} \text{ and } \theta = \phi(\mu_0/\sigma_0^2 + \sum_i x_i/\sigma^2)$$

Large sample properties

As $n \to \infty$

$$1/\phi = (1/\sigma_0^2 + n/\sigma^2) \to n/\sigma^2$$

So posterior variance $\to \sigma^2/n$

Posterior mean $\theta = \phi(\mu_0/\sigma_0^2 + \bar{x}/(\sigma^2/n)) \to \bar{x}$

And so the posterior distribution

$$\mu \,|\, \mathbf{x} \to N(\bar{x}, \sigma^2/n)$$

Compared to $\overline{X} \,|\, \mu \sim N(\mu, \sigma^2/n)$

in the Frequentist setting

Sufficient Statistic

- Intuitively, a sufficient statistic for a parameter is a statistic that captures all the information about a given parameter contained in the sample.

- Sufficiency Principle: If $T(\boldsymbol{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\boldsymbol{X}$ only through the value of $T(\boldsymbol{X})$.

- That is, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are two sample points such that $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then the inference about θ should be the same whether $\boldsymbol{X} = \boldsymbol{x}$ or $\boldsymbol{X} = \boldsymbol{y}$.

- Definition: A statistic $T(\boldsymbol{x})$ is a sufficient statistic for θ if the conditional distribution of the sample $\boldsymbol{X}$ given $T(\boldsymbol{x})$ does not depend on θ.

- Definition: Let $X_1, X_2, \ldots X_n$ denote a random sample of size n from a distribution that has a pdf $f(x,\theta)$, $\theta \in \Omega$.

  Let $Y_1=u_1(X_1, X_2, \ldots X_n)$ be a statistic whose pdf or pmf is $f_{Y1}(y_1,\theta)$. Then Y_1 is a sufficient statistic for θ if and only if

$$\frac{f(x_1;\theta)\,f(x_2;\theta)\cdots f(x_n;\theta)}{f_{Y_1}\left[u_1(x_1,x_2,\ldots x_n);\theta\right]} = H(x_1,x_2,\ldots x_n)$$

- Example: Normal sufficient statistic:

  Let $X_1, X_2, \ldots X_n$ be independently and identically distributed $N(\mu,\sigma^2)$ where the variance is known. The sample mean

$$T(\underline{X}) = \bar{X} = 1/n \sum_{i=1}^{n} X_i$$

  is the sufficient statistic for μ.

- Starting with the joint distribution function

$$f\left(\underline{x}\middle|\mu\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right]$$

$$=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left[-\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right]$$

- Next, we add and subtract the sample average yielding

$$f\left(\underline{x}\middle|\mu\right)=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left[-\sum_{i=1}^{n}\frac{\left(x_i-\bar{x}+\bar{x}-\mu\right)^2}{2\sigma^2}\right]$$

$$=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left[-\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+n\left(\bar{x}-\mu\right)^2}{2\sigma^2}\right]$$

- Where the last equality derives from

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(\bar{x}-\mu\right)=\left(\bar{x}-\mu\right)\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=0$$

- Given that the distribution of the sample mean is

$$q\left(T\left(\underline{X}\right)\middle|\theta\right)=\frac{1}{\left(2\pi\sigma^2/n\right)^{1/2}}\exp\left[-\frac{n\left(\bar{x}-\mu\right)^2}{2\sigma^2}\right]$$

- The ratio of the information in the sample to the information in the statistic becomes

$$\frac{f\left(\underline{x}\middle|\theta\right)}{q\left(T\left(\underline{x}\right)\middle|\theta\right)}=\frac{\dfrac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left[-\dfrac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+n\left(\bar{x}-\mu\right)^2}{2\sigma^2}\right]}{\dfrac{1}{\left(2\pi\,\sigma^2/n\right)^{1/2}}\exp\left[-\dfrac{n\left(\bar{x}-\mu\right)^2}{2\sigma^2}\right]}$$

$$\frac{f\left(\underline{x}\middle|\theta\right)}{q\left(T\left(\underline{x}\right)\middle|\theta\right)} = \frac{1}{n^{1/2}\left(2\pi\sigma^2\right)^{n-1/2}} \exp\left[-\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{2\sigma^2}\right]$$

which is a function of the data X_1, X_2, … X_n only, and does not depend on μ. Thus we have shown that the sample mean is a sufficient statistic for μ.

- Theorem (**Factorization Theorem**) Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exists functions $g(t|\theta)$ and $h(\boldsymbol{x})$ such that, for all sample points x and all parameter points θ

$$f\left(\underline{x}\middle|\theta\right) = g\left(T\left(\underline{x}\right)\middle|\theta\right)h\left(\underline{x}\right)$$

Posterior Distribution Through Sufficient Statistics

Theorem: The **posterior distribution** depends only on **sufficient statistics.**

Proof: let T(**X**) be a sufficient statistic for θ, then $f(\mathbf{x}\,|\,\theta) = f(T(\mathbf{x})\,|\,\theta)H(\mathbf{x})$

$$f(\theta\,|\,\mathbf{x}) = \frac{f(\theta)f(\mathbf{x}\,|\,\theta)}{\int f(\theta)f(\mathbf{x}\,|\,\theta)d\theta} = \frac{f(\theta)f(T(\mathbf{x})\,|\,\theta)H(\mathbf{x})}{\int f(\theta)f(T(\mathbf{x})\,|\,\theta)H(\mathbf{x})d\theta}$$

$$= \frac{f(\theta)f(T(\mathbf{x})\,|\,\theta)}{\int f(\theta)f(T(\mathbf{x})\,|\,\theta)d\theta} = f(\theta\,|\,T(\mathbf{x}))$$

Posterior Distribution Through Sufficient Statistics

Example: Posterior for Normal distribution mean (with known variance)

Now, instead of using the entire sample, we can derive the posterior distribution using the sufficient statistic

$$T(\mathbf{x}) = \bar{\mathbf{x}}$$

Exercise: Please derive the posterior distribution using this approach.

Girls Heights Example

- 10 girls aged 18 had both their heights and weights measured.
- Their heights (in cm) where as follows:

169.6,166.8,157.1,181.1,158.4,165.6,166.7,156.5,168.1,165.3

We will assume the variance is known to be 50.

Two individuals gave the following prior distributions for the mean height

Individual 1 $\mu \sim N(165, 2^2)$

Individual 2 $\mu \sim N(170, 3^2)$

Constructing posterior 1

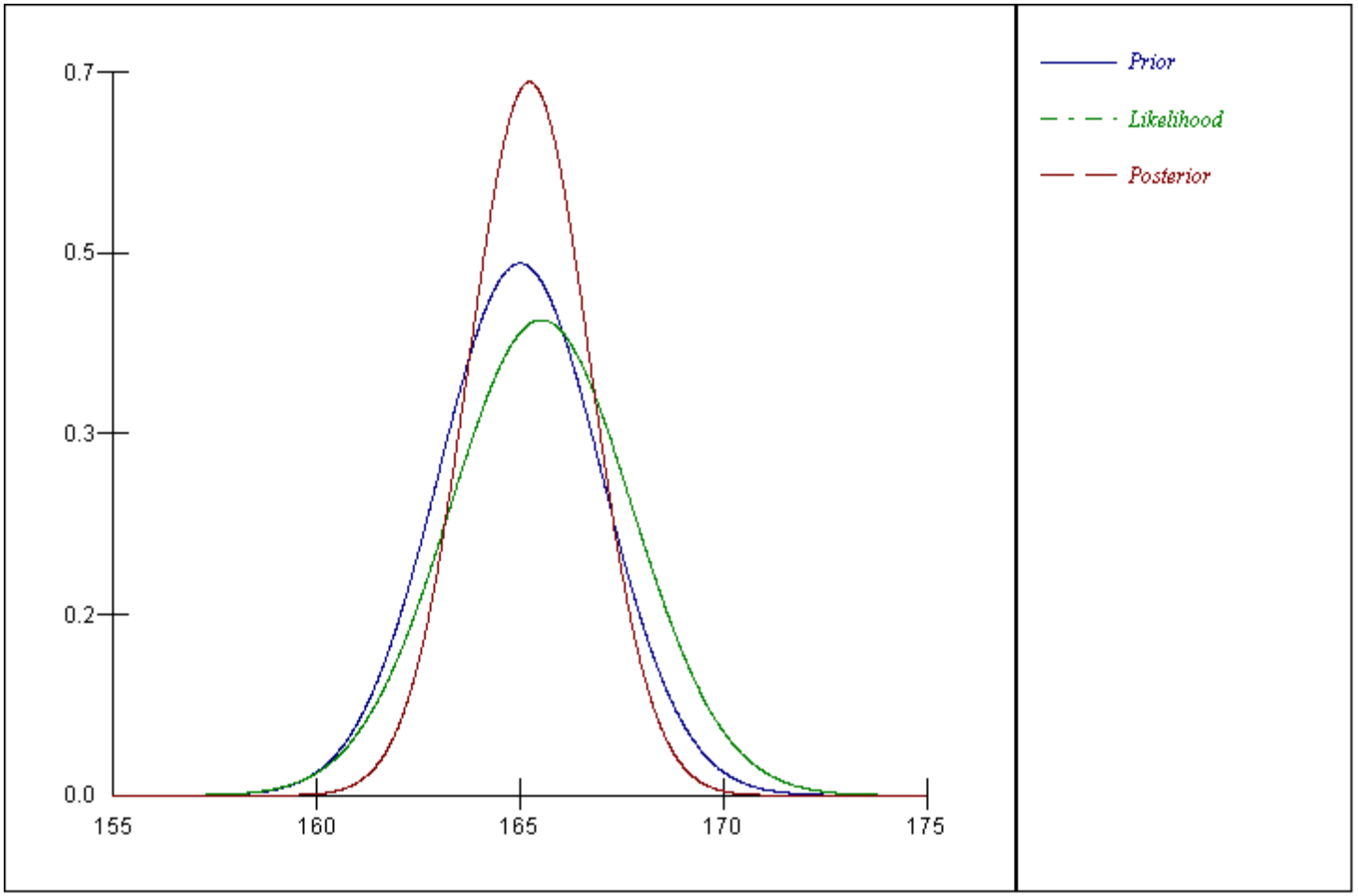
- To construct the posterior we use the formulae we have just calculated
- From the prior, $\mu_0 = 165, \sigma_0^2 = 4$
- From the data, $\bar{x} = 165.52, \sigma^2 = 50, n = 10$
- The posterior is therefore

$$\mu \mid \underline{x} \sim N(\theta_1, \phi_1)$$

$$\textbf{where } \phi_1 = (\tfrac{1}{4} + \tfrac{10}{50})^{-1} = 2.222,$$

$$\theta_1 = \phi_1(\tfrac{165}{4} + \tfrac{1655.2}{50}) = 165.23.$$

Prior and posterior comparison



Constructing posterior 2

- Again to construct the posterior we use the earlier formulae we have just calculaed
- From the prior, $\mu_0 = 170, \sigma_0^2 = 9$
- From the data, $\bar{x} = 165.52, \sigma^2 = 50, n = 10$
- The posterior is therefore

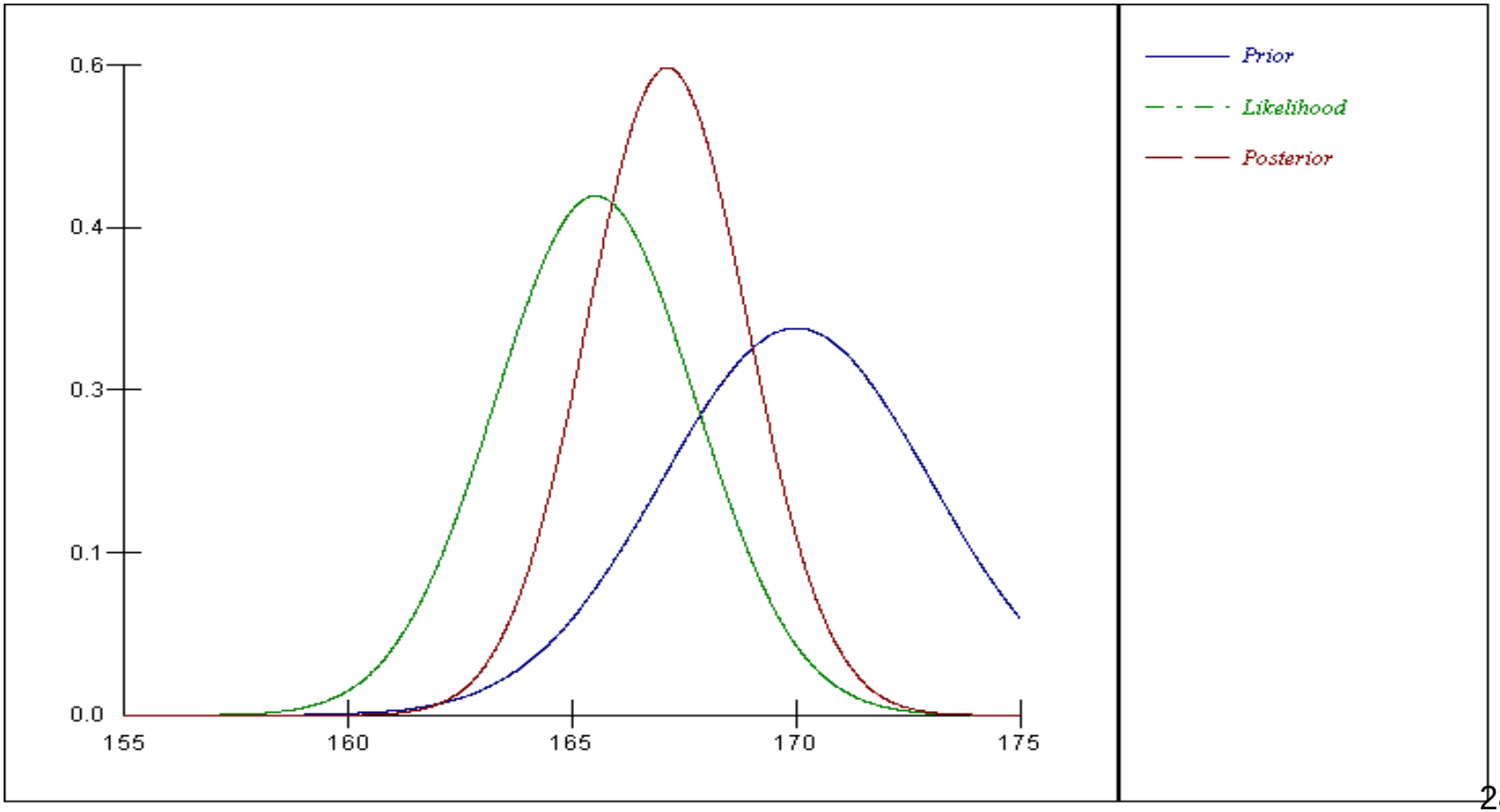
$$\mu \mid \underline{x} \sim N(\theta_2, \phi_2)$$

$$\textbf{where } \phi_2 = (\tfrac{1}{9} + \tfrac{10}{50})^{-1} = 3.214,$$

$$\theta_2 = \phi_2(\tfrac{170}{9} + \tfrac{1655.2}{50}) = 167.12.$$

Prior 2 comparison

Note this prior is not as close to the data as prior 1 and hence posterior is somewhere between prior and likelihood -- represented by the pdf of $\overline{X}$



Other conjugate examples

- When the posterior is in the same family as the prior we have *conjugacy*. Examples include:

Likelihood	Parameter	Prior	Posterior
Normal	Mean	Normal	Normal
Binomial	Probability	Beta	Beta
Poisson	Mean	Gamma	Gamma

In all cases

- The posterior mean is a compromise between the prior mean and the MLE
- The posterior s.d. is less than both the prior s.d. and the s.e. (MLE)

'A Bayesian is one who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule'

-- Senn, 1997--

As $n \rightarrow \infty$

- The posterior mean $\rightarrow$ the MLE
- The posterior s.d. $\rightarrow$ the s.e. (MLE)
- The posterior does not depend on the prior.

Non-informative priors

- We often do not have any prior information, although true Bayesian's would argue we always have some prior information!
- We would hope to have good agreement between the Frequentist approach and the Bayesian approach with a non-informative prior.
- Diffuse or flat priors are often better terms to use as no prior is strictly non-informative!
- For our example of an unknown mean, candidate priors are a Uniform distribution over a large range or a Normal distribution with a huge variance.

Improper priors

- The limiting prior of both the Uniform and Normal is a Uniform prior on the whole real line.
- Such a prior is defined as **improper** as it is not strictly a probability distribution and doesn't integrate to 1.
- Some care has to be taken with improper priors however in many cases they are acceptable provided they result in a proper posterior distribution.
- Uniform priors are often used as non-informative priors however it is worth noting that a uniform prior on one scale can be very informative on another.
- For example: If we have an unknown variance we may put a uniform prior on the variance, standard deviation or log(variance) which will all have different effects.

Point and Interval Estimation

- In Bayesian inference the outcome of interest for a parameter is its full posterior distribution however we may be interested in summaries of this distribution.
- A simple point estimate would be the mean of the posterior. (although the median and mode are alternatives.)
- Interval estimates are also easy to obtain from the posterior distribution and are given several names, for example credible intervals, Bayesian confidence intervals and Highest density regions (HDR). All of these refer to the same quantity.

Definition (Mean Square Error)

Let $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_n)$ denote the vector of observations having joint density $f(\mathbf{x}|\theta)$ with the unknown parameter θ.

Let $\hat{\boldsymbol{\theta}} = h(\mathbf{x})$ be an estimator of the parameter θ. Then the *Mean Square Error* of $h(\mathbf{x})$ is defined to be:

$$M.S.E._{h(\mathbf{x})}(\boldsymbol{\theta}) = E\left[(h(\mathbf{x}) - \boldsymbol{\theta})^2\right]$$

$$= \iint (h(\mathbf{x}) - \boldsymbol{\theta}))^2 \, f(\mathbf{x}\,|\,\boldsymbol{\theta})\, f(\boldsymbol{\theta}) d\mathbf{x} d\boldsymbol{\theta}$$

Bayes Estimator

Theorem: The Bayes estimator that will Minimize the mean square error is

$$h(\mathbf{x}) = \hat{\boldsymbol{\theta}} = E[\boldsymbol{\theta} \mid \mathbf{x}]$$

- That is, the posterior mean.

Bayes Estimator

Lemma: Suppose Z and W are real random variables, then

$$\min_h E\left[(Z - h(W))^2\right] = E\left[(Z - E(Z \mid W))^2\right]$$

That is, the posterior mean E(Z|W) will minimize the quadratic loss (mean square error).

Bayes Estimator

Proof of the Lemma:

$$E(Z-h(W))^2 = E(Z-E(Z|W)+E(Z|W)-h(W))^2$$

$$= E(Z-E(Z|W))^2$$

$$+ 2E\left[(Z-E(Z|W))(E(Z|W)-h(W))\right]$$

$$+ E(E(Z|W)-h(W))^2$$

Conditioning on W, the cross term is zero. Thus

$$E(Z-h(W))^2 = E(Z-E(Z|W))^2 + E(E(Z|W)-h(W))^2$$

Bayes Estimator

Proof of the Theorem (*shorter version):

$$E\left[\left(\boldsymbol{h}(\underline{\boldsymbol{X}})-\theta\right)^2\right]$$

$$=E\left[\left(h(\underline{X})-E(\theta|\underline{X})+E(\theta|\underline{X})-\theta\right)^2\right]$$

$$=E\left[\left(\boldsymbol{h}(\underline{\boldsymbol{X}})-E(\theta|\underline{X})\right)^2\right]+E\left[\left(E(\theta|\underline{X})-\theta\right)^2\right]$$

(* As shown previously, the cross-product =0)

Thus it is minimized by $\boldsymbol{h}(\underline{\boldsymbol{X}})=E(\theta|\underline{X})$

Bayes Estimator

Proof of the Theorem (longer version):

$$E\left[\left(\boldsymbol{h}(\underline{\boldsymbol{X}})-\theta\right)^2\right]=E_{\underline{X}}\left\{E_{\theta|\underline{X}}\left[\left(h(\underline{X})-\theta\right)^2|\underline{X}\right]\right\}$$

=

$$E_{\underline{X}}\left\{E_{\theta|\underline{X}}\left[\left(h(\underline{X})-E(\theta|\underline{X})+E(\theta|\underline{X})-\theta\right)^2|\underline{X}\right]\right\}$$

$$=E_{\underline{X}}\left\{E_{\theta|\underline{X}}\left[\left(\boldsymbol{h}(\underline{\boldsymbol{X}})-E(\theta|\underline{X})\right)^2|\underline{X}\right]\right\}+$$

$$E_{\underline{X}}\left\{E_{\theta|\underline{X}}\left[\left(E(\theta|\underline{X})-\theta\right)^2|\underline{X}\right]\right\}$$ (cross-product =0)

Thus it is minimized by $\boldsymbol{h}(\underline{\boldsymbol{X}})=E(\theta|\underline{X})$

Bayes Estimator

Recall:

Theorem: The **posterior distribution** depends only on **sufficient statistics.**

Therefore, let T(X) be a sufficient statistic, then

$$\hat{\boldsymbol{\theta}} = E[\boldsymbol{\theta} \mid \mathbf{x}] = E[\boldsymbol{\theta} \mid T(\mathbf{x})]$$

Credible Intervals

- If we consider the heights example with our first prior then our posterior is

$$\mu|\underline{x} \sim N(165.23, 2.222),$$

and a 95% credible interval for μ is

$$165.23 \pm 1.96 \times \text{sqrt}(2.222) = (162.31, 168.15).$$

Similarly prior 2 results in a 95% credible interval for μ is

$$(163.61, 170.63).$$

Note that credible intervals can be interpreted in the more natural way that there is a probability of 0.95 that the interval contains μ rather than the Frequentist conclusion that 95% of such intervals contain μ.

Hypothesis Testing

Another big issue in statistical modelling is the ability to test hypotheses and model comparisons in general.

The Bayesian approach is in some ways more straightforward. For an unknown parameter θ

we simply calculate the posterior probabilities

$$p_0 = P(\theta \in \Theta_0 \mid x), \;\; p_1 = P(\theta \in \Theta_1 \mid x)$$

and decide between H_0 and H_1 accordingly.

We also require the prior probabilities to achieve this

$$\pi_0 = P(\theta \in \Theta_0), \;\; \pi_1 = P(\theta \in \Theta_1)$$

Acknowledgement

Part of this presentation was based on publicly available resources posted on-line.