STA 321
Spring 2015

Lecture 3

Thursday, January 22

Law of total probability

- For E and F are evenets,

$$P(E) = P(E \cap F) + P(E \cap F^c)$$

- Example: *A machine produces parts that are either good (90%), slightly defective (2%), or obviously defective (8%). Now assume that a one-year warranty is given for the parts that are shipped to customers. Suppose that a good part fails within the first year with probability 0.01, while a slightly defective part fails within the first year with probability 0.10. What is the probability that a customer receives a part that fails within the first year and therefore is entitled to a warranty replacement?*

Partition

- A collection of events $\{A_1, A_2, \ldots, A_k\}$ to be said a partition of a sample space S if $A_i \cap A_j$ is empty set.

Example: A is any event. Then $\{A, A^c\}$ is a partition.

Example: *A machine produces parts that are either good (90%), slightly defective (2%), or obviously defective (8%).*

Bayes Theorem

- Bayesian statistics named after Rev. Thomas Bayes (1702-1761)
- Bayes Theorem for probability events A and B

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

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

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

Example – coin tossing

- Let A be the event of 2 Heads in three tosses of a fair coin. B be the event of 1st coin is a Head.
- Three coins have 8 equally probable patterns {HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}
- A = {HHT,HTH,THH} $\rightarrow$p(A)=3/8
- B = {HHH,HTH,HTH,HTT} $\rightarrow$p(B)=1/2
- A|B = {HHT,HTH}|{HHH,HTH,HTH,HTT} $\rightarrow$p(A|B)=1/2
- B|A = {HHT,HTH}|{HHT,HTH,THH} $\rightarrow$p(B|A)=2/3
- P(A|B) = P(B|A)P(A)/P(B) = (2/3*3/8)/(1/2) = 1/2

Example 2 – Diagnostic testing

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

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 Testing 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 HIV.

Being Bayesian!

- So the vital issue in this example is *how should this test result change our prior belief that the patient is HIV positive?*
- The disease prevalence (p=0.001) can be thought of as a '*prior*' probability.
- Observing a positive result causes us to modify this probability to p=0.045 which is our '*posterior*' probability that the patient is HIV positive.
- This use of Bayes theorem applied to *observables* is uncontroversial however its use in general statistical analyses where *parameters* are unknown quantities is more controversial.

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 positing a model which specifies $p(x|\theta)$

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** $p(\theta)$

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**.

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

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: $x_i \sim N(\mu, \sigma^2), i = 1, \ldots, n.$

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 p(μ|*x*).

Posterior for Normal distribution mean

So we have

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

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

and hence

$$p(\mu \mid x) = p(\mu) p(x \mid \mu)$$

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

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

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

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:

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

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

Precisions and means

- In Bayesian statistics the precision = 1/variance is often more important than the variance.
- For the Normal model we have

$$1/\phi = (1/\sigma_0^2 + n/\sigma^2) \text{ and } \theta = \phi(\mu_0/\sigma_0^2 + \bar{x}/(\sigma^2/n))$$

In other words the posterior precision = sum of prior precision and data precision, and the posterior mean is a (precision weighted) average of the prior mean and data mean.

Large sample properties

As $n \to \infty$

Posterior precision

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

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 posterior distribution

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

Compared to $p(\bar{x} \mid \mu) = N(\mu, \sigma^2/n)$ in the frequentist setting

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 $p_1(\mu) \sim N(165, 2^2)$

Individual 2 $p_2(\mu) \sim N(170, 3^2)$

Constructing posterior 1

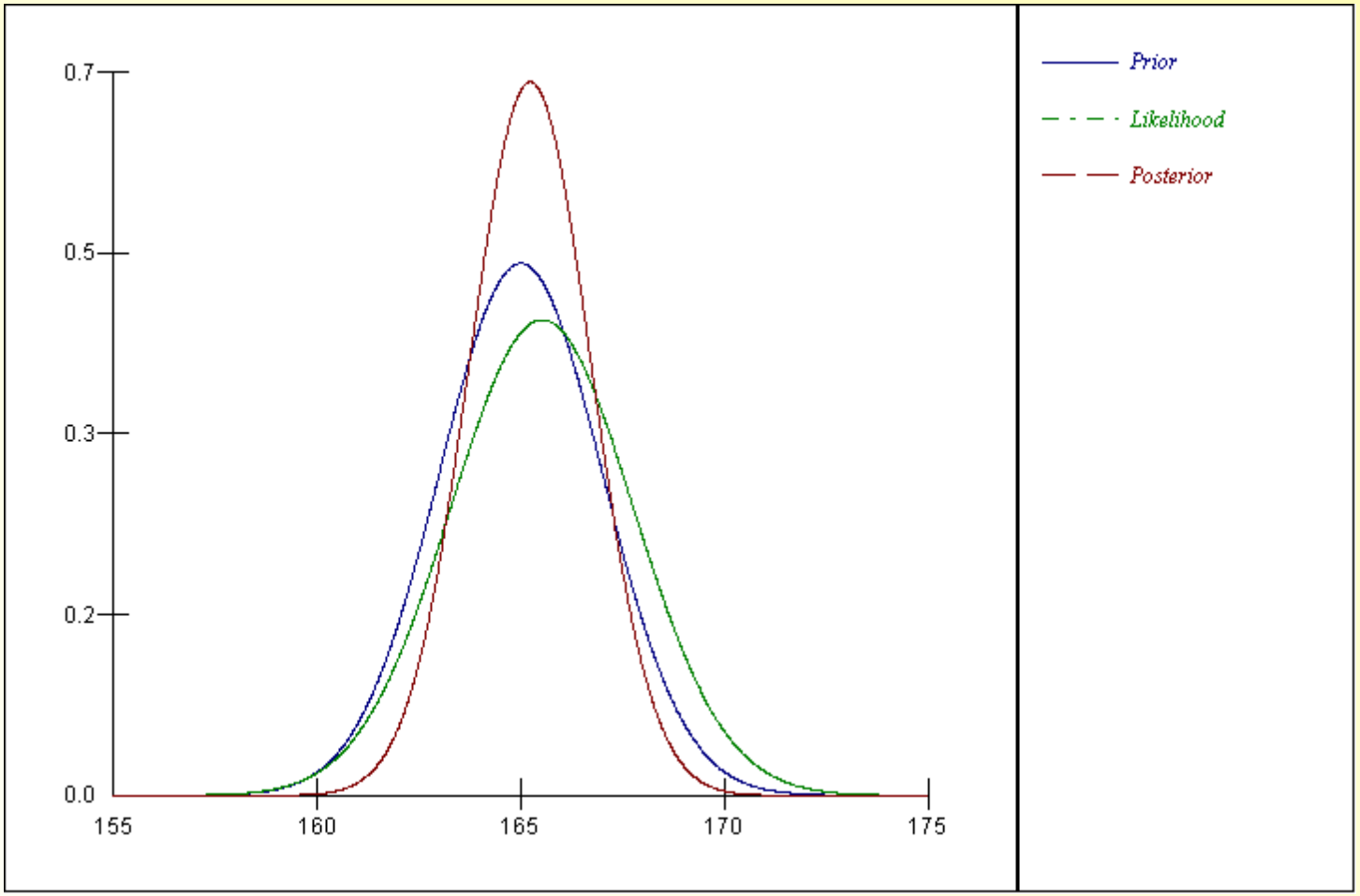
- To construct the posterior we use the formulae we have just calculated
- From the prior, $\mu_0 = 165, \sigma_0^2 = 4$
- From the data, $\bar{x} = 165.52, \sigma^2 = 50, n = 10$
- The posterior is therefore

$$p(\mu \mid x) \sim N(\theta_1, \phi_1)$$

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

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

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

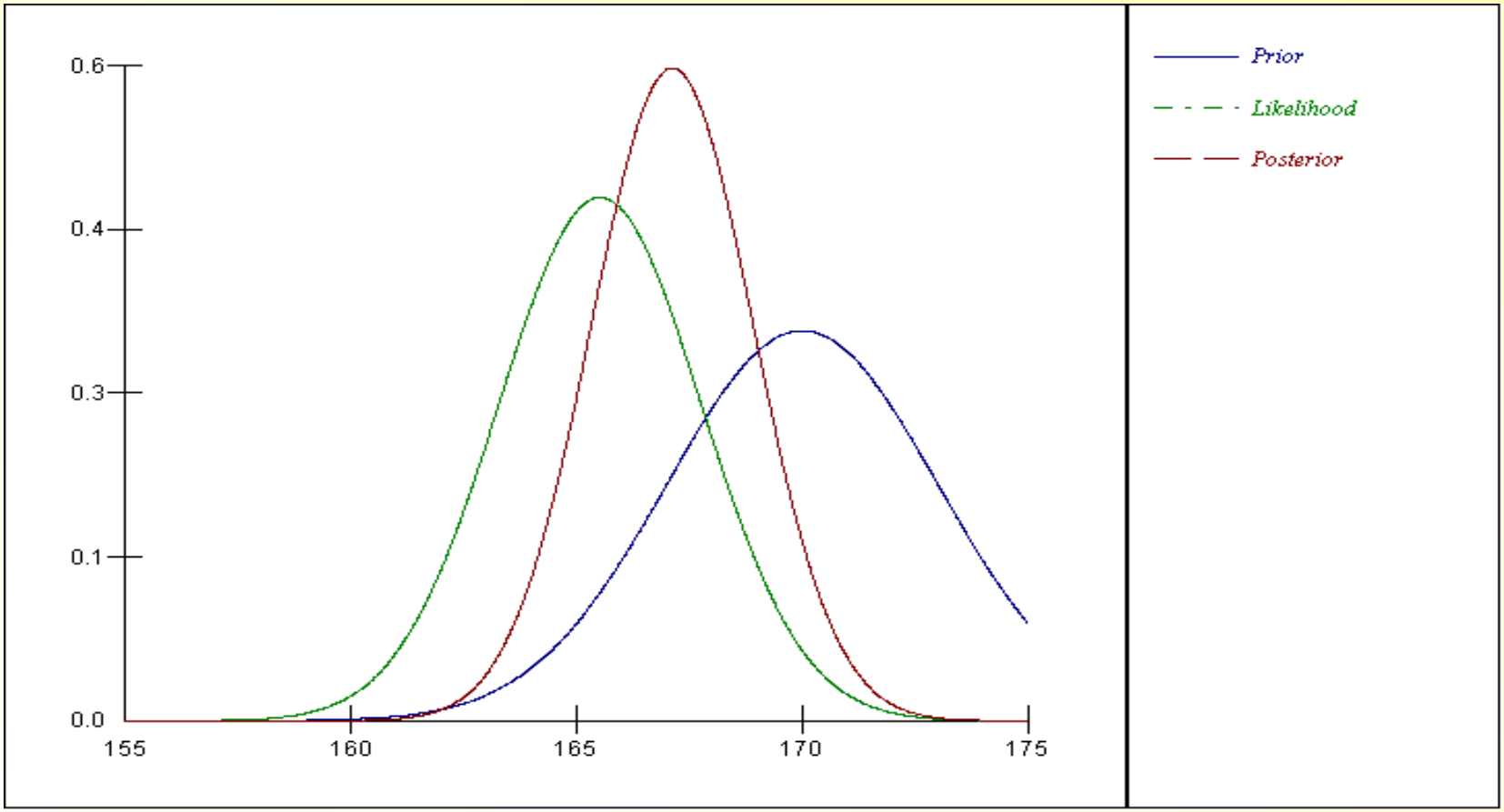
$$p(\mu \mid x) \sim N(\theta_2, \phi_2)$$

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

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

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.



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
Normal	Precision	Gamma	Gamma
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)

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.

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.