What are Bayesian methods?

- Bayesian methods have been widely applied in many areas:
  - medicine / epidemiology
  - genetics
  - ecology
  - environmental sciences
  - social and political sciences
  - finance
  - archaeology
  - .....

- Motivations for adopting Bayesian approach vary:
  - natural and coherent way of thinking about science and learning
  - pragmatic choice that is suitable for the problem in hand

Spiegelhalter et al (2004) define a Bayesian approach as

> 'the explicit use of external evidence in the design, monitoring, analysis, interpretation and reporting of a [scientific investigation]'

They argue that a Bayesian approach is:

- more flexible in adapting to each unique situation
- more efficient in using all available evidence
- more useful in providing relevant quantitative summaries

than traditional methods

Example

A clinical trial is carried out to collect evidence about an unknown 'treatment effect'

Conventional analysis

- p-value for H_0: treatment effect is zero
- Point estimate and CI as summaries of size of treatment effect

Aim is to learn what this trial tells us about the treatment effect

Bayesian analysis

- Asks: 'how should this trial change our opinion about the treatment effect?'

The Bayesian analyst needs to explicitly state

- a reasonable opinion concerning the plausibility of different values of the treatment effect *excluding* the evidence from the trial (the **prior distribution**)
- the support for different values of the treatment effect based *solely* on data from the trial (the **likelihood**),

and to combine these two sources to produce

- a final opinion about the treatment effect (the **posterior distribution**)

The final combination is done using Bayes theorem, which essentially weights the likelihood from the trial with the relative plausibilities defined by the prior distribution

One can view the Bayesian approach as a formalisation of the process of learning from experience

Posterior distribution forms basis for all inference — can be summarised to provide

- point and interval estimates of treatment effect
- point and interval estimates of any function of the parameters
- probability that treatment effect exceeds a clinically relevant value
- prediction of treatment effect in a new patient
- prior information for future trials
- inputs for decision making
- ....

Bayes theorem and its link with Bayesian inference

Bayes' theorem Provable from probability axioms

Let A and B be events, then

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

If A_i is 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: use of Bayes theorem in 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 patient testing positive actually has HIV?

Let A be the event that patient is truly HIV positive, $\overline{A}$ be the event that they are truly HIV negative.

Let B be the event that they test positive.

We want $p(A|B)$.

"95% sensitivity" means that $p(B|A) = .95$.

"98% specificity" means that $p(B|\overline{A}) = .02$.

Now Bayes theorem says

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|\overline{A})p(\overline{A})}.$$

Hence $p(A|B) = \frac{.95 \times .001}{.95 \times .001 + .02 \times .999} = .045$.

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

- Our intuition is poor when processing probabilistic evidence
- The vital issue is *how should this test result change our belief that patient is HIV positive?*
- The disease prevalence can be thought of as a *'prior'* probability ($p = 0.001$)
- Observing a positive result causes us to modify this probability to $p = 0.045$. This is our *'posterior'* probability that patient is HIV positive.
- Bayes theorem applied to *observables* (as in diagnostic testing) is uncontroversial and established
- More controversial is the use of Bayes theorem in general statistical analyses, where *parameters* are the unknown quantities, and their prior distribution needs to be specified — this is **Bayesian inference**

Bayesian inference

Makes fundamental distinction between

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

θ can be statistical parameters, missing data, mismeasured data ...

$\rightarrow$ parameters are treated as random variables

$\rightarrow$ 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

As with any statistical analysis, we start by positing a model which specifies

$$p(x \mid \theta)$$

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

From a Bayesian point of view

- θ is unknown so should have a **probability distribution** reflecting our uncertainty about it before seeing the data
  $\rightarrow$ need to specify a **prior distribution** $p(\theta)$

- x is known so we should condition on it
  $\rightarrow$ use Bayes theorem to obtain conditional probability distribution for unobserved quantities of interest given the data:

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

  This is the **posterior distribution**

The prior distribution $p(\theta)$, expresses our uncertainty about θ **before** seeing the data.

The posterior distribution $p(\theta \mid x)$, expresses our uncertainty about θ **after** seeing the data.

Inference on proportions using a continuous prior

Suppose we now observe r positive responses out of n patients.

Assuming patients are independent, with common unknown response rate θ, leads to a binomial likelihood

$$p(r|n,\theta) \;=\; \left(\begin{array}{c} n \\ r \end{array} \right) \theta^r (1-\theta)^{n-r} \;\propto\; \theta^r (1-\theta)^{n-r}$$

θ needs to be given a continuous prior distribution.

Suppose that, before taking account of the evidence from our trial, we believe all values for θ are equally likely (is this plausible?) $\Rightarrow \theta \sim$ Unif(0, 1) i.e. $p(\theta) = \frac{1}{1-0} = 1$

Posterior is then

$$p(\theta|r,n) \propto \theta^r (1-\theta)^{(n-r)} \times 1$$

This has form of the *kernel* of a Beta(r+1, n-r+1) distribution (see lect 1), where

$$\theta \;\sim\; \text{Beta}(a,b) \;\equiv\; \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \theta^{a-1}(1-\theta)^{b-1}$$

To represent external evidence that some response rates are more plausible than others, it is mathematically convenient to use a Beta(a, b) prior distribution for θ

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Combining this with the binomial likelihood gives a posterior distribution

$$\begin{aligned} p(\theta \mid r, n) &\propto p(r \mid \theta, n)p(\theta) \\ &\propto \theta^{r}(1-\theta)^{n-r}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{r+a-1}(1-\theta)^{n-r+b-1} \\ &\propto \text{Beta}(r+a,\ n-r+b) \end{aligned}$$

Comments

- When the prior and posterior come from the same family of distributions the prior is said to be **conjugate** to the likelihood
  - Occurs when prior and likelihood have the same 'kernel'

- Recall from lecture 1 that a Beta(a, b) distribution has

$$\begin{aligned} \text{mean} &= a/(a+b), \\ \text{variance} &= ab/\left[(a+b)^2(a+b+1)\right] \end{aligned}$$

  Hence posterior mean is $E(\theta|r,n) = (r+a)/(n+a+b)$

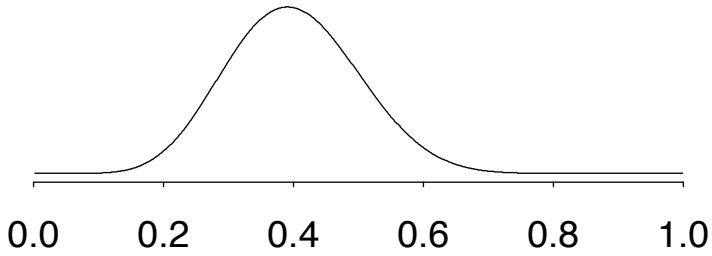
- a and b are equivalent to observing a priori $a-1$ successes in $a+b-2$ trials $\rightarrow$ can be elicited

- With fixed a and b, as r and n increase, $E(\theta|r,n) \rightarrow r/n$ (the MLE), and the variance tends to zero
  - This is a general phenomenon: as n increases, posterior distribution gets more concentrated and the likelihood dominates the prior

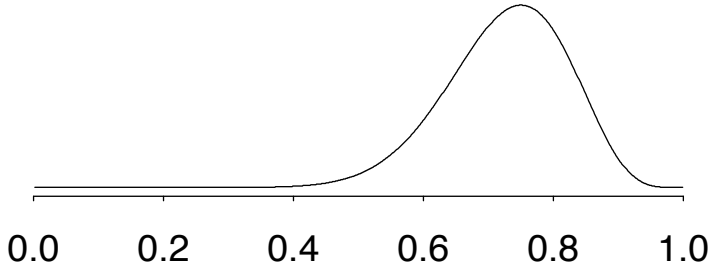
- A Beta(1, 1) is equivalent to Uniform(0, 1)

Example: Drug

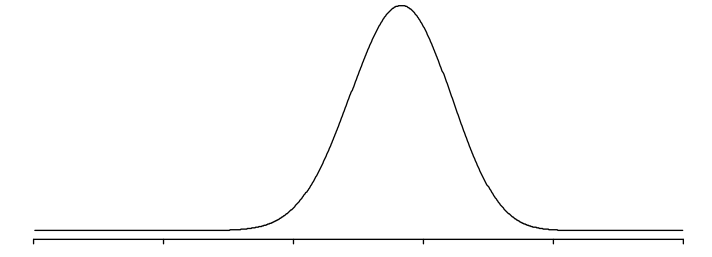
- Recall example from lecture 1, where we consider early investigation of a new drug
- Experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible
- We interpreted this as a distribution with mean = 0.4, standard deviation 0.1 and showed that a Beta(9.2,13.8) distribution has these properties
- Suppose we now treat $n = 20$ volunteers with the compound and observe $y = 15$ positive responses



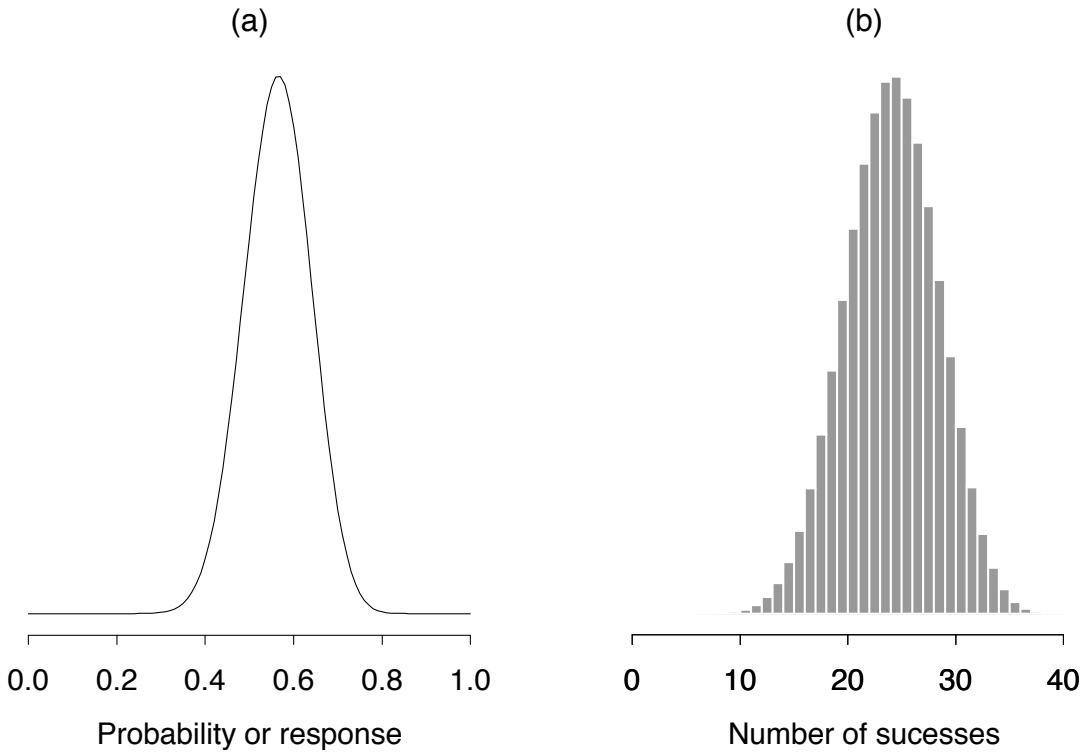
Beta(9.2, 13.8) prior distribution supporting response rates between 0.2 and 0.6



Likelihood arising from a Binomial observation of 15 successes out of 20 cases



Parameters of the Beta distribution are updated to (a+15, b+20-15) = (24.2, 18.8): mean 24.2/(24.2+18.8) = 0.56



(a) Beta posterior distribution after having observed 15 successes in 20 trials

(b) predictive Beta-Binomial distribution of the number of successes $\tilde{y}_{40}$ in the next 40 trials with mean 22.5 and standard deviation 4.3

Suppose we would consider continuing a development program if the drug managed to achieve at least a further 25 successes out of these 40 future trials

From Beta-binomial distribution, can calculate $P(\tilde{y}_{40} \geq 25) = 0.329$

Drug (continued): learning about parameters from data using Markov chain Monte-Carlo (MCMC) methods in WInBUGS

- Using MCMC (e.g. in WinBUGS), no need to explicitly specify posterior
- Can just specify the prior and likelihood separately
- WinBUGS contains algorithms to evaluate the posterior given (almost) arbitrary specification of prior and likelihood
  - posterior doesn't need to be closed form
  - but can (usually) recognise conjugacy when it exists

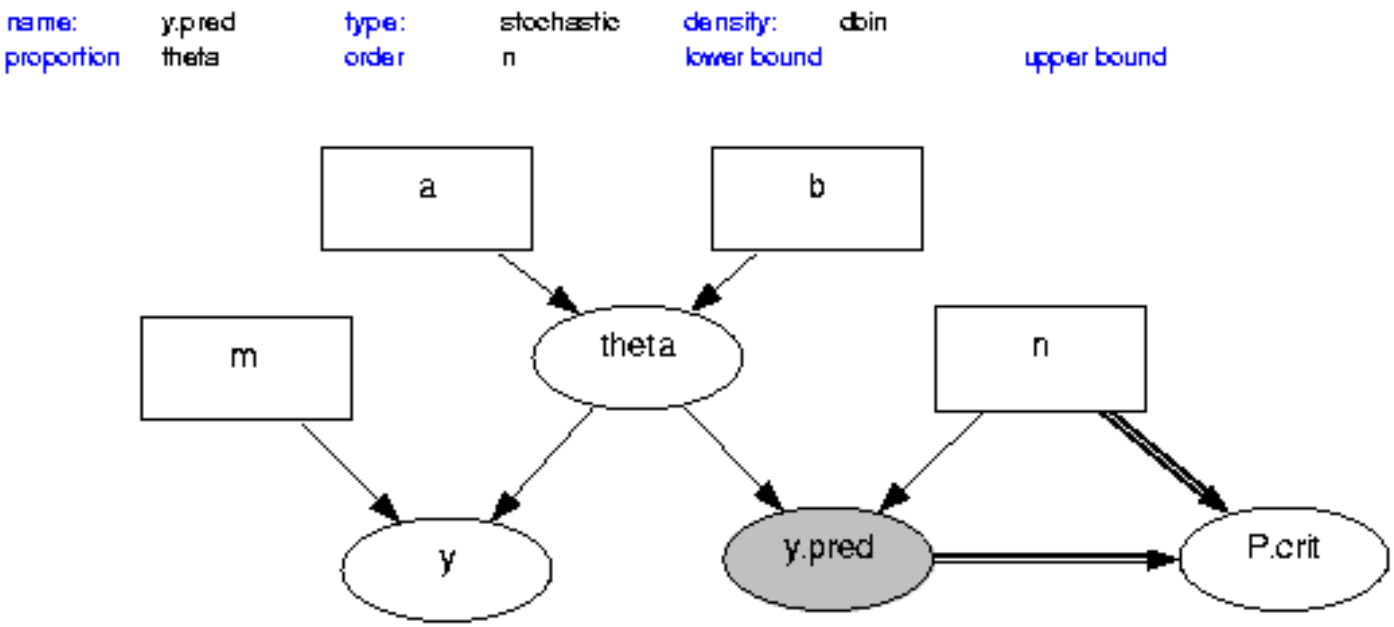
The drug model can be written

$$\begin{aligned}
\theta &\sim \text{Beta}[a, b] && \text{prior distribution} \\
y &\sim \text{Binomial}[\theta, m] && \text{sampling distribution} \\
y_{\text{pred}} &\sim \text{Binomial}[\theta, n] && \text{predictive distribution} \\
P_{\text{crit}} &= P(y_{\text{pred}} \geq n_{\text{crit}}) && \text{Probability of exceeding critical threshold}
\end{aligned}$$

```
# In BUGS syntax:

# Model description '
model {
  theta     ~ dbeta(a,b)              # prior distribution
  y         ~ dbin(theta,m)           # sampling distribution
  y.pred    ~ dbin(theta,n)           # predictive distribution
  P.crit   <- step(y.pred-ncrit+0.5)  # =1 if y.pred >= ncrit, 0 otherwise
}
```

Graphical representation of models



Note that adding data to a model is simply extending the graph.

Data files

Data can be written after the model description, or held in a separate .txt or .odc file

```
 list( a = 9.2,    # parameters of prior distribution
  b = 13.8,
  y = 15,      # number of successes
  m = 20,      # number of trials
  n = 40,      # future number of trials
 ncrit = 25)  # critical value of future successes
```

Alternatively, in this simple example, we could have put all data and constants into model description:

```
model{
  theta     ~ dbeta(9.2,13.8)      # prior distribution
  y         ~ dbin(theta,20)       # sampling distribution
  y.pred    ~ dbin(theta,40)       # predictive distribution
  P.crit   <- step(y.pred-24.5)    # =1 if y.pred >= ncrit, 0 otherwise

  y         <- 15
}
```

The WinBUGS data formats

WinBUGS accepts data files in:

1. Rectangular format (easy to cut and paste from spreadsheets)

```
 n[] r[]
 47  0
148 18
 ...
360 24
END
```

2. S-Plus format:

```
list(N=12,n = c(47,148,119,810,211,196,
          148,215,207,97,256,360),
     r = c(0,18,8,46,8,13,9,31,14,8,29,24))
```

Generally need a 'list' to give size of datasets etc.

Initial values

- WinBUGS can automatically generate initial values for the MCMC analysis using *gen inits*
- Fine if have informative prior information
- If have fairly 'vague' priors, better to provide reasonable values in an initial-values list

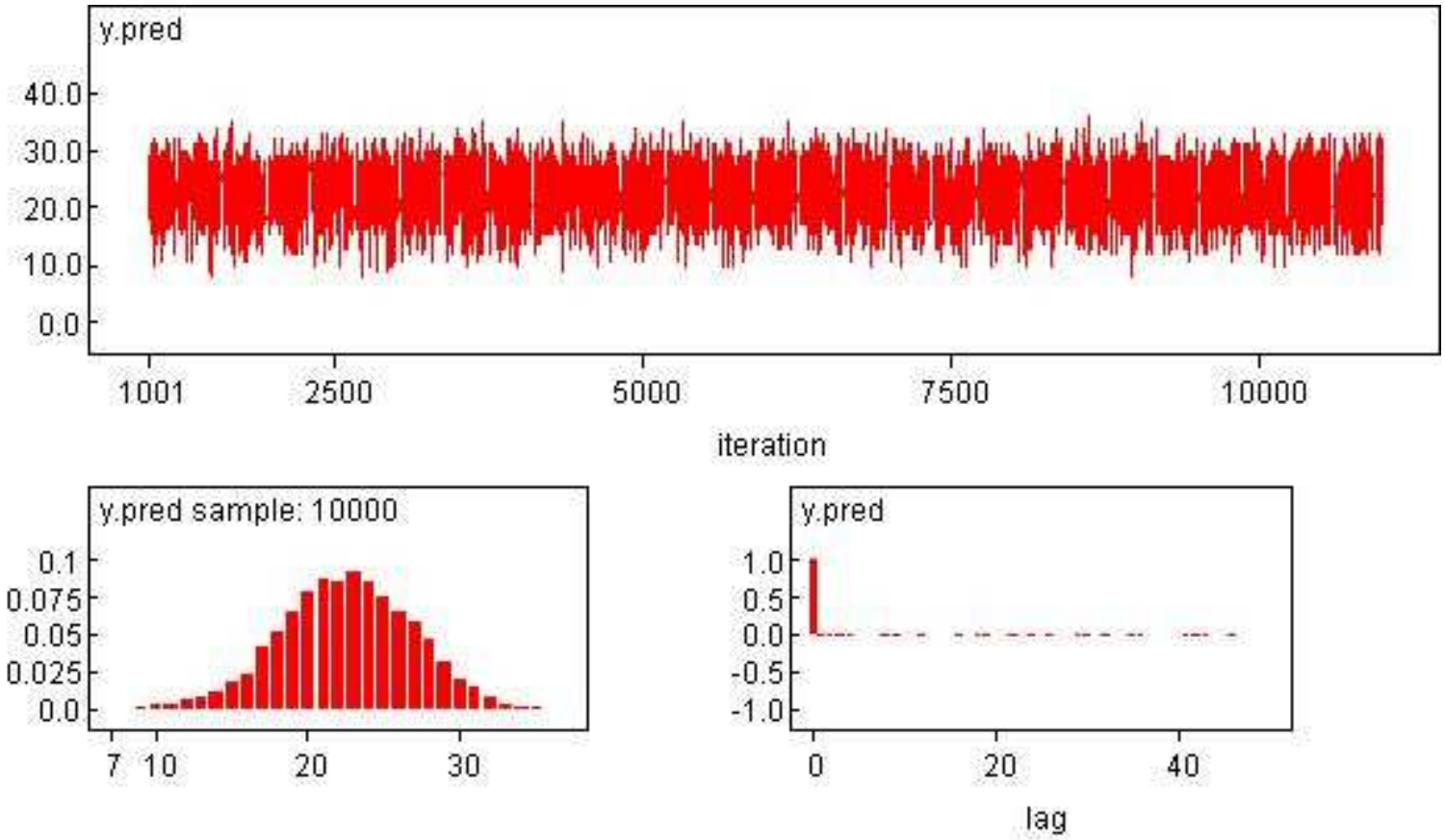
Initial values list can be after model description or in a separate file

```
list(theta=0.1)
```

Running WinBUGS for MCMC analysis (single chain)

1. Open *Specification tool* from *Model* menu.
2. Program responses are shown on bottom-left of screen.
3. Highlight `model` by double-click. Click on *Check model*.
4. Highlight start of data. Click on *Load data*.
5. Click on *Compile*.
6. Highlight start of initial values. Click on *Load inits*.
7. Click on *Gen Inits* if more initial values needed.
8. Open *Update* from *Model* menu.
9. Click on *Update* to burn in.
10. Open *Samples* from *Inference* menu.
11. Type nodes to be monitored into *Sample Monitor*, and click *set* after each.
12. Perform more updates.
13. Type * into *Sample Monitor*, and click *stats* etc to see results on all monitored nodes.

WinBUGS output



WinBUGS output and exact answers

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta	0.5633	0.07458	4.292E-4	0.4139	0.5647	0.7051	1001	30000
y.pred	22.52	4.278	0.02356	14.0	23.0	31.0	1001	30000
P.crit	0.3273	0.4692	0.002631	0.0	0.0	1.0	1001	30000

Exact answers from conjugate analysis

- θ: mean 0.563 and standard deviation 0.075
- Y^{pred}: mean 22.51 and standard deviation 4.31.
- Probability of at least 25: 0.329

MCMC results are within Monte Carlo error of the true values

Bayesian inference using the Normal distribution

Known variance, unknown mean

Suppose we have a sample of Normal data $x_i \sim \mathrm{N}(\theta, \sigma^2)$ $(i = 1, ..., n)$.

For now assume σ^2 is known and θ has a Normal prior $\theta \sim \mathrm{N}(\mu, \sigma^2/n_0)$

Same standard deviation σ is used in the likelihood and the prior. Prior variance is based on an 'implicit' sample size n_0

Then straightforward to show that the posterior distribution is

$$\theta|\boldsymbol{x} \sim \mathrm{N}\left(\frac{n_0\mu + n\overline{x}}{n_0 + n}, \frac{\sigma^2}{n_0 + n}\right)$$

- As n_0 tends to 0, the prior variance becomes larger and the distribution becomes 'flatter', and in the limit the prior distribution becomes essentially uniform over $-\infty, \infty$

- Posterior mean $(n_0\mu + n\overline{x})/(n_0 + n)$ is a weighted average of the prior mean μ and parameter estimate $\overline{x}$, weighted by their precisions (relative 'sample sizes'), and so is always a compromise between the two

- Posterior variance is based on an implicit sample size equivalent to the sum of the prior 'sample size' n_0 and the sample size of the data n

- As $n \to \infty$, $p(\theta|\boldsymbol{x}) \to \mathrm{N}(\overline{x}, \sigma^2/n)$ which does not depend on the prior

- Compare with frequentist setting, the MLE is $\widehat{\theta} = \overline{x}$ with $\mathrm{SE}(\widehat{\theta}) = \sigma/\sqrt{n}$, and sampling distribution

$$p(\widehat{\theta} \mid \theta) = p(\overline{x}|\theta) = \mathrm{N}(\theta, \sigma^2/n)$$

Example: THM concentrations

- Regional water companies in the UK are required to take routine measurements of trihalomethane (THM) concentrations in tap water samples for regulatory purposes
- Samples are tested throughout the year in each water supply zone
- Suppose we want to estimate the average THM concentration in a particular water zone, z
- Two independent measurements, x_{z1} and x_{z2} are taken and their mean, $\overline{x}_z$ is 130 $\mu g/l$
- Suppose we know that the assay measurement error has a standard deviation $\sigma_{[e]} = 5\mu g/l$
- What should we estimate the mean THM concentration to be in this water zone?

Let the mean THM conc. be denoted θ_z.

A standard analysis would use the sample mean $\overline{x}_z = 130\mu g/l$ as an estimate of θ_z, with standard error $\sigma_{[e]}/\sqrt{n} = 5/\sqrt{2} = 3.5\mu g/l$

A 95% confidence interval is $\overline{x}_z \pm 1.96 \times \sigma_{[e]}/\sqrt{n}$, i.e. 123.1 to 136.9 $\mu g/l$.

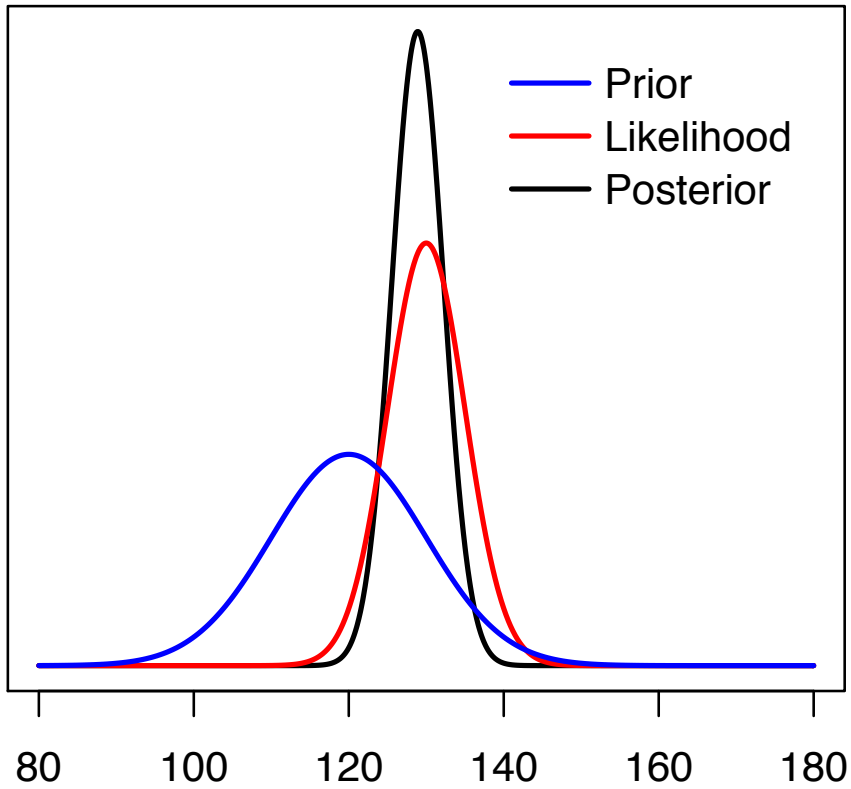
Suppose historical data on THM levels in other zones supplied from the same source showed that the mean THM concentration was 120 $\mu g/l$ with standard deviation 10 $\mu g/l$

- suggests Normal$(120, 10^2)$ prior for θ_z
- if we express the prior standard deviation as $\sigma_{[e]}/\sqrt{n_0}$, we can solve to find $n_0 = (\sigma_{[e]}/10)^2 = 0.25$
- so our prior can be written as $\theta_z \sim$ Normal$(120, \sigma^2_{[e]}/0.25)$

Posterior for θ_z is then

$$\begin{aligned} p(\theta_z|\boldsymbol{x}) &= \text{Normal}\left(\frac{0.25 \times 120 + 2 \times 130}{0.25 + 2}, \ \frac{5^2}{0.25 + 2}\right) \\ &= \text{Normal}(128.9, \ 3.33^2) \end{aligned}$$

giving 95% interval for θ_z of 122.4 to 135.4$\mu g/l$


Prior
Likelihood
Posterior
80
100
120
140
160
180
mean THM concentration, ug/l (theta)

Prediction

Denoting the posterior mean and variance as $\mu_n = (n_0\mu + n\bar{x})/(n_0 + n)$ and $\sigma_n^2 = \sigma^2/(n_0 + n)$, the *predictive distribution* for a new observation $\tilde{x}$ is

$$p(\tilde{x}|\boldsymbol{x}) = \int p(\tilde{x}|\boldsymbol{x}, \theta)p(\theta|\boldsymbol{x})d\theta$$

which generally simplifies to

$$p(\tilde{x}|\boldsymbol{x}) = \int p(\tilde{x}|\theta)p(\theta|\boldsymbol{x})d\theta$$

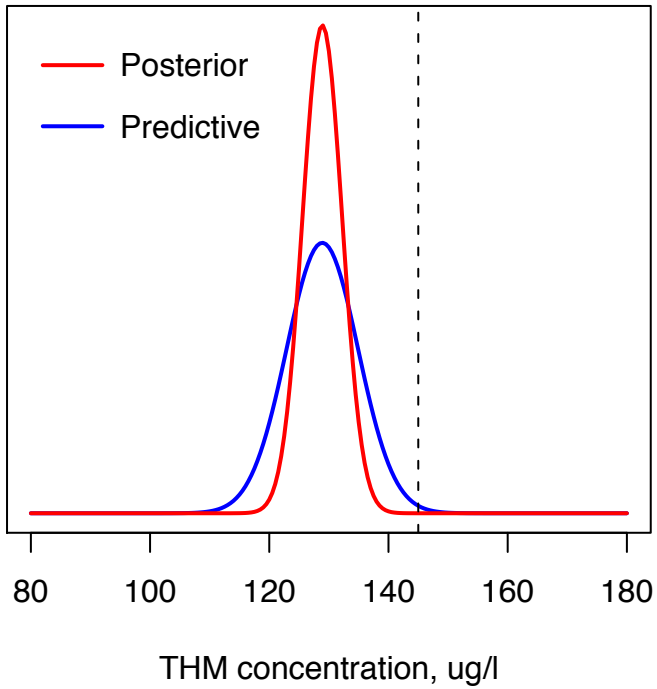
which can be shown to give

$$p(\tilde{x}|\boldsymbol{x}) \sim \mathsf{N}\left(\mu_n, \sigma_n^2 + \sigma^2\right)$$

So the predictive distribution is centred around the posterior mean with variance equal to sum of the posterior variance and the sample variance of $\tilde{x}$

Example: THM concentration (continued)

- Suppose the water company will be fined if THM levels in the water supply exceed $145\mu g/l$
- Predictive distribution for THM concentration in a future sample taken from the water zone is $\mathrm{N}(128.9, 3.33^2 + 5^2) = \mathrm{N}(128.9, 36.1)$
- Probability that THM concentration in future sample exceeds $145\mu g/l$ is $1 - \Phi[(145 - 128.9)/\sqrt{(36.1)}] = 0.004$



Bayesian inference using count data

Suppose we have an independent sample of counts $x_1, ..., x_n$ which can be assumed to follow a Poisson distribution with unknown mean μ:

$$p(\boldsymbol{x}|\mu) = \prod_i \frac{\mu^{x_i} e^{-\mu}}{x_i!}$$

The kernel of the Poisson likelihood (as a function of μ) has the same form as that of a Gamma(a, b) prior for μ:

$$p(\mu) = \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu}$$

Note: A Gamma(a, b) density has mean a/b and variance a/b^2

This implies the following posterior

$$\begin{aligned} p(\mu \mid \boldsymbol{x}) &\propto p(\mu)\, p(\boldsymbol{x} \mid \mu) \\ &= \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu} \prod_{i=1}^{n} e^{-\mu} \frac{\mu^{x_i}}{x_i!} \\ &\propto \mu^{a+n\overline{x}-1}\, e^{-(b+n)\mu} \\ &= \text{Gamma}(a + n\overline{x},\, b + n). \end{aligned}$$

The posterior is another (different) Gamma distribution.

The Gamma distribution is said to be the *conjugate* prior.

$$E(\mu \mid \boldsymbol{x}) = \frac{a + n\overline{x}}{b + n} = \overline{x} \left(\frac{n}{n+b} \right) + \frac{a}{b} \left(1 - \frac{n}{n+b} \right)$$

So posterior mean is a compromise between the prior mean a/b and the MLE $\overline{x}$

Example: London bombings during WWII

- Data below are the number of flying bomb hits on London during World War II in a 36 km^2 area of South London

- Area was partitioned into 0.25 km^2 grid squares and number of bombs falling in each grid was counted

Hits, x	0	1	2	3	4	7
Number of areas, n	229	211	93	35	7	1

Total hits, $\sum_i n_i x_i = 537$
Total number of areas, $\sum_i n_i = 576$

- If the hits are random, a Poisson distribution with constant hit rate θ should fit the data

- Can think of $n = 576$ observations from a Poisson distribution, with $\overline{x} = 537/576 = 0.93$

The 'invariant' Jeffreys prior (see later) for the mean θ of a Poisson distribution is $p(\theta) \propto 1/\sqrt{\theta}$, which is equivalent to an (improper) Gamma(0.5,0) distribution. Therefore

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \text{Gamma}(a + n\overline{x}, b + n) = \text{Gamma}(537.5, 576) \\ \mathbb{E}(\theta|\boldsymbol{y}) &= \frac{537.5}{576} = 0.933; \qquad \text{Var}(\theta|\boldsymbol{y}) = \frac{537.5}{576^2} = 0.0016 \end{aligned}$$

Note that these are almost exactly the same as the MLE and the square of the SE(MLE)

Summary

For all these examples, we see that

- the posterior mean is a compromise between the prior mean and the MLE
- the posterior s.d. is less than each of 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 $n \to \infty$,

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

These observations are generally true, when the MLE exists and is unique

Choosing prior distributions

When the posterior is in the same family as the prior then we have what is known as *conjugacy*. This has the advantage that prior parameters can usually be interpreted as a *prior sample*. Examples include:

Likelihood	Parameter	Prior	Posterior
Normal	mean	Normal	Normal
Normal	precision	Gamma	Gamma
Binomial	success prob.	Beta	Beta
Poisson	rate or mean	Gamma	Gamma

- Conjugate prior distributions are mathematically convenient, but do not exist for all likelihoods, and can be restrictive

- Computations for non-conjugate priors are harder, but possible using MCMC (see next lecture)

Calling WinBUGS from other software

- Scripts enable WinBUGS 1.4 to be called from other software
- Interfaces developed for R, Splus, SAS, Matlab
- See `www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`
- Andrew Gelman's `bugs` function for R is most developed - reads in data, writes script, monitors output etc. Now packaged as `R2WinBUGS`.
- OpenBUGS site http://mathstat.helsinki.fi/openbugs/ provides an open source version, including `BRugs` package which works from within R