Advanced Topics in Statistics

~ Bayesian inference methods ~

2

BAYESIAN METHOD.

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.

Bayesian methods have been widely applied in many areas:

- medicine / epidemiology
- genetics
- ecology
- environmental sciences
- social and political sciences
- finance
- archaeology
- ...

Bayesian approach.

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.

Example.

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

Conventional analysis

- Calculates p-value for 'H_0: treatment effect is zero'.
- Finds 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?'

Bayesian analysis - Learning from experience.

The Bayesian analyst needs to explicitly state

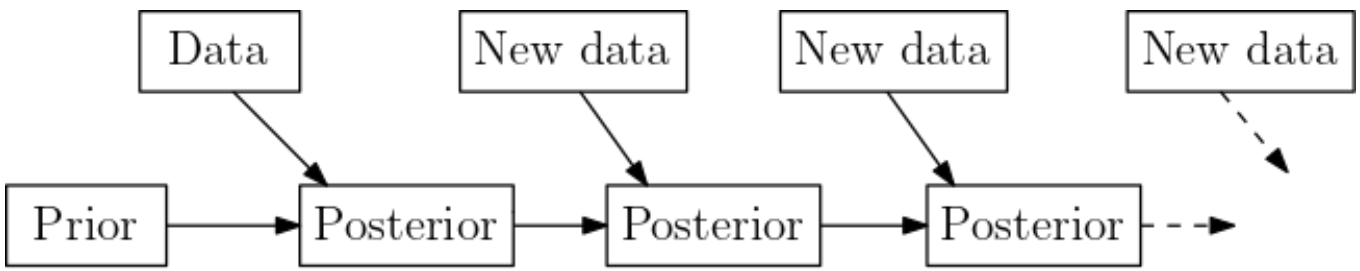
- a reasonable opinion concerning the plausibility of different values of the treatment effect *excluding* the evidence from the trial (the prior distribution),
- the support for different values of the treatment effect based *solely* on data from the trial (the likelihood),

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.

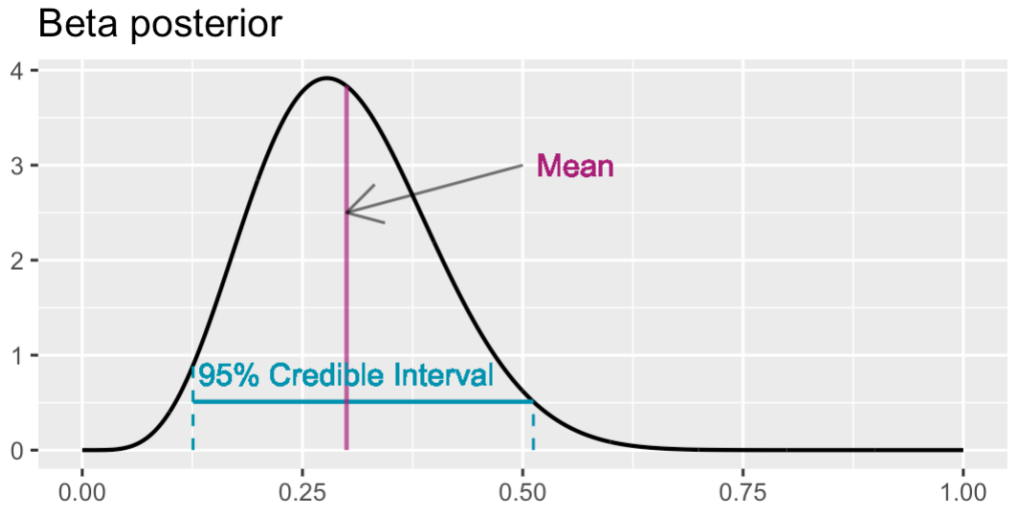


Bayesian analysis - the Posterior.

In Bayesian analysis, the posterior distribution forms the basis for all inference.

It 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, etc.



Bayes Theorem for observables.

Bayes Formula. Let A and B be events, then

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

Bayes Theorem. 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

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

Bayes theorem applied to *observables* (as in diagnostic testing) is uncontroversial and established.

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 a patient testing positive actually has HIV?

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 a patient testing positive actually has HIV?

Let A be the event that the patient is truly HIV positive, A^C 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 $p(B|A^C) = .02$.

Now Bayes theorem says

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

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

- 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 the patient is HIV positive.

Bayes Theorem in statistical analyses.

- In the diagnostic testing example the vital issue is how should this test result change our belief that the patient is HIV positive? This is the idea that Bayesian statistics extends to general statistical analysis.
- But while Bayes theorem applied to *observables* (as in diagnostic testing) is uncontroversial and established, it is more controversial in general statistical analyses. (Hence the Frequentist-Bayesian debate).

Bayesian inference makes fundamental distinction between

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

  Note that θ can denote any unknown quantity, such as statistical parameters, missing data, mismeasured data, etc.

  These unknown quantities are all treated as random variables.

In the Bayesian framework, we make probability statements about model parameters.

(Recall that in the frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data).

Bayesian Inference.

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

$$p(x|\theta).$$

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.

  Therefore, we need to specify a prior distribution $p(\theta)$.

- x is known so we should condition on it.

  We can use Bayes theorem to obtain conditional probability distributions for unobserved quantities of interest given the data, this is the posterior distribution, $p(\theta|x)$.

Note that the prior distribution $p(\theta)$ expresses our uncertainty about θ **before** seeing the data, while the posterior distribution $p(\theta|x)$ expresses our uncertainty about θ **after** seeing the data.

Bayesian Inference - Bayes Theorem.

Using the previous notation, Bayes Theorem takes the form

$$p(\theta|x) = \frac{p(\theta)\, p(x|\theta)}{\int p(\theta)\, p(x|\theta)\, \mathrm{d}\theta}.$$

- For given data the denominator of the above expression is just a normalising constant.
- Since a probability density function always integrates to one, the numerator, in this case, fully characterises the distribution.
- Therefore, in Bayesian inference the normalisation constant is often ignored, and only the terms that depend on the parameter are considered. This part of the distribution is called the kernel.
- We often write

  $$p(\theta|x) \propto p(\theta)\, p(x|\theta),$$

  which means that the posterior is proportional to the product of the prior and the likelihood:

  $$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

Conjugacy.

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.

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 later).

POSTERIOR PREDICTIVE DISTRIBUTION.

Recall, that before observing the data, we can provide its predictive distribution by integrating out the unknown parameter

$$p(x) = \int p(x|\theta)p(\theta)\mathrm{d}\theta.$$

This formula can be adapted to provide prediction for *future observations* once data have been collection.

This so-called posterior predictive distribution for a new observation $\tilde{x}$ is defined as

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

which generally simplifies to

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

The posterior predictive distribution accounts for our posterior uncertainty about θ.

INFERENCE ON PROPORTIONS.

A typical scenario when we are interested in proportions is when we are assessing a treatment effect in a fixed number of patients.

Suppose we observe r positive responses out of n patients, which can be thought of as r successes out of n trials.

Assuming patients are independent, with common unknown response rate θ, leads to a **Binomial(θ, n) likelihood**

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

Since theoretically the parameter θ can take any value in the interval $[0,1]$, it needs to be given a continuous prior distribution that's restricted to the unit interval.

We have already seen that for proportions it is convenient to use a **Beta(a,b) prior** distribution,

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

Combining this with the binomial likelihood gives a **Beta$(r+a, n-r+b)$ posterior** distribution

$$p(\theta|r,n) \propto \theta^{r+a-1}(1-\theta)^{n-r+b-1} \propto Beta(r+a, n-r+b).$$

Inference on proportions - Comments.

Comments.

- A Beta$(1, 1)$ is equivalent to Uniform$(0, 1)$. Thus a Beta prior is suitable even when we believe that all values for θ are equally likely.
- The Beta prior is a conjugate prior for the binomial likelihood.
- As a result of conjugacy, the prior parameters can be interpreted as a prior sample:

  a and b are equivalent to observing a priori $a - 1$ successes in $a + b - 2$ trials.
- Recall that a Beta(a, b) distribution has

$$\text{mean} = \frac{a}{a+b}, \qquad \text{variance} = \frac{ab}{(a+b)^2(a+b+1)}.$$

  Hence the posterior mean is $E(\theta|r, n) = (r + a)/(n + a + b)$.
- With fixed a and b, as r and n increase, $E(\theta|r, n) \to r/n$ (which is the maximum likelihood estimate of θ), and the variance tends to zero.

  This is a general phenomenon: as n increases, the posterior distribution gets more concentrated and the likelihood dominates the prior.

Inference on proportions - Drug example.

Recall our previous Drug example, where we considered the 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. (Note that $y = 15$ is our data).

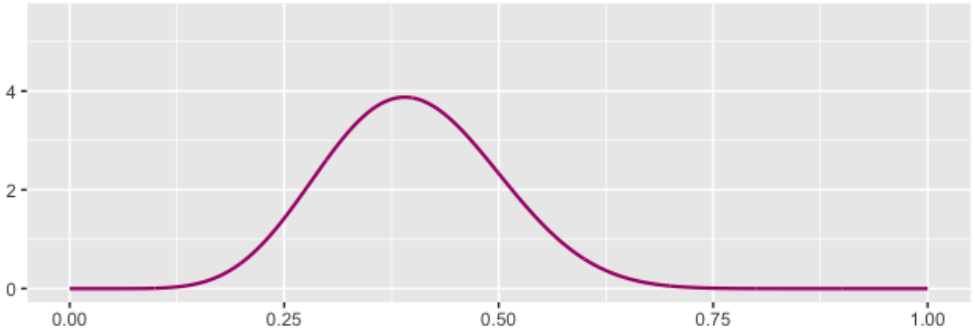
  The Binomial observation then gives the following likelihood

$$\text{likelihood} = \binom{20}{15} \theta^{15} (1-\theta)^{20-15}.$$

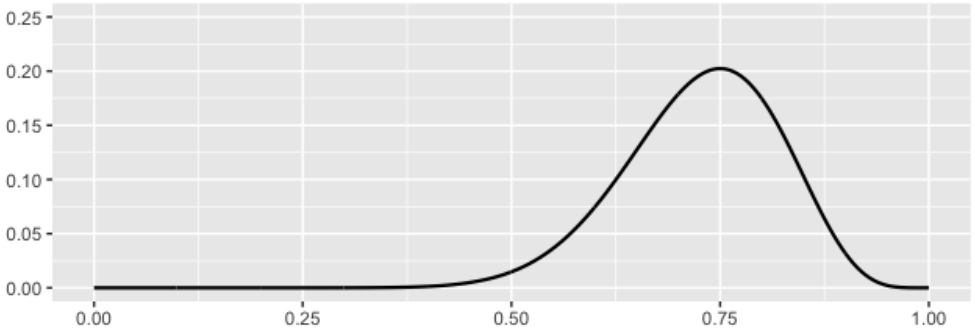
- Using the previously derived results we get that the posterior distribution of the patient response rate is Beta with updated parameters $(a + 15, b + 20 - 15) = (24.2, 18.8)$.

  The mean is $24.2/(24.2 + 18.8) = 0.56$, which gives a point estimate of the response rate.

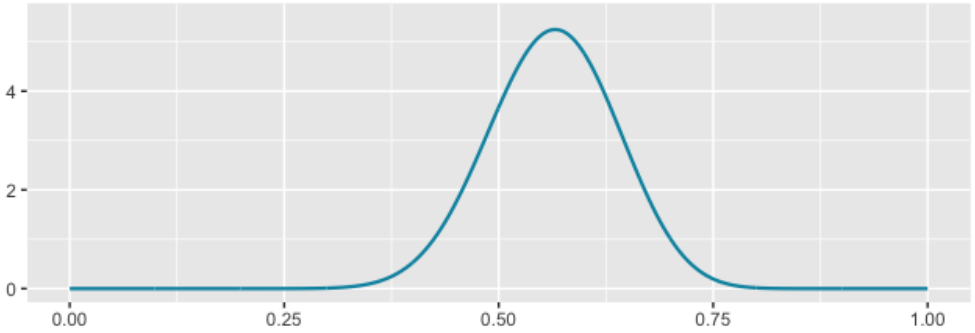
Drug example.



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

Drug example - Prediction.

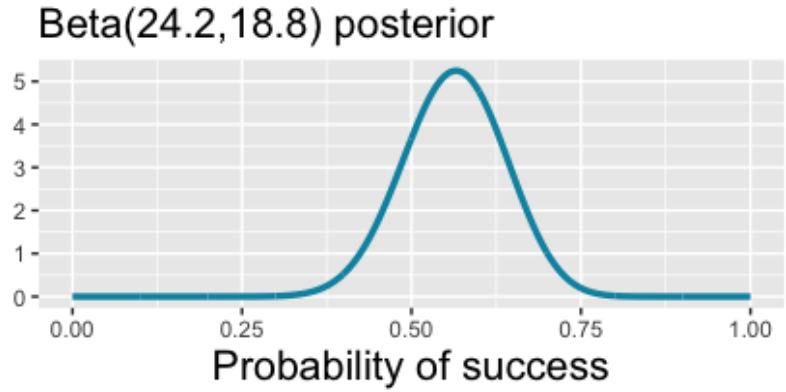
Suppose we would consider continuing the development programme if the drug managed to achieve at least a further 25 successes out of 40 future trials.
What is the probability that we can continue the drug development programme?

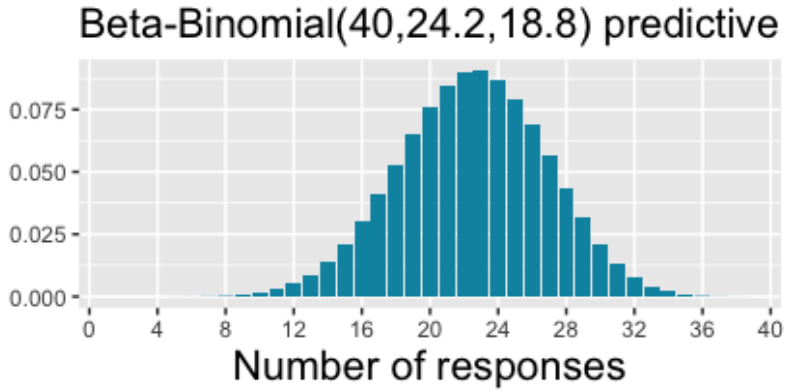
As we've seen before, we can use the posterior distribution to predict future observations. The posterior of the success rate in our case is Beta(24.2, 18.8).

One can show that the posterior predictive distribution, similarly to the prior predictive distribution, is Beta-Binomial. (Note that this is generally true for conjugate priors).

The predictive Beta-Binomial distribution of the number of successes $\tilde{y}_{40}$ in the next 40 trials has mean 22.5 and standard deviation 4.3.

From the Beta-Binomial distribution, we can then calculate $P(\tilde{y}_{40} \geq 25) = 0.329$.





Inference using the Normal distribution.

Suppose we have a sample of Normal data:

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

We will assume that the variance σ^2 is known and the mean θ is unknown.

Theoretically, the unknown parameter can take any real value, therefore we choose a normal prior for θ. Also, recall that the normal distribution in this setting is a conjugate prior:

$$\theta \sim N\left(\mu, \frac{\sigma^2}{n_0}\right),$$

where σ is the same standard deviation as in the likelihood, and μ and n_0 are known.

Recall that prior parameters can usually be interpreted as a prior sample. This prior sample is what's captured by the 'implicit' sample size n_0.

Then using the properties of the normal distribution, it is straightforward to show that the posterior distribution is also normal,

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

Inference using the Normal distribution - Comments.

- 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 $(-\infty, \infty)$.

  This is the reason we can use a normal distribution with large variance as a vague prior over the real line.

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

- The 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.

- Overall, as $n \to \infty$, we get $p(\theta|\boldsymbol{x}) \to N(\bar{x}, \sigma^2/n)$ which no longer depends on the prior.

  That is, the larger the sample size, the more the likelihood dominates the posterior.

- Compare this with the frequentist setting, where the MLE is $\hat{\theta} = \bar{x}$ with $\text{SE}(\hat{\theta}) = \sigma/\sqrt{n}$, and sampling distribution $N(\bar{x}, \sigma^2/n)$

Inference using the Normal distribution - THM Example.

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, $\bar{x}_z$ is 130 $\mu g/l$.
- Furthermore, 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 concentration be denoted by θ_z.

Note that a frequentist analysis would use the sample mean $\bar{x}_z = 130\mu g/l$ as an estimate of θ_z, with standard error $\sigma_e/\sqrt{n} = 5/\sqrt{2} = 3.5\mu g/l$.

Thus, a 95% confidence interval in the frequentist setting would be $\bar{x}_z \pm 1.96 \times \sigma_e/\sqrt{n}$, i.e. 123.1 to 136.9 $\mu g/l$.

THM Example - Bayesian Inference.

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 $\mu g/l$, which suggests a $N(120, 10^2)$ prior for θ_z.

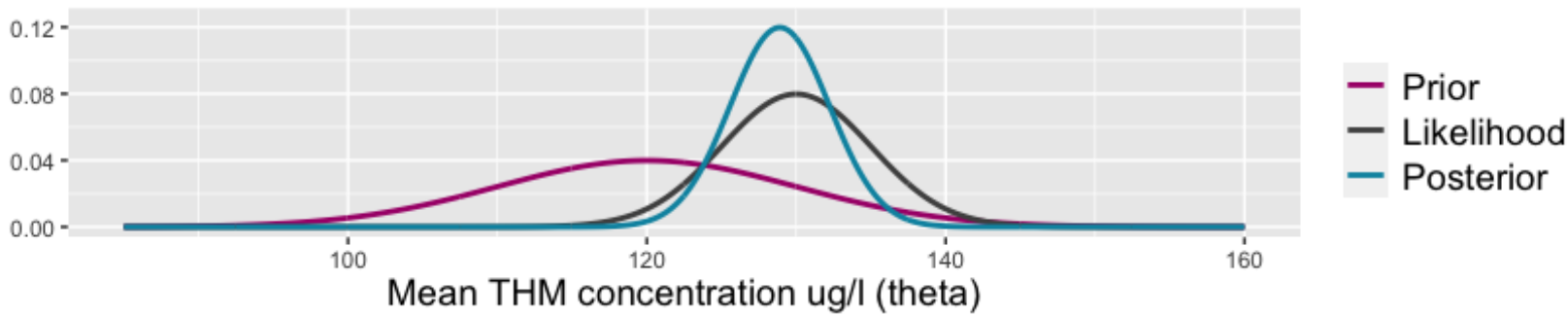
Note that expressing the prior standard deviation as $\sigma_e/\sqrt{n_0}$, gives $n_0 = (\sigma_e/10)^2 = 0.25$.

So our prior can be written as $\theta_z \sim N(120, \sigma_e^2/0.25)$.

Posterior for θ_z is then

$$p(\theta_z|\boldsymbol{x}) = N\left(\frac{0.25 \times 120 + 2 \times 130}{0.25 + 2}, \quad \frac{5^2}{0.25 + 2}\right) = (128.9, 3.33^2).$$

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



THM Example - 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 can be shown to be

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

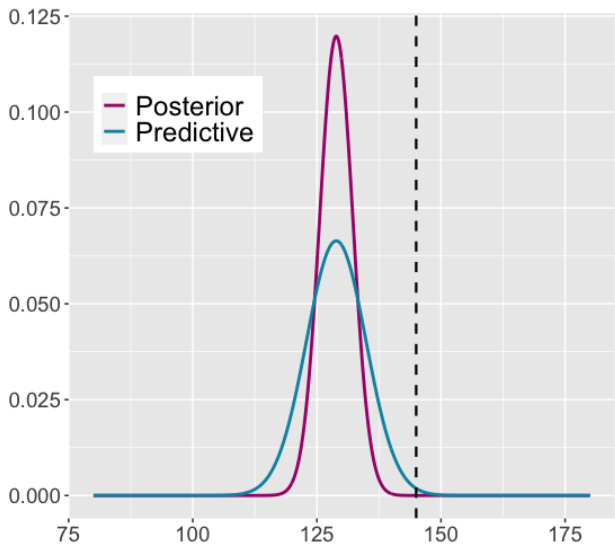
So the predictive distribution is centred around the posterior mean with variance equal to the sum of the posterior variance and the sample variance of $\tilde{x}$.

- Suppose the water company will be fined if THM levels in the water supply exceed $145\mu g/l$. What is the probability of this happening?
- Using the above formula the predictive distribution for THM concentration in a future sample taken from the water zone is

$$N(128.9, 3.33^2 + 5^2) = N(128.9, 36.1).$$

- Thus, the probability that THM concentration in future sample exceeds $145\mu g/l$ is

$$1 - \Phi[(145 - 128.9)/\sqrt{36.1}] = 0.004.$$



INFERENCE USING COUNT DATA

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

- Thus the likelihood is

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

- The unknown parameter μ can only take positive values, therefore we choose a Gamma prior.

  Also, 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 that the Gamma(a, b) density has mean a/b and variance a/b^2.

- This implies the following posterior

$$p(\mu|\boldsymbol{x}) \propto \mu^{a+n\bar{x}-1} e^{-(b+n)\mu} \propto Gamma(a + n\bar{x}, b + n).$$

  So the posterior is another (different) Gamma distribution.

Inference using count data - Comments.

- The Gamma distribution is a conjugate prior to the Poisson likelihood.
- Note that the posterior mean is

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

  Therefore the posterior mean is a compromise between the prior mean a/b and the MLE $\bar{x}$.

- When we need a vague prior for count data we can use a Gamma$(0.5, 0)$ prior, which is a vague prior on the positive half-line.

  The Gamma$(0.5, 0)$ prior is a so-called Jeffreys prior. (Jeffreys' priors is a family of priors, whose key feature is that they are invariant to reparameterisation).

  Its density function satisfies $p(\theta) \propto 1/\sqrt{\theta}$, which implies that the Gamma$(0.5, 0)$ prior is an improper prior.

  Improper means that the integral of the prior values is not finite (or doesn't exist).

  These functions don't form a proper probability distribution, but we can still use them as priors in Bayesian inference. The posterior will still usually be proper.

EXAMPLE - LONDON BOMBING 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 the 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, $\sum_i n_i x_i = 537$, and total number of areas, $\sum_i n_i = 576$.
- If the hits are assumed to be random, a Poisson distribution with constant hit rate θ should fit the data.
- We can think of $n = 576$ observations from a Poisson distribution, with $\bar{x} = 537/576 = 0.93$.

We will use the improper Gamma$(0.5, 0)$ distribution as prior for the mean θ of the Poisson distribution. This gives

$$p(\theta|\boldsymbol{y}) = Gamma(a + n\bar{x}, b + n) = Gamma(537.5, 576),$$

$$E(\theta|\boldsymbol{y}) = \frac{537.5}{576} = 0.933; \qquad Var(\theta|\boldsymbol{y}) = \frac{537.5}{576^2} = 0.0016.$$

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 standard deviation is less than each of the prior standard deviations and the standard error (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)*

Furthermore, 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.

In all three situations we considered, the prior was conjugate to the likelihood. This allowed us to find the posterior and the predictive distribution analytically.

JAGS uses indirect sampling to build up the posterior. The algorithm doesn't require conjugate priors, and can be used even when the posterior doesn't have a closed form.

Bayesian inference in JAGS

When using JAGS there's no need to explicitly specify the posterior.

- We can just specify the prior and likelihood separately.
- JAGS contains algorithms to evaluate the posterior given (almost) arbitrary specification of prior and likelihood.

We will use the Drug example to demonstrate how JAGS can be used in Bayesian inference.

Recall that we were interested in the probability of having at least 25 successes out of 40 future trials.

In general, the model can be summarised as follows

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

Defining the model.

Drug model:

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

Line by line we can translate the previous model definition to JAGS syntax to get

```
# Model description
jags.mod <- function(){
  theta ~ dbeta(a,b)                      # prior disribution
  y ~ dbin(theta,n)                       # sampling distribution
  y.pred ~ dbin(theta,m)                  # predictive distribution
  P.crit <- ifelse(y.pred>=ncrit,1,0) # =1 if y.pred>=ncrit, 0 o/w
}
```

In the above definition the value of $a, b, n, m, ncrit$ and y are known. The latter is our data.

These known quantities can either be specified within the model definition, or can be passed onto JAGS in a list.

DATA.

Data and known quantities can be included in the model definition as follows.

```
jags.mod <- function(){
  theta ~ dbeta(9.2,13.8)              # prior disribution
  y ~ dbin(theta,20)                   # sampling distribution
  y.pred ~ dbin(theta,40)              # predictive distribution
  P.crit <- ifelse(y.pred>=25,1,0)     # =1 if y.pred>=ncrit, 0 o/w

  y <- 15                              # data
}
```

However including the data in the model definition is only feasible for fairly simple models. A more robust way of passing the known information onto JAGS is putting these quantities into a list.

```
# data
a=9.2; b=13.8
n=20; m=40
ncrit=25; y=15

jags.data <- list("a","b","y","n","m","ncrit")
```

Note that elements of the list can also be vectors.

FITTING THE MODEL.

Once the model is defined we can set the nodes we want to monitor. These are usually those parameters we later want to make inferences about. Monitoring the nodes is what makes sure that the sampled values are saved.

```
# Parameters we want to monitor
jags.param <- c("theta","y.pred","P.crit")
```

In JAGS stochastic nodes need to be initialised. JAGS can automatically generate initial values if these are not specified. This is fine when we have informative priors.

However with fairly 'vague' priors, it's better to provide reasonable values in an initial-values list. (This will be discussed later).

```
# Specify initial values
jags.inits <- function(){
  list("theta" = 0.7, "y.pred" = 20)
  }
```

The next step is to fit the model using

```
jags.mod.fit <- jags(data = jags.data, inits = jags.inits,
                     parameters.to.save = jags.param,
                     n.iter = 10000, model.file = jags.mod)
```

Output

Printing the fitted model gives the following numerical summary. (Recall that the theoretical value of `P.crit` was 0.329).

	mu.vect	sd.vect	2.5%	25%	50%	75%	97.5%
P.crit	0.334	0.472	0.000	0.000	0.000	1.000	1.000
theta	0.564	0.074	0.417	0.513	0.566	0.615	0.704
y.pred	22.577	4.272	14.000	20.000	23.000	26.000	31.000

Finally we can look at trace plots, histograms/density plots of the monitored nodes.