Chapter 4

Bayesian Inference

Estimation: Point estimates

There are many useful summaries for a typical value of a random variable with a particular distribution.

For example,

- the mean;
- the mode;
- the median.

The **mode** is used more often as a summary than is the case in frequentist statistics.

Estimation: Interval estimates

A more useful summary of the posterior distribution is one which also reflects its **variation**.

For example, a $100(1-\alpha)\%$ **Bayesian confidence interval** for θ is any region C_α that satisfies $\Pr(\theta \in C_\alpha | \boldsymbol{x}) = 1 - \alpha$.

If θ is a continuous quantity with posterior probability density function $\pi(\theta|\boldsymbol{x})$ then

$$\int_{C_\alpha} \pi(\theta|\boldsymbol{x})\, d\theta = 1 - \alpha.$$

Estimation: Interval estimates

The usual correction is made for discrete θ, that is, we take the largest region C_α such that $\Pr(\theta \in C_\alpha | \boldsymbol{x}) \leq 1 - \alpha$.

Bayesian confidence intervals are sometimes called **credible regions** or **plausible regions**.

Clearly these intervals are not unique, since there will be many intervals with the correct probability coverage for a given posterior distribution.

Mozart, Earthquakes, speed cameras, glaciers

Think back to the examples in Chapters 2 and 3: the music expert, the geologist, the hydrologist, the glaciologist...

In each example, we used **Bayes' Theorem for distributions** to combine the beliefs of the expert, with our data, to form the **posterior distribution**.

How should we feedback our findings to the experts?

Mozart, Earthquakes, speed cameras, glaciers

We often used a single posterior summary – usually the mean (often in the form of the **Bayes Linear rule**), or the mode.

But we really need more than this!

We *could* supply the expert with the full posterior.

Often, we use **Bayesian confidence intervals**.

Example: Flood defence in New Orleans

We will think about this in more detail in **Case Study 2**.

Hurricane Katrina destroyed centuries–old sea walls (levees) in New Orleans in August 2005.

Example: Flood defence in New Orleans

A Bayesian analysis of sea–surge in the Gulf of Mexico could be used to help design the new sea walls. This would:

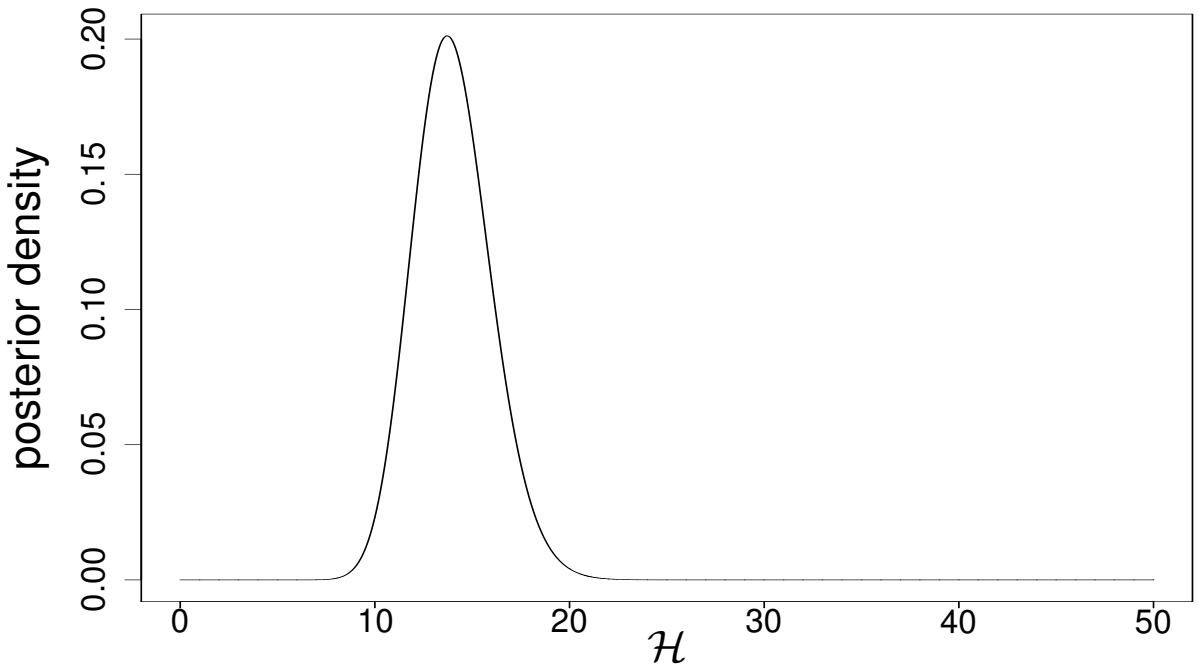
- Use data on the height of the sea (notably sea–surge)
- Supplement the analysis with the knowledge of an expert – perhaps an **oceanographer** – through a suitable prior distribution for the parameter of interest

Example: Flood defence in New Orleans

A recent Bayesian analysis by researchers at Tulane University, New Orleans (**JRSS, 2012**), found that

$$\mathcal{H}|\boldsymbol{x} \sim Ga(49, 3.5),$$

where $\mathcal{H}$ is the height (in feet) of the new sea wall, estimated to protect New Orleans against the **once in a hundred year** storm surge:



Example: Flood defence in New Orleans

What do we tell the engineers?

The gamma distribution for $\mathcal{H}|\boldsymbol{x}$ covers the entire positive real line...

- "Your sea wall needs to be be anywhere between 0 feet and ∞ feet tall"?
- "Your sea wall needs to be

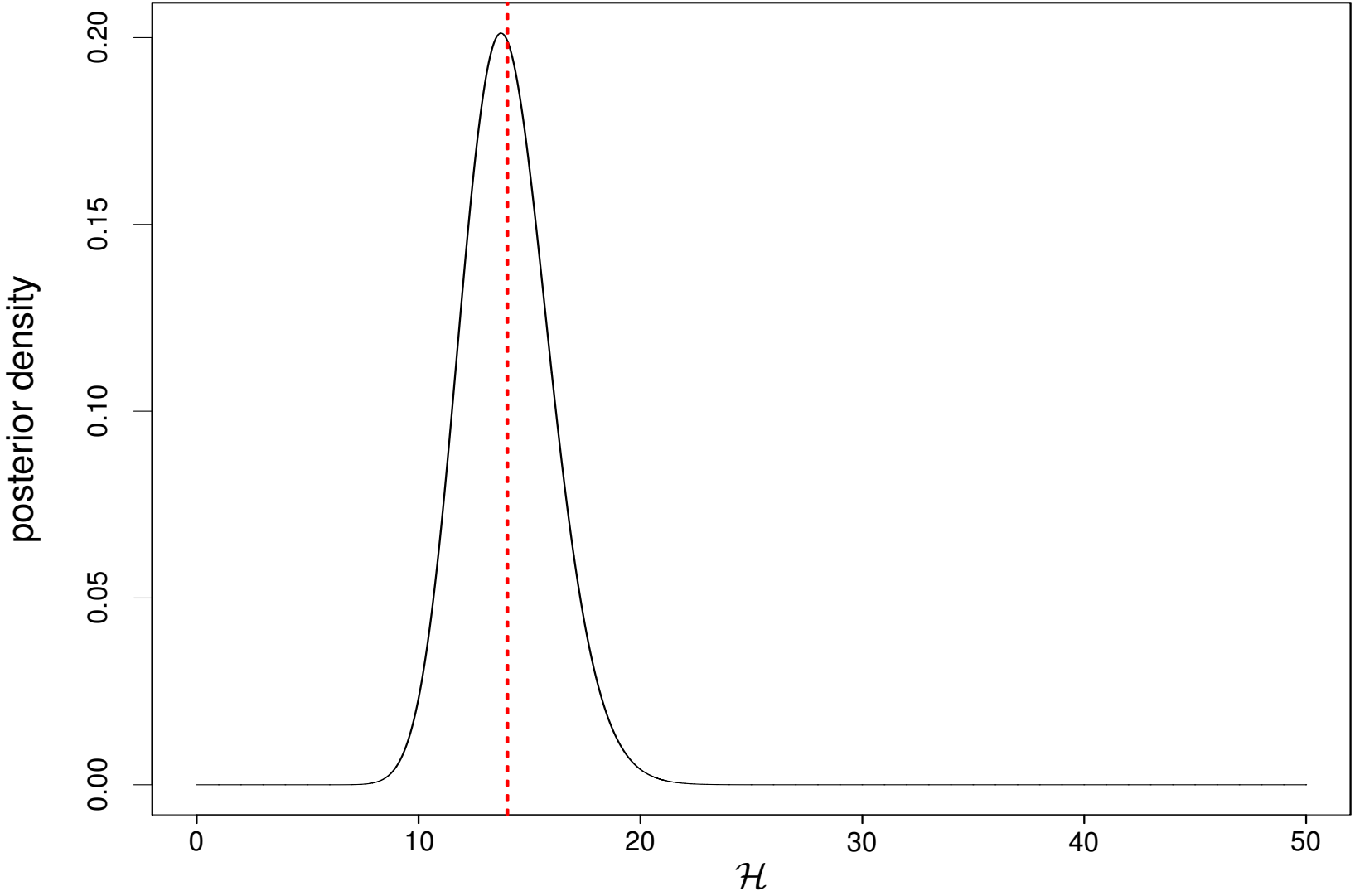
$$E(\mathcal{H}|\boldsymbol{x}) = \frac{49}{3.5} = 14 \text{ feet tall"?}$$

- "Your sea wall needs to be

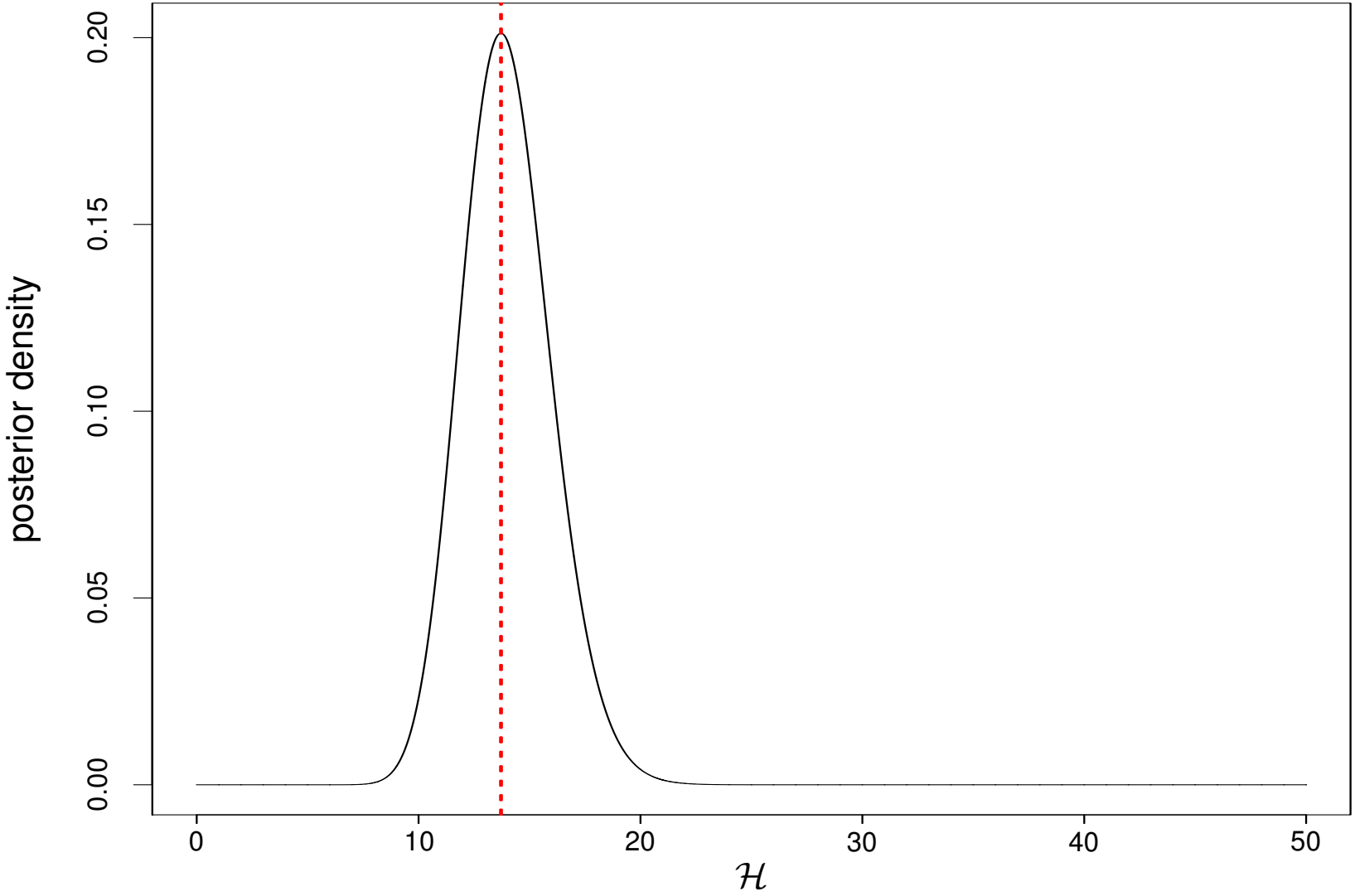
$$Mode(\mathcal{H}|\boldsymbol{x}) = \frac{49-1}{3.5} = 13.7 \text{ feet tall"?}$$

- $SD(\mathcal{H}|\boldsymbol{x}) = \sqrt{\frac{49}{3.5^2}} = 2$ feet

Example: Flood defence in New Orleans



Example: Flood defence in New Orleans



Estimation: Interval estimates

A $100(1-\alpha)\%$ **highest density interval** (HDI) for θ is the region $C_\alpha = \{\theta : \ \pi(\theta|\boldsymbol{x}) \geq \gamma\}$ where γ is chosen so that $\Pr(\theta \in C_\alpha|\boldsymbol{x}) = 1-\alpha$.

This region is sometimes called a **most plausible Bayesian confidence interval**.

If the posterior distribution has many modes then it is possible that the HDI will be the union of several disjoint regions.

For example, the HDI could take the form $C_\alpha = (a,b) \cup (c,d) \cup (e,f)$, where $a < b < c < d < e < f$.

Interpretation of confidence intervals

Suppose C_B is a 95% **Bayesian** confidence interval for θ and C_F is a 95% **frequentist** confidence interval for θ.

These intervals do not have the same interpretation:

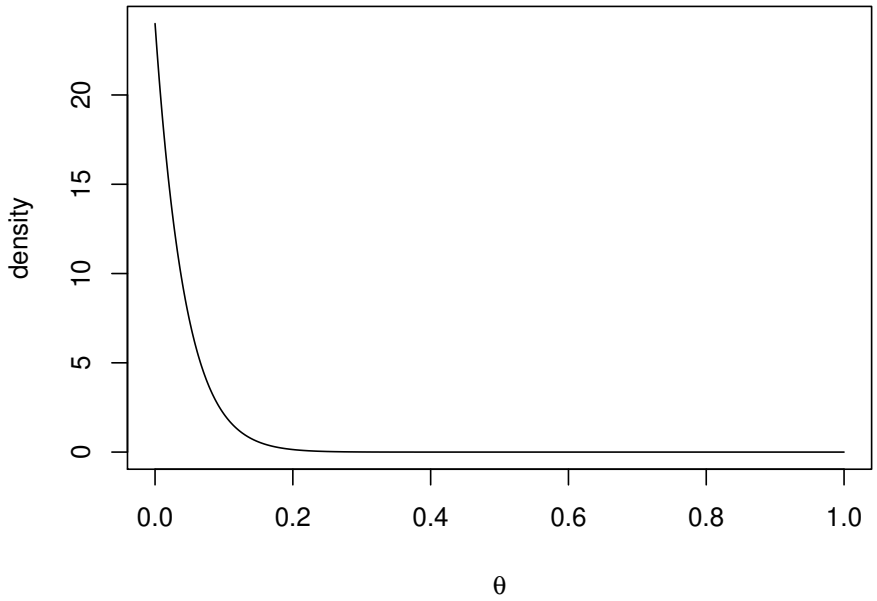
- the probability that C_B contains θ is 0.95;
- the probability that C_F contains θ is either 0 or 1 — since θ does not have a (non-degenerate) probability distribution;
- the interval C_F covers the true value θ on 95% of occasions — in repeated applications of the formula.

Example 4.1

Suppose that the posterior distribution for θ is a *Beta*$(1, 24)$ distribution, with probability density function

$$\pi(\theta|\boldsymbol{x}) = 24\,(1-\theta)^{23}, \quad 0 < \theta < 1.$$

A plot of this distribution is given in Figure 4.1 (See also the gravestones question in Chapter 5):



Determine the $100(1-\alpha)\%$ HDI for θ.

Solution to Example 4.1 (1/2)

The HDI must include those values of θ with highest posterior density and so must take the form $C_\alpha = (0, b)$.

The end–point b must satisfy

$$\int_0^b 24\,(1-\theta)^{23}\,d\theta = 1 - \alpha.$$

Now

$$\int_0^b 24\,(1-\theta)^{23}\,d\theta = \Big[-(1-\theta)^{24}\Big]_0^b = 1 - (1-b)^{24}.$$

Solution to Example 4.1 (2/2)

Hence

$$1-(1-b)^{24} = 1-\alpha \quad \Longrightarrow \quad 1-b = \alpha^{1/24} \quad \Longrightarrow \quad b = 1-\alpha^{1/24}.$$

Therefore, a $100(1-\alpha)\%$ HDI for θ is $(0, 1-\alpha^{1/24})$.

For example, the 95% HDI for θ is

$$(0, 1-(0.05^{1/24})) \longrightarrow (0, 0.117).$$

Example 4.2

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ from a $N(\mu, 1/\tau)$ distribution (where τ is known).

We have seen that, assuming vague prior knowledge, the posterior distribution is $\mu|\boldsymbol{x} \sim N(\bar{x}, 1/(n\tau))$.

Determine the $100(1-\alpha)\%$ HDI for μ.

Solution to Example 4.2 (1/1)

This distribution has a symmetric bell shape and so the HDI takes the form $C_\alpha = (a, b)$ with end–points

$$a = \bar{x} - \frac{z_{\alpha/2}}{\sqrt{n\tau}} \qquad \text{and} \qquad b = \bar{x} + \frac{z_{\alpha/2}}{\sqrt{n\tau}},$$

where z_α is the upper α-quantile of the $N(0, 1)$ distribution.

Therefore, the 95% HDI for μ is

$$\left(\bar{x} - \frac{1.96}{\sqrt{n\tau}}, \bar{x} + \frac{1.96}{\sqrt{n\tau}}\right).$$

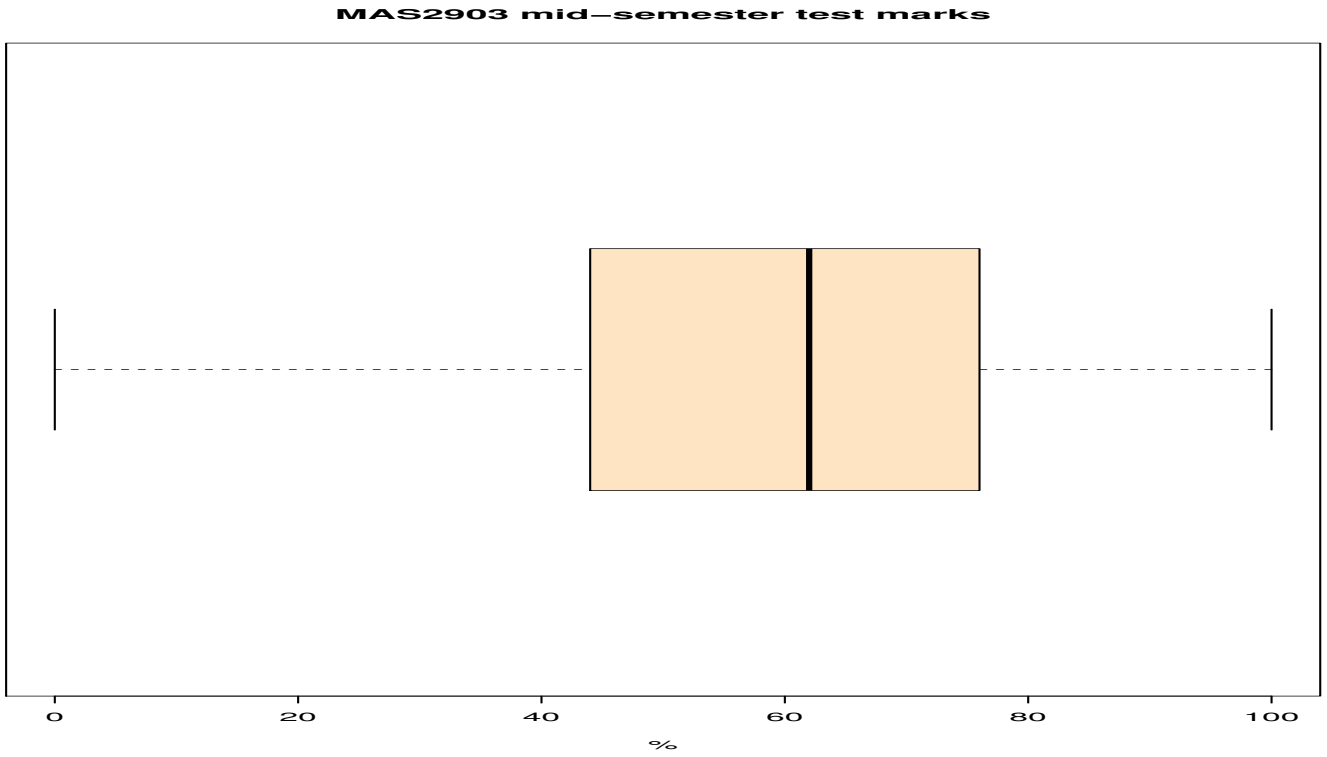
Note that this interval is numerically identical to the 95% frequentist confidence interval for the (population) mean of a normal random sample with known variance. However, the interpretation is very different.

Mid-semester test feedback

- It was hard, so WELL DONE!
- Not sure you were expecting *MATCH* output (although I did say you should revise up to page 66)
- Expect *MATCH* output in the exam!
- Official marks in NESS have been adjusted slightly to account for overly-harsh marking – done fairly, across the board

Mid-semester test feedback

Q1	Median	Mean	Q3	Max.	St. dev.
44	62	58	76	100	26



Specific feedback: Question 1 (3 marks)

- Generally very well done
- Some people not reading carefully – just assuming that the spinner was classical, and that the jockey was a Bayesian...
- Some students choosing a different word for each scenario

Specific feedback: Question 2 (8 marks)

- Done very well by almost everyone
- Language: “Speed cameras = 16/47" (as opposed to $\Pr(SC|R) = 16/47$)

Specific feedback: Question 3 (18 marks)

- Hardly anyone wrote “roulette method” (and instead wrote ”Prior knowledge” or ”Gamma distribution”
- Lots of students didn’t know where to find parameter values
- Hardly anyone could phrase the question correctly, **using the percentiles**
- Need to state the domain – i.e. $\theta > 0$ (-1/2 mark) COME ON!
- Comments at end – done quite well, by those who attempted it. Should also comment on prior/data discrepancies

Specific feedback: Question 4 (14 marks)

- Part (a)
  - Not including limits on the sum/product (so WHAT?!)
  - Not pulling the 3 outside the product to give 3^n
  - Other errors: Thinking

$$\prod_{i=1}^{n} x_i^2 = \bar{x}^2 \quad \text{or} \quad n\bar{x}^2$$

  - Also, pulling the 3 out of the product and not raising it to the power n
- Part (b)
  - Writing $T = \sum_{i=1}^{n} x_i^3$ – i.e. lower case x
  - Generally, just errors in notation

Specific feedback: Question 4 (14 marks)

- Part (c)
  - Most students found the correct proportional function but then never went on to say

$$\lambda|\boldsymbol{x} \sim InvGa(a+n, b+t).$$

  - Students often thought the posterior was a standard gamma distribution
  - General algebraic mistakes – e.g.

$$e^{-b/\lambda} \cdot e^{(-1/\lambda)t} = e^{-(b+1)t/\lambda}$$

Specific feedback: Question 5 (7 marks)

- Not having square root over $d + n\tau$
- Incorrectly identifying τ and d
- Mistakes getting inequalities the wrong way round
- General arithmetic mistakes

Assignment feedback

- **Question 4** (Air France disaster)
  - Done well by everyone
  - Some issues with “odds” – see car example in Chapter 1!
- **Question 9** (Rayleigh likelihood and sufficient statistic)
  - Main problem – students not choosing the correct functions h and g and ending up with two sufficient statistics... only one parameter!
  - Otherwise done well

Assignment feedback

- **Question 10** (Henry Cavendish on the Earth's density)
  - Generally done well
  - A lot of

$$\Pr(5.45 < \mu < 5.55) = \Pr(\mu < 5.55) - \Pr(\mu > 5.45)$$

- **Question 11** (Posterior $\propto$ Prior $\times$ Likelihood)
  - Exponential prior – PDF should not contain x! We have $\theta \sim \text{Exp}(1)$, so

$$\pi(\theta) = 1 \cdot e^{-1 \cdot \theta} = e^{-\theta}.$$

  - Confusion over the binomial range

Assignment feedback

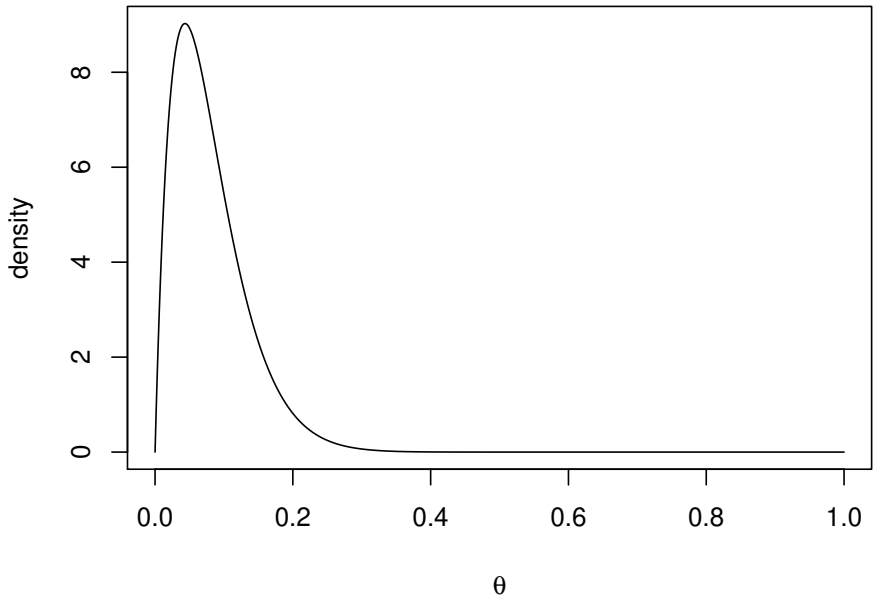
- **Question 12** (Bicycle tyre punctures)
  - Main comment – students bnot commenting on the change from prior to posterior!

Example 4.3

Suppose that the posterior distribution for θ is a *Beta*$(2, 23)$ distribution, with probability density function

$$\pi(\theta|\boldsymbol{x}) = 552\,\theta(1-\theta)^{22}, \quad 0 < \theta < 1.$$

A plot of this distribution is given in Figure 4.2.



Determine the $100(1-\alpha)\%$ HDI for θ.

Solution to Example 4.3

The HDI must take the form $C_\alpha = (a, b)$ if it is to include the values of θ with the highest probability density.

Therefore, the end–points a and b must satisfy

$$\int_a^b 552\,\theta(1-\theta)^{22}\,d\theta = 1-\alpha \qquad \text{and} \qquad 552a(1-a)^{22} = 552b(1-b)^{22}$$

Unfortunately, there is no simple analytical solution to these equations and so numerical methods have to be employed to determine a and b.

Computation of HDIs for unimodal distributions

Suppose that we require the HDI (a, b) for a unimodal distribution with density $f(\cdot)$ and distribution function $F(\cdot)$.

We have seen that if one of the end–points is known (because of the shape of the distribution) or the distribution is symmetric then the solution is in terms of the distribution's percentage points.

Computation of HDIs for unimodal distributions

When this is not the case, the problem requires a numerical scheme to find a and b satisfying

$$F(b) - F(a) = 1 - \alpha \qquad \text{and} \qquad f(a) = f(b).$$

Thus,

$$F(b) - F(a) - (1 - \alpha) = 0 \qquad \text{and} \qquad f(b) - f(a) = 0.$$

Any numerical solver, such as `optim` in R, will search through different a and b until the above conditions are met.

We can help the computer by squaring both (ensuring both differences will always be positive); we can also combine both squares:

$$\{F(b) - F(a) - (1 - \alpha)\}^2 + \{f(b) - f(a)\}^2,$$

and then ask R to find a and b such that this sum is minimsed.

Computation of HDIs for unimodal distributions

In fact, in order to use `optim` in R we also need to introduce a 'tuning parameter' k, and hence minimise:

$$g(a, b) = \left\{F(b) - F(a) - (1 - \alpha)\right\}^2 + k\left\{f(b) - f(a)\right\}^2,$$

where $k > 0$ tries to ensure that both terms are zeroed.

Computation of HDIs for unimodal distributions

In R:

```
g=function(x)
     {
      a=x[1]
      b=x[2]
       (pbeta(b,2,23)-pbeta(a,2,23)-0.95)**2
       +0.0001*(dbeta(b,2,23)-dbeta(a,2,23))**2
     }
```

Then:

```
initiala=qbeta(0.025,2,23)
initialb=qbeta(0.975,2,23)
res=optim(c(initiala,initialb),g,
     method="L-BFGS-B",lower=0,upper=1)
a=res$par[1]
b=res$par[2]
```

Computation of HDIs for unimodal distributions

This gives

- $a = 0.002211733$
- $b = 0.1840109$
- $F(b) - F(a) = 0.9500041$
- $f(b) - f(a) = -0.004484121$.

Thus the 95% HDI is $(0.002211733, 0.1840109)$.

Welcome back! Plan for the rest of term...

- This week:
  - Monday: Lecture (page 93-96/7 in notes)
  - Tuesday: Drop-in for assignment? Or lecture?
  - Thursday: Drop-in for assignment? Or lecture (page 96/7-100)?
  - Friday: Assignment 2 due in by 4pm
- Next week:
  - No Monday lecture as it's Bank Holiday!
  - Double lecture on Tuesday instead – half case study, half lecture: **2–4pm**, LT1
  - Thursday: Last problems class – remaining questions from Chapter 5

Welcome back! Plan for the rest of term...

- Week after:
  - Monday: Revision
  - Tuesday: Revision
  - Thursday: Revision, and assignment 2 feedback
- Week after that: Exam!

Prediction

Much of statistical inference (both Frequentist and Bayesian) is aimed towards making statements about a parameter θ.

Often the inferences are used as a yardstick for similar **future experiments**.

For example, we may want to **predict** the outcome when the experiment is performed again.

Prediction

Clearly there will be uncertainty about the future outcome of an experiment.

Suppose this future outcome Y is described by a probability (density) function $f(y|\theta)$.

There are several ways we could make inferences about what values of Y are likely.

Prediction

For example, if we have an estimate $\hat{\theta}$ of θ we might base our inferences on $f(y|\theta = \hat{\theta})$ (i.e. a **frequentist** approach).

Obviously this is not the best we can do, as such inferences ignore the fact that it is very unlikely that $\theta = \hat{\theta}$.

Prediction

Implicit in the Bayesian framework is the concept of the **predictive distribution**.

This distribution describes how likely are different outcomes of a future experiment.

The predictive probability (density) function is calculated as

$$f(y|\boldsymbol{x}) = \int_{\Theta} f(y|\theta)\,\pi(\theta|\boldsymbol{x})\,d\theta$$

when θ is a continuous quantity.

From this equation, we can see that the predictive distribution is formed by weighting the possible values of θ in the future experiment $f(y|\theta)$ by how likely we believe they are to occur $\pi(\theta|\boldsymbol{x})$.

Prediction

If the true value of θ were known, say θ_0, then any prediction can do no better than one based on $f(y|\theta = \theta_0)$.

However, as (generally) θ is unknown, the predictive distribution is used as the next best alternative.

Prediction

We can use the predictive distribution to provide a useful range of plausible values for the outcome of a future experiment.

This **prediction interval** is similar to a HDI interval.

A $100(1-\alpha)\%$ **prediction interval** for Y is the region $C_\alpha = \{y : \ f(y|\boldsymbol{x}) \geq \gamma\}$ where γ is chosen so that $\Pr(Y \in C_\alpha|\boldsymbol{x}) = 1-\alpha$.

Example 4.4

Suppose that X is the number of expensive goods sold in a shop over 24 days.

If θ is the expected number of sales per day then it may be plausible that $X|\theta \sim Po(24\,\theta)$.

Also, suppose our prior distribution for θ is as given in Table 4.1.

	θ	$\pi(\theta)$
"great"	1/2	0.2
"good"	1/4	0.5
"poor"	1/8	0.3

Example 4.4

Clearly, we believe that the most likely value of θ is 1/4, indicating that we would expect around 6 expensive goods to be sold in any 24 day period.

Suppose now that we observe that $x = 10$ expensive goods were sold in the last 24 days.

This will impact our beliefs about θ.

We can calculate the posterior distribution for θ as follows.

Example 4.4

The likelihood term is

$$\Pr(X = 10|\theta) = \frac{(24\theta)^{10}e^{-24\theta}}{10!} = \begin{cases} 0.1048 & \text{if } \theta = 1/2 \\ 0.0413 & \text{if } \theta = 1/4 \\ 0.0008 & \text{if } \theta = 1/8 \end{cases}$$

and so, using Bayes' Theorem,

$$\begin{aligned}\pi(\theta = 1/2|x = 10) = \frac{\Pr(X = 10|\theta = 1/2)\pi(\theta = 1/2)}{[\Pr(X = 10|\theta = 1/2)\pi(\theta = 1/2)} \\ + \Pr(X = 10|\theta = 1/4)\pi(\theta = 1/4) \\ + \Pr(X = 10|\theta = 1/8)\pi(\theta = 1/8)]\end{aligned}$$

$$= \frac{0.1048 \times 0.2}{0.1048 \times 0.2 + 0.0413 \times 0.5 + 0.0008 \times 0.3}$$

$$= 0.501$$

Example 4.4

Similarly,

$$\begin{aligned} \pi(\theta = 1/4|x = 10) &= \cdots = 0.493 \qquad \text{and} \\ \pi(\theta = 1/8|x = 10) &= \cdots = 0.006. \end{aligned}$$

Thus, the posterior distribution for θ is as shown in Table 4.2, with the most likely value of θ now being 1/2, and standard deviation $SD(\theta|x = 10) = 0.126$.

	θ	$\pi(\theta\|x = 10)$
"great"	1/2	0.501
"good"	1/4	0.493
"poor"	1/8	0.006

Example 4.4

Suppose now we want to predict the number of sales Y in the next 24 days.

If there have been no changes in the sales process (no special advertising campaigns etc) then we can take $Y|\theta \sim Po(24\,\theta)$.

Determine the **predictive probability function** for Y.

Solution to Example 4.4 (1/2)

As θ is discrete, the predictive probability function for Y is

$$
\begin{aligned}
f(y|x=10) &= \sum_{\theta=\frac{1}{2},\frac{1}{4},\frac{1}{8}} f(y|\theta)\,\pi(\theta|x=10) \\
&= \left(\frac{12^y e^{-12}}{y!} \times 0.501\right) + \left(\frac{6^y e^{-6}}{y!} \times 0.493\right) \\
&\qquad + \left(\frac{3^y e^{-3}}{y!} \times 0.006\right) \\
&= \frac{0.501(12^y e^{-12}) + 0.493(6^y e^{-6}) + 0.006(3^y e^{-3})}{y!}
\end{aligned}
$$

Solution to Example 4.4 (2/2)

For example:

$$\begin{aligned}\Pr(Y = 10|x = 10) &= \frac{0.501(12^{10}e^{-12}) + 0.493(6^{10}e^{-6}) + 0.006(3^{10}e^{-3})}{10!} \\ &= 0.073.\end{aligned}$$

Example 4.4

This probability can be compared with a more **naive** predictive probability calculated assuming that $\theta = \hat{\theta}$, the likelihood mode.

Here, $\hat{\theta} = 1/2$ and so $Y|\theta = \hat{\theta} \sim Po(24\hat{\theta}) \equiv Po(12)$, whence

$$\Pr(Y = 10|\theta = \hat{\theta}) = \frac{12^{10}e^{-12}}{10!} = 0.1048.$$

In the same manner, we can calculate the **entire predictive distribution** and **entire naive predictive distribution**; see Table 4.3.

Example 4.4

y	correct $f(y\|x = 10)$	naive $f(y\|\theta = \hat{\theta})$
0	0.002	0.000
1	0.008	0.000
2	0.024	0.000
3	0.046	0.002
4	0.070	0.005
5	0.086	0.013
6	0.092	0.025
7	0.090	0.044
8	0.084	0.066
9	0.078	0.087
10	0.073	0.105
11	0.068	0.114
12	0.063	0.114
13	0.055	0.106
14	0.046	0.090
15	0.037	0.072
16	0.027	0.054
17	0.019	0.038
18	0.013	0.026
19	0.008	0.016
20	0.005	0.010
.	.	.

Example 4.4

Notice that the correct predictive probability distribution has more probability out in the tails of its distribution – that is, the probabilities of 0, 1, 2, . . . are larger than their “naive” equivalents.

- Common occurrence
- Due to ignoring the uncertainty about the parameter estimate

Example 4.4

Essentially, the naive predictive distribution is a predictive distribution which, instead of using the correct posterior distribution, uses the degenerate posterior distribution

$$\pi^*(\theta|x = 10) = \begin{cases} 1 & \text{if } \theta = 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

a distribution with standard deviation $SD_{\pi^*}(\theta|x = 10) = 0$.

The *correct* posterior standard deviation of θ is $SD_{\pi}(\theta|x = 10) = 0.126$.

Example 4.4

Therefore, the predictive distribution using the naive posterior π^* is, loosely speaking, **too confident** that it "knows" the value of θ and so produces a predictive distribution with too small a standard deviation:

$$SD(Y|x = 10) = \begin{cases} 4.26 & \text{using the correct } \pi(\theta|x = 10) \\ 3.46 & \text{using the naive } \pi^*(\theta|x = 10). \end{cases}$$

Example 4.4

We can also use the above table of predictive probabilities to determine a **prediction interval** for Y.

Using the correct predictive distribution, and recalling the highest density feature of prediction intervals, we obtain

$$\Pr(2 \leq Y \leq 17 | x = 10) = 0.959.$$

The corresponding naive calculation is

$$\Pr(6 \leq Y \leq 19 | \theta = \hat{\theta}) = 0.958.$$

Hence the correct (approximate) 96% prediction interval for Y is $\{2, 3, \ldots, 17\}$.

The naive version, and hence narrower interval, is $\{6, 7, \ldots, 19\}$.

Definition (Beta–binomial distribution)

The random variable Y follows a Beta–binomial $BetaBin(n, a, b)$ distribution (n positive integer, $a > 0$, $b > 0$) if it has probability function

$$f(y|n, a, b) = \binom{n}{y} \frac{B(y + a, b + n - y)}{B(a, b)}, \qquad y = 0, 1, \ldots, n,$$

where $B(a, b)$ is the beta function defined earlier. It can be shown that

$$E(Y) = \frac{na}{a + b} \qquad \text{and} \qquad Var(Y) = \frac{nab(a + b + n)}{(a + b)^2(a + b + 1)}.$$

Example 4.5

Suppose that X is the number of defective items in a sample of size 5.

If the items are defective independently of one another and they each have the same probability θ of being defective then $X|\theta \sim Bin(5, \theta)$.

Suppose we believe that defective items are quite unlikely and so take a $Beta(1, 19)$ prior distribution with mean and standard deviation

$$E(\theta) = 0.05 \qquad \text{and} \qquad SD(\theta) = 0.048.$$

Example 4.5

Suppose we take a sample of size 5 and observe $x = 1$ defective item.

In this case, the likelihood mode is $\hat{\theta} = 1/5 = 0.2$, higher than the prior mean.

We can combine the prior with our data via Bayes' Theorem, to show that the posterior distribution is a $Beta(2, 23)$ distribution, with mean and standard deviation

$$E(\theta|x = 1) = 0.08 \qquad \text{and} \qquad SD(\theta|x = 1) = 0.053.$$

The posterior mean is larger than the prior mean and the standard deviation has also increased (slightly).

Example 4.5

If we observe another sample of 5 items, what is the predictive probability distribution of the number found to be defective?

Solution to Example 4.5

Suppose the number of defective items in this future sample is Y, with $Y|\theta \sim Bin(5, \theta)$.

The predictive probability function of Y is, for $y = 0, 1, 2, \ldots, 5$,

$$\begin{aligned} f(y|x=1) &= \int_{\Theta} f(y|\theta)\,\pi(\theta|x=1)\,d\theta \\ &= \int_0^1 \binom{5}{y} \theta^y (1-\theta)^{5-y} \times \frac{\theta(1-\theta)^{22}}{B(2,23)}\,d\theta \\ &= \binom{5}{y} \frac{1}{B(2,23)} \int_0^1 \theta^{y+1}(1-\theta)^{27-y}\,d\theta \\ &= \binom{5}{y} \frac{B(y+2, 28-y)}{B(2,23)}, \end{aligned}$$

that is, $Y|x = 1 \sim BetaBin(5, 2, 23)$.

Example 4.5

We can compare this predictive distribution with a naive predictive distribution based on an estimate of θ, for example, the likelihood mode or the posterior mode.

Here we shall base our naive predictive distribution on the posterior mode $\hat{\theta} = 1/23$, that is, use the distribution

$$Y|\theta = \hat{\theta} \sim Bin(5, 1/23).$$

Thus, the naive predictive probability function is, for $y = 0, 1, \ldots, 5$,

$$f(y|\theta = \hat{\theta}) = \binom{5}{y} \hat{\theta}^y (1 - \hat{\theta})^{5-y} = \binom{5}{y} \frac{22^{5-y}}{23^5}.$$

Example 4.5

Numerical values for the predictive and naive predictive probability functions are given in Table 4.4:

y	correct $f(y\|x=1)$	naive $f(y\|\theta=\hat{\theta})$
0	0.680	0.801
1	0.252	0.182
2	0.058	0.017
3	0.009	0.001
4	0.001	0.000
5	0.000	0.000

Example 4.5

As before, the naive predictive distribution

- allows only one value for θ: $\theta = 1/23$
- has a standard deviation of zero: $SD(\theta|x = 1) = 0$

The **correct** posterior standard deviation is $SD(\theta|x = 1) = 0.053$.

This results in

$$SD(Y|x = 1) = \begin{cases} 0.652 & \text{using the correct predictive dist.} \\ 0.456 & \text{using the naive predictive dist.,} \end{cases}$$

these values being calculated from $BetaBin(5, 2, 23)$ and binomial $Bin(5, 1/23)$ distributions.

Example 4.5

Using the numerical table of predictive probabilities, we can see that $\{0, 1\}$ is a 93.2% prediction set/interval.

This is to be contrasted with the more "optimistic" calculation using the naive predictive distribution which shows that $\{0, 1\}$ is a 98.3% prediction set/interval.

Predictive distribution

In the previous example, a non–trivial integral had to be evaluated.

However, when the past data $\boldsymbol{x}$ and future data y are independent (given θ) and we use a conjugate prior distribution, another (easier) method can be used to determine the predictive distribution.

Predictive distribution

Using Bayes' Theorem, the posterior density for θ given $\boldsymbol{x}$ and y is

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}, y) &= \frac{\pi(\theta) f(\boldsymbol{x}, y|\theta)}{f(\boldsymbol{x}, y)} \\
&= \frac{\pi(\theta) f(\boldsymbol{x}|\theta) f(y|\theta)}{f(\boldsymbol{x}) f(y|\boldsymbol{x})} \qquad \text{since } \boldsymbol{X} \text{ and } Y \text{ are independent} \\
&= \frac{\pi(\theta|\boldsymbol{x})\, f(y|\theta)}{f(y|\boldsymbol{x})}.
\end{aligned}$$

Predictive distribution

Rearranging, we obtain

$$f(y|\boldsymbol{x}) = \frac{f(y|\theta)\pi(\theta|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x}, y)}.$$

The right–hand–side of this equation looks as if it depends on θ, but, in fact, any terms in θ will be cancelled between the numerator and denominator.

Example 4.5 revisited

We had

$$\theta \sim Beta(1, 19), \quad X|\theta \sim Bin(5, \theta), \quad Y|\theta \sim Bin(5, \theta),$$

from which we obtain

$$\theta|x = 1 \sim Beta(2, 23), \quad \theta|x = 1, y \sim Beta(y + 2, 28 - y).$$

Example 4.5 revisited

Therefore, for $y = 0, 1, 2, \ldots, 5$

$$
\begin{aligned}
f(y|x=1) &= \frac{f(y|\theta)\pi(\theta|x=1)}{\pi(\theta|x=1,y)} \\
&= \frac{\binom{5}{y}\theta^y(1-\theta)^{5-y} \times \frac{\theta(1-\theta)^{22}}{B(2,23)}}{} \frac{\theta^{y+1}(1-\theta)^{27-y}}{B(y+2,28-y)} \\
&= \binom{5}{y}\frac{\theta^y(1-\theta)^{5-y}\theta(1-\theta)^{22}B(y+2,28-y)}{\theta^{y+1}(1-\theta)^{27-y}B(2,23)} \\
&= \binom{5}{y}\frac{B(y+2,28-y)}{B(2,23)}.
\end{aligned}
$$

Definition (Inverse–Beta distribution)

The random variable Y follows a Inverse-Beta $InBe(a, b, c)$ distribution ($a > 0$, $b > 0$, $c > 0$) if it has probability density function

$$f(y|a, b, c) = \frac{c^b y^{a-1}}{B(a, b)(y + c)^{a+b}} \qquad y > 0,$$

where $B(a, b)$ is the beta function defined earlier. It can be shown that

$$E(Y) = \frac{ac}{b-1} \qquad \text{and} \qquad Var(Y) = \frac{ac^2(a + b - 1)}{(b-1)^2(b-2)}.$$

The distribution gets its name because $Y/(Y + c) \sim Beta(a, b)$. Also note that if $Y \sim InBe(a, b, 1)$ then $cY \sim InBe(a, b, c)$.

Example 4.6

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ from a $Ga(k, \theta)$ distribution, where k is known, and our prior beliefs are described by a $Ga(g, h)$ distribution.

The likelihood function is

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \frac{\theta^k x_i^{k-1} e^{-\theta x_i}}{\Gamma(k)} \propto \theta^{nk} e^{-n\bar{x}\theta}.$$

Therefore, using Bayes Theorem, the posterior density is

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)\, f(\boldsymbol{x}|\theta) \\
&\propto \theta^{g-1} e^{-h\theta} \times \theta^{nk} e^{-n\bar{x}\theta}, \quad \theta > 0 \\
&\propto \theta^{g+nk-1} e^{-(h+n\bar{x})\theta}, \quad \theta > 0.
\end{aligned}$$

Example 4.6

Hence, the posterior distribution is a $Ga(G = g + nk, H = h + n\bar{x})$ distribution.

Notice that this implies that the gamma distribution is the conjugate prior distribution for the model "random sample from a $Ga(k, \theta)$ distribution, with k known".

Determine the predictive distribution for a future outcome Y.

Solution to Example 4.6 (1/4)

The predictive density function is, for $y > 0$,

$$f(y|\boldsymbol{x}) = \frac{f(y|\theta)\pi(\theta|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x}, y)}.$$

Now

- $y|\theta \sim Ga(\theta, k)$
- $\theta|\boldsymbol{x} \sim Ga(G, H)$
- $\theta|\boldsymbol{x}, y \sim ??$

Solution to Example 4.6 (2/4)

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}, y) &\propto \pi(\theta|\boldsymbol{x}) \times f(y|\theta) \\
&\propto \theta^{G-1} e^{-H\theta} \times \theta^{k} y^{k-1} e^{-\theta y} \\
&\propto \theta^{G+k-1} e^{-(H+y)\theta}
\end{aligned}$$

giving $\theta|\boldsymbol{x}, y \sim Ga(G+k, H+y)$.

Solution to Example 4.6 (3/4)

Therefore, the predictive density function is, for $y > 0$,

$$
\begin{aligned}
f(y|\boldsymbol{x}) &= \frac{f(y|\theta)\pi(\theta|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x}, y)} \\
&= \frac{\dfrac{\theta^k y^{k-1} e^{-\theta y}}{\Gamma(k)} \times \dfrac{H^G \theta^{G-1} e^{-H\theta}}{\Gamma(G)}}{\dfrac{(H+y)^{G+k} \theta^{G+k-1} e^{-(H+y)\theta}}{\Gamma(G+k)}}
\end{aligned}
$$

Solution to Example 4.6 (4/4)

Giving

$$
\begin{aligned}
f(y|\boldsymbol{x}) &= \frac{\theta^{G+k-1}e^{-(H+y)\theta}y^{k-1}H^G\Gamma(G+k)}{\theta^{G+k-1}e^{-(H+y)\theta}(H+y)^{G+k}\Gamma(k)\Gamma(g)} \\
f(y|\boldsymbol{x}) &= \frac{\Gamma(G+k)H^G y^{k-1}}{\Gamma(k)\Gamma(G)(H+y)^{G+k}} \\
&= \frac{H^G y^{k-1}}{B(k,G)(H+y)^{G+k}},
\end{aligned}
$$

that is, $Y|\boldsymbol{x} \sim InBe(k, G, H)$.

Example 4.6

Consider the case where the data follow an exponential distribution, that is, where $k = 1$.

Determine the predictive density function and the $100(1 - \alpha)\%$ prediction interval for Y.

Solution to Example 4.6 (1/3)

Now

$$B(1, G) = \frac{\Gamma(G)\Gamma(1)}{\Gamma(G+1)} = \frac{(G-1) \times (G-2) \times \cdots}{G \times (G-1) \times (G-2) \times \cdots} = \frac{1}{G}.$$

So the predictive density function is

$$f(y|\boldsymbol{x}) \quad = \quad \frac{GH^G}{(H+y)^{G+1}}, \quad y > 0.$$

The prediction interval includes the most likely values of Y and so, given the shape of the density, must take the form $(0, b)$. The end–point b must satisfy

$$\int_0^b \frac{GH^G}{(H+y)^{G+1}}\, dy = 1 - \alpha.$$

Solution to Example 4.6 (2/3)

Using the substitution $z = H + y$, we have

$$\begin{aligned}
\int_0^b \frac{GH^G}{(H+y)^{G+1}}\,dy &= \int_H^{b+H} \frac{GH^G}{z^{G+1}}\,dz \\
&= \left[-\frac{H^G}{z^G}\right]_H^{b+H} \\
&= 1 - \frac{H^G}{(b+H)^G}.
\end{aligned}$$

Solution to Example 4.6 (3/3)

Therefore, we require

$$1 - \frac{H^G}{(b+H)^G} = 1 - \alpha \qquad \Longrightarrow \frac{b+H}{H} = \alpha^{-1/G}$$

$$\Longrightarrow b = H\left\{\alpha^{-1/G} - 1\right\}.$$

Hence, a $100(1-\alpha)\%$ prediction interval for Y is

$$\left(0, H\left\{\alpha^{-1/G} - 1\right\}\right).$$

The End!!